LQR Learning Pipelines

KIOS Graduate Training School 2024

Florian Dörfler

ETH zürich

Context & acknowledgements

- collaboration with Claudio dePersis & Pietro Tesi to develop an explicit version of regularized DeePC
  → **data-driven & regularized LQR**

Pietro Tesi (Florence)

Alessandro Chiuso (Padova)

Claudio de Persis (Groningen)

- extension to adaptive LQR with Feiran Zhao, Keyou You, Linbin Huang, & Alessandro Chiuso
  → **data-enabled policy optimization**

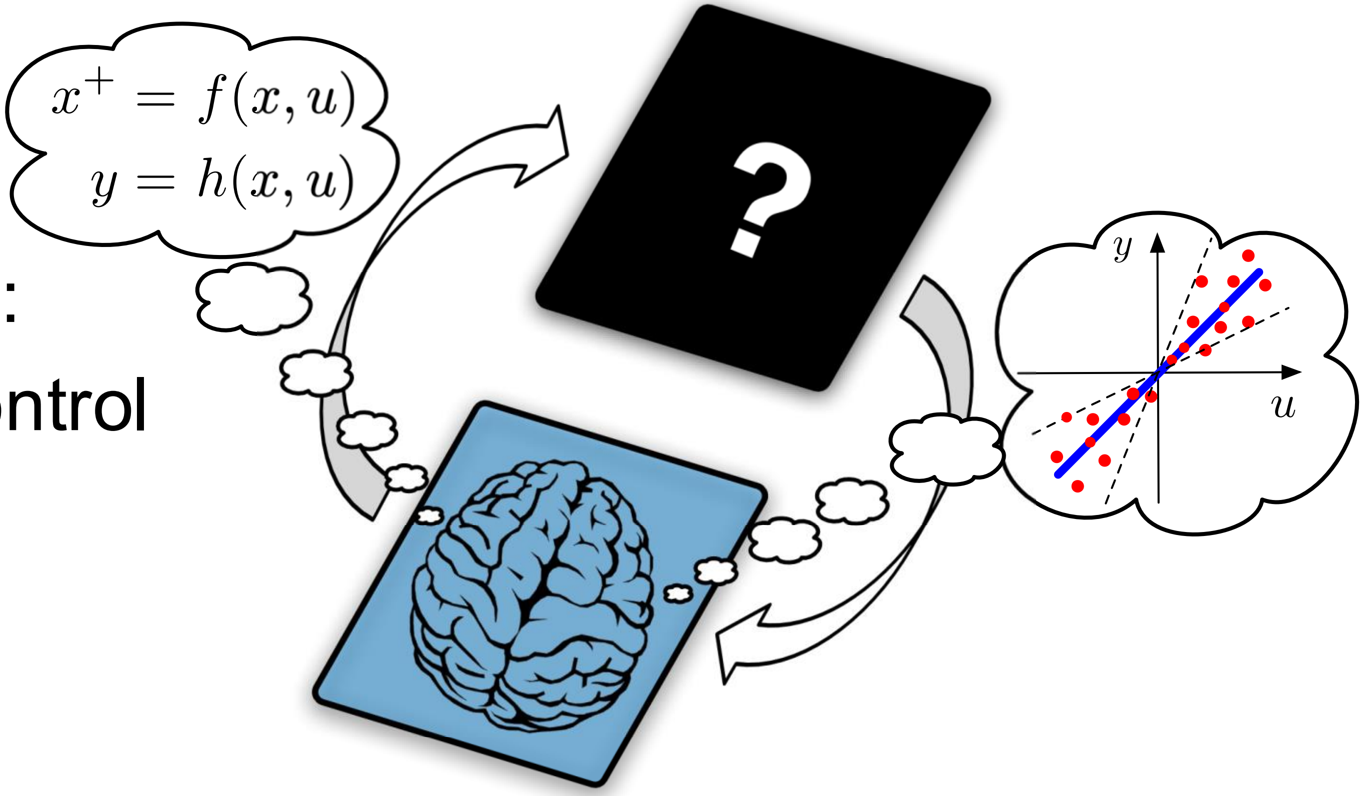
Feiran Zhao (Tsinghua)

Keyou You (Tsinghua)

Linbin Huang (Zhejiang)

- revisit ***old open problems with new perspectives***

Data-driven pipelines



- **indirect** (model-based) approach:
  data $\xrightarrow{\text{ID}}$ model + uncertainty → control

- **direct** (model-free) approach:
  direct MRAC, RL, behavioral, …

- **episodic & batch** algorithms:
  collect batch of data → design policy

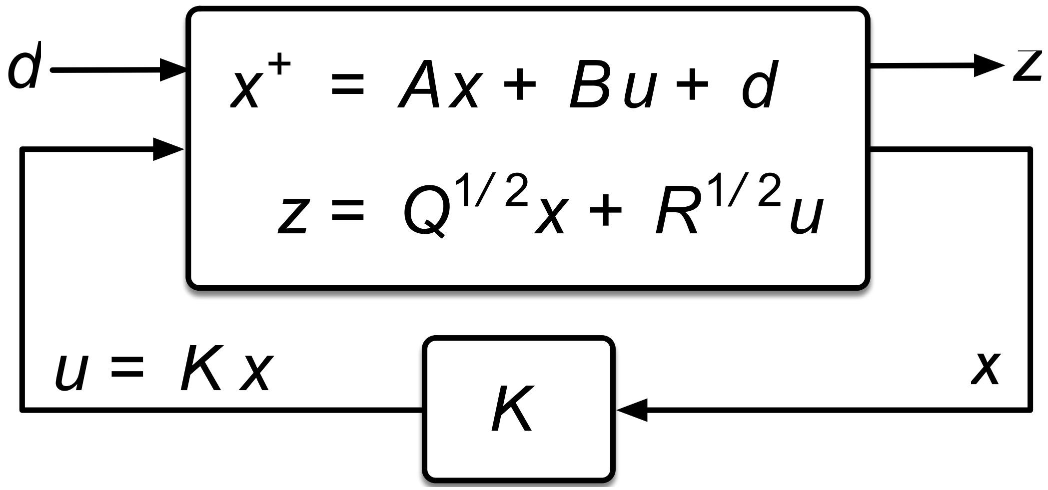
- **online & adaptive** algorithms:
  measure → update policy → actuate

well-documented **trade-offs** concerning
- complexity: data, compute, & analysis
- goal: optimality vs (robust) stability
- practicality: modular vs end-to-end …

→ **gold(?) standard**: direct, adaptive, optimal yet robust, cheap, & tractable

LQR

- **cornerstone** of automatic control



Equivalent LQR formulations:

- $J(K) = \sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t = \sum_{t=0}^{\infty} x_t^T Q x_t + x_t^T K^T R K x_t$
- solution to $x_{t+1} = (A+BK)x_t$ is $x_t = (A+BK)^t x_0$

$\rightsquigarrow J(K) = \sum_{t=0}^{\infty} x_0^T (A+BK)^{Tt} (Q + K^T R K)(A+BK)^t x_0$

- Recall the closed-loop observability Gramian: $W = \sum_{t=0}^{\infty} ((A+BK)^T)^t (Q + K^T R K)(A+BK)^t$

- W can also be obtained as the unique positive definite solution to the Lyapunov equation:

$$(A+BK)^T W (A+BK) - W + Q + K^T R K = 0$$

⇝ equivalent reformulation of $J(u) = x_0^T W x_0 = \mathrm{trace}(W \underbrace{x_0 x_0^T})$

covariance of (random) x_0

- yet another reformulation using $\mathrm{tr}(x_t^T Q x_t) = \mathrm{tr}(Q\, x_t x_t^T)$

$$J(K) = \mathrm{tr}(Q \cdot P) + \mathrm{tr}(K^T R K P)$$

where $P = \sum_{t=0}^{\infty} x_t x_t^T$ (state covariance)

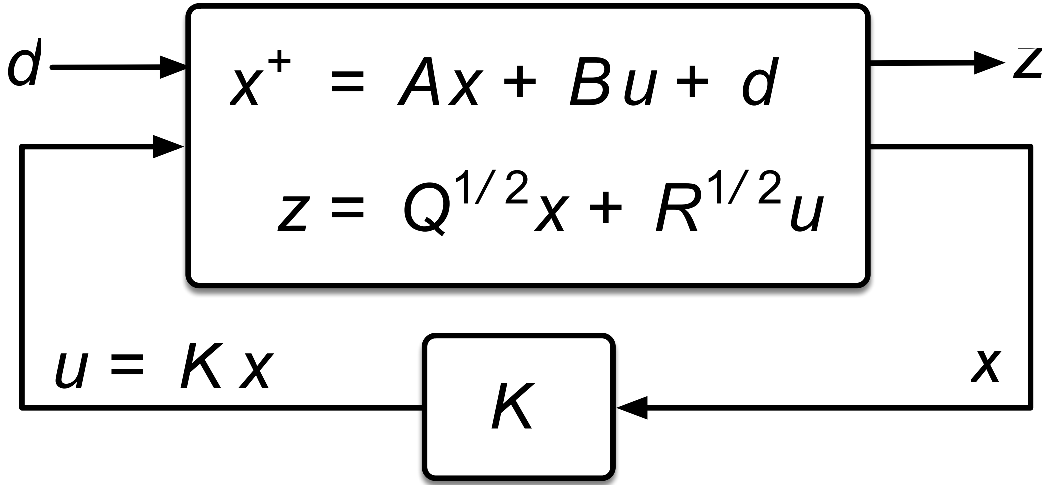
$$= \sum_{t=0}^{\infty} (A+BK)^t x_0 x_0^T \left((A+BK)^T\right)^t$$

- recall that the above is the controllability Gramian which can be calculated uniquely as positive definite solution to

$$(A+BK) P (A+BK)^T - P + x_0 x_0^T = 0$$

side note: as it turns the actual value of $x_0 x_0^T$ does not matter for the final result, and often one simply sets it to be identity

LQR



- **cornerstone** of automatic control

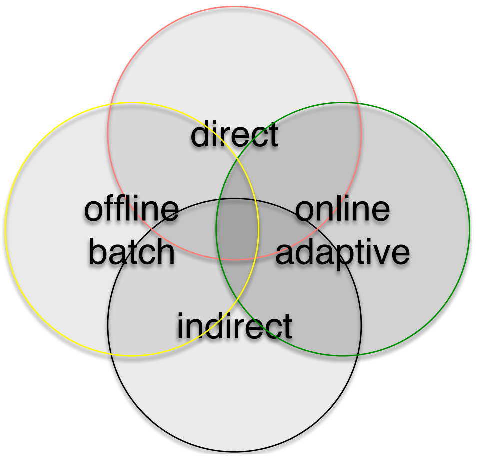
- $\mathcal{H}_2$ **parameterization**

  (can be posed as convex SDP,

  as differentiable program, as…)

$$
\begin{aligned}
\underset{P \succeq I,\, K}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top R K P\right) \\
\text{subject to} \quad & (A+BK)P(A+BK)^\top - P + I \preceq 0
\end{aligned}
$$

- <u>**the**</u> **benchmark** for all data-driven control approaches in last decades but there is **no direct & adaptive LQR**



Contents

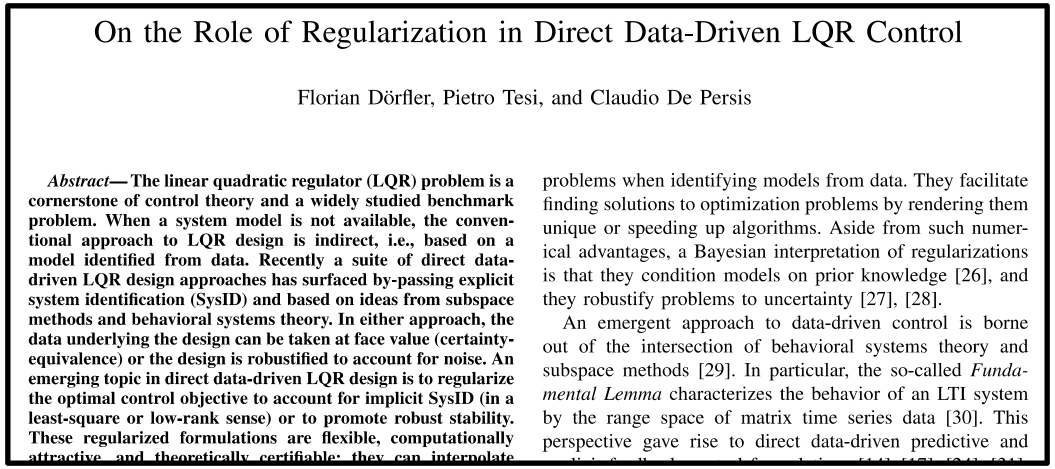
1. **model-based pipeline with model-free elements**
   → data-driven parametrization & robustifying regularization

2. **model-free pipeline with model-based elements**
   → adaptive method: policy gradient & sample covariance

3. **case studies**: academic & power systems/electronics
   → LQR is academic example but can be made useful

Contents

1. **regularizations** bridging direct & indirect data-driven LQR
   → story of a *model-based pipeline with model-free elements*



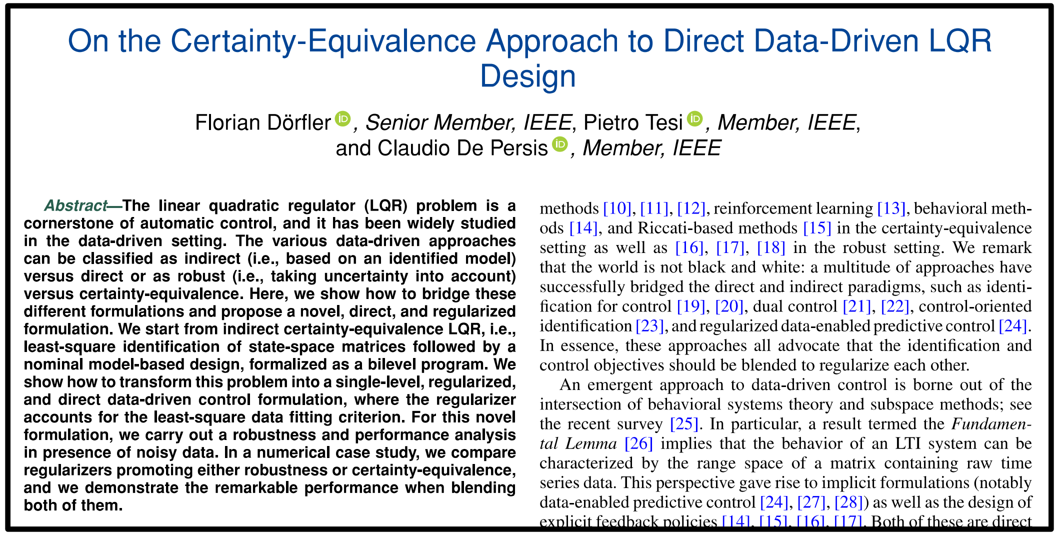
On the Role of Regularization in Direct Data-Driven LQR Control

Florian Dörfler, Pietro Tesi, and Claudio De Persis

***Abstract*— The linear quadratic regulator (LQR) problem is a cornerstone of control theory and a widely studied benchmark problem. When a system model is not available, the conventional approach to LQR design is indirect, i.e., based on a model identified from data. Recently a suite of direct data-driven LQR design approaches has surfaced by-passing explicit system identification (SysID) and based on ideas from subspace methods and behavioral systems theory. In either approach, the data underlying the design can be taken at face value (certainty-equivalence) or the design is robustified to account for noise. An emerging topic in direct data-driven LQR design is to regularize the optimal control objective to account for implicit SysID (in a least-square or low-rank sense) or to promote robust stability. These regularized formulations are flexible, computationally attractive, and theoretically certifiable; they can interpolate**

problems when identifying models from data. They facilitate finding solutions to optimization problems by rendering them unique or speeding up algorithms. Aside from such numerical advantages, a Bayesian interpretation of regularizations is that they condition models on prior knowledge [26], and they robustify problems to uncertainty [27], [28].

An emergent approach to data-driven control is borne out of the intersection of behavioral systems theory and subspace methods [29]. In particular, the so-called *Fundamental Lemma* characterizes the behavior of an LTI system by the range space of matrix time series data [30]. This perspective gave rise to direct data-driven predictive and



On the Certainty-Equivalence Approach to Direct Data-Driven LQR Design

Florian Dörfler, *Senior Member, IEEE*, Pietro Tesi, *Member, IEEE*, and Claudio De Persis, *Member, IEEE*

***Abstract*—The linear quadratic regulator (LQR) problem is a cornerstone of automatic control, and it has been widely studied in the data-driven setting. The various data-driven approaches can be classified as indirect (i.e., based on an identified model) versus direct or as robust (i.e., taking uncertainty into account) versus certainty-equivalence. Here, we show how to bridge these different formulations and propose a novel, direct, and regularized formulation. We start from indirect certainty-equivalence LQR, i.e., least-square identification of state-space matrices followed by a nominal model-based design, formalized as a bilevel program. We show how to transform this problem into a single-level, regularized, and direct data-driven control formulation, where the regularizer accounts for the least-square data fitting criterion. For this novel formulation, we carry out a robustness and performance analysis in presence of noisy data. In a numerical case study, we compare regularizers promoting either robustness or certainty-equivalence, and we demonstrate the remarkable performance when blending both of them.**

methods [10], [11], [12], reinforcement learning [13], behavioral methods [14], and Riccati-based methods [15] in the certainty-equivalence setting as well as [16], [17], [18] in the robust setting. We remark that the world is not black and white: a multitude of approaches have successfully bridged the direct and indirect paradigms, such as identification for control [19], [20], dual control [21], [22], control-oriented identification [23], and regularized data-enabled predictive control [24]. In essence, these approaches all advocate that the identification and control objectives should be blended to regularize each other.

An emergent approach to data-driven control is borne out of the intersection of behavioral systems theory and subspace methods; see the recent survey [25]. In particular, a result termed the *Fundamental Lemma* [26] implies that the behavior of an LTI system can be characterized by the range space of a matrix containing raw time series data. This perspective gave rise to implicit formulations (notably data-enabled predictive control [24], [27], [28]) as well as the design of

with Pietro Tesi (Florence) &

Claudio de Persis (Groningen)

Indirect & certainty-equivalence LQR

- collect **I/O data** (X_0, U_0, X_1) with D_0 unknown & PE: rank $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix} = n + m$

$$U_0 := \begin{bmatrix} u(0) & u(1) & \dots & u(T-1) \end{bmatrix}$$
$$D_0 := \begin{bmatrix} d(0) & d(1) & \dots & d(T-1) \end{bmatrix}$$
$$\longrightarrow \boxed{X_1 = AX_0 + BU_0 + D_0} \longrightarrow$$
$$X_0 := \begin{bmatrix} x(0) & x(1) & \dots & x(T-1) \end{bmatrix}$$
$$X_1 := \begin{bmatrix} x(1) & x(2) & \dots & x(T) \end{bmatrix}$$

- **indirect & certainty-equivalence LQR** (optimal in MLE setting)

$$\begin{aligned} \underset{P \succeq I, K}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top R K P\right) \\ \text{subject to} \quad & (\hat{A} + \hat{B}K)P(\hat{A} + \hat{B}K)^\top - P + I \preceq 0 \end{aligned}$$

certainty-equivalent LQR

$$\begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} = \arg\min_{B,A} \left\| X_1 - \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \right\|_F$$

least squares SysID

Recall indirect approach on the board

- **I/O data** (X_0, U_0, X_1) with D_0 unknown & PE: rank $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix} = n + m$

$$U_0 := \begin{bmatrix} u(0) & u(1) & \dots & u(T-1) \end{bmatrix} \qquad X_0 := \begin{bmatrix} x(0) & x(1) & \dots & x(T-1) \end{bmatrix}$$

$$D_0 := \begin{bmatrix} d(0) & d(1) & \dots & d(T-1) \end{bmatrix} \qquad X_1 := \begin{bmatrix} x(1) & x(2) & \dots & x(T) \end{bmatrix}$$

$$\boxed{X_1 = AX_0 + BU_0 + D_0}$$

$$\begin{aligned} \underset{P \succeq I, K}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top R K P\right) \\ \text{subject to} \quad & (\hat{A} + \hat{B}K)P(\hat{A} + \hat{B}K)^\top - P + I \preceq 0 \end{aligned}$$

$$\begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} = \underset{B,A}{\arg\min} \left\| X_1 - \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \right\|_F^2$$

$$= X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger$$ ← Moore-Penrose (right) inverse

$$= X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \left(\begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \right)^{-1}$$

invertible due to PE

$$\frac{\partial}{\partial(\cdot)} = 0: \quad 0 = 2\left(X_1 - \begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}\right) \cdot \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top$$

$$\rightsquigarrow \begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} = X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \left(\begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \right)^{-1}$$

are uniquely determined

$\rightsquigarrow$ model-based design

Derivation of a direct approach on the board

- **I/O data** (X_0, U_0, X_1) with D_0 unknown & PE: rank $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix} = n + m$

$$U_0 := \begin{bmatrix} u(0) & u(1) & \dots & u(T-1) \end{bmatrix} \qquad D_0 := \begin{bmatrix} d(0) & d(1) & \dots & d(T-1) \end{bmatrix} \longrightarrow \boxed{X_1 = AX_0 + BU_0 + D_0} \longrightarrow X_0 := \begin{bmatrix} x(0) & x(1) & \dots & x(T-1) \end{bmatrix} \qquad X_1 := \begin{bmatrix} x(1) & x(2) & \dots & x(T) \end{bmatrix}$$

- PE implies that $\forall K \; \exists G$ so that $\begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G$

- subspace relations for closed-loop matrix

$$A + BK = \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G = \begin{bmatrix} A X_0 + B U_0 \end{bmatrix} G = (X_1 - D_0) G$$

⇝ can replace $A + BK$ in any LMI by $(X_1 - D_0) G$

⇝ data driven parameterization of linear control design

Direct approach from subspace relations in data

- **PE data:** rank $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix} = n + m \;\Rightarrow\; \forall\, K \;\exists\, G$ s.t. $\boxed{\begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G}$

$$U_0 := \begin{bmatrix} u(0) & u(1) & \dots & u(T-1) \end{bmatrix} \qquad X_0 := \begin{bmatrix} x(0) & x(1) & \dots & x(T-1) \end{bmatrix}$$

$$D_0 := \begin{bmatrix} d(0) & d(1) & \dots & d(T-1) \end{bmatrix} \qquad X_1 := \begin{bmatrix} x(1) & x(2) & \dots & x(T) \end{bmatrix}$$

$$\boxed{X_1 = AX_0 + BU_0 + D_0}$$

- **subspace relations**

$$A + BK = \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} K \\ I \end{bmatrix} \boxed{=} \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G \boxed{=} (X_1 - D_0) G$$

- **data-driven LQR** LMIs by substituting $A + BK = (X_1 - D_0)G$

  → certainty equivalence by neglecting noise D_0: $\boxed{A + BK = X_1 G}$

Indirect vs direct

Indirect

$$\begin{aligned} \underset{P \succeq I, K}{\text{minimize}} \quad & \text{trace}\,(QP) + \text{trace}\left(K^\top R K P\right) \\ \text{subject to} \quad & (\hat{A} + \hat{B}K)P(\hat{A} + \hat{B}K)^\top - P + I \preceq 0 \\ & \begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} = \underset{B,A}{\arg\min} \left\| X_1 - \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \right\|_F \end{aligned}$$

$= X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger$

$\rightsquigarrow$ $\underset{P, K}{\text{minimize}}$ $\text{trace}\,(QP) + \text{trace}\,(K^\top R K P)$

s.t. $\underbrace{\underbrace{X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger}_{\begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix}} \begin{bmatrix} K \\ I \end{bmatrix}}_{\hat{A} + \hat{B}K} \cdot P \cdot (\ldots)^\top - P + I \leq 0$

direct

$$\begin{aligned} \underset{P \succeq I, K, G}{\text{minimize}} \quad & \text{trace}\,(QP) + \text{trace}\left(K^\top R K P\right) \\ \text{subject to} \quad & X_1 G P G^\top X_1^\top - P + I \preceq 0 \\ & \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G \end{aligned}$$

- issue $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}$ has a big null space
  $\rightsquigarrow$ solution G is not unique
- pick least norm solution satisfying
  $\underbrace{\left(I - \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}\right)}_{=: \Pi \text{ projects orthogonal to null space}} G = 0$

Equivalence: direct + *xxx* ⇔ indirect

- **direct** approach

  → optimizer has nullspace $\ker \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}$

  → orthogonality constraint

$$\begin{aligned}
\underset{P \succeq I, K, G}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top R K P\right) \\
\text{subject to} \quad & X_1 G P G^\top X_1^\top - P + I \preceq 0 \\
& \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G \\
& \left(I - \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \right) G = 0
\end{aligned}$$

$$G = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger \begin{bmatrix} K \\ I \end{bmatrix}$$

equivalent constraints:

$$\left(X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger \begin{bmatrix} K \\ I \end{bmatrix} \right) P \left(\ldots \right)^\top - P + I \preceq 0$$

- **indirect** approach

$$\begin{aligned}
\underset{P \succeq I, K}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top R K P\right) \\
\text{subject to} \quad & (\hat{A} + \hat{B}K) P (\hat{A} + \hat{B}K)^\top - P + I \preceq 0 \\
& \begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} = \arg\min_{B,A} \left\| X_1 - \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \right\|_F
\end{aligned}$$

$$\begin{bmatrix} \hat{B} & \hat{A} \end{bmatrix} = X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger$$

Convex reformulation of the control design problem

$$\begin{aligned}
\underset{P \succeq I, K, G}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top R K P\right) \\
\text{subject to} \quad & X_1 G P G^\top X_1^\top - P + I \preceq 0 \\
& \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G \\
& \Pi G = 0
\end{aligned}$$

$\hookrightarrow \operatorname{trace}(R^{1/2} K P K^\top R^{1/2})$

(2) can be pushed to constraint via epigraph formulation

(1) $K = U_0 G$ can be eliminated

(4) remove $P = X_0 Y$

(5) interpret $\geq$ & $\leq$ as Schur complements

e.g.,

$\begin{bmatrix} a & b \\ b^\top & d \end{bmatrix} \geq 0$

$\Leftrightarrow d - b^\top a^{-1} b \geq 0$ if $a > 0$

(3) substitute $Y = GP$ or $G = Y \cdot P^{-1}$ $\rightsquigarrow$ $K = U_0 G = U_0 Y P^{-1}$

$$\begin{aligned}
\underset{Y, X, P}{\text{minimize}} \quad & \operatorname{trace}(QP) + \operatorname{trace}(X) \\
& X_1 Y P^{-1} Y^\top X_1 - P + I \leq 0 \\
& X - R^{1/2} U_0 Y \underbrace{P^{-1} P P^{-1}}_{} U_0 Y^\top U_0^\top R^{1/2} \geq 0 \\
& I = X_0 G = X_0 Y \overset{\llcorner P^{-1}}{P^{-1}} \Leftrightarrow P = X_0 Y \\
& \Pi G = 0
\end{aligned}$$

Regularized, direct, & certainty-equivalent LQR

- orthogonality constraint

$$\Pi = I - \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}$$

lifted to regularizer

$$\begin{aligned} \underset{P \succeq I, K, G}{\text{minimize}} \quad & \text{trace}\,(QP) + \text{trace}\left(K^\top RKP\right) + \lambda \cdot \|\Pi G\| \\ \text{subject to} \quad & X_1 G P G^\top X_1^\top - P + I \preceq 0 \\ & \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G \end{aligned}$$

- **equivalent** to indirect certainty-equivalent LQR design for λ suff. large
- λ **interpolates** between direct & indirect approaches
- **multi-criteria interpretation**: λ interpolates control & SysID objectives
- however, certainty-equivalence formulation may not be **robust (?)**

Robustness-promoting regularization

- **effect of noise** entering data: $A + BK = (X_1 - D_0)G$
  Lyapunov constraint $X_1 GPG^\top X_1^\top - P + I \preceq 0$
  becomes $(X_1 - D_0)GPG^\top(X_1 - D_0)^\top - P + I \preceq 0$

for robustness $GPG^\top$ should be small

- previous **certainty-equivalence regularizer** $\|\Pi G\|$ achieves small $\|G\|$

- **robustness-promoting regularizer** [de Persis & Tesi, '21]

$$
\begin{aligned}
\underset{P \succeq I, K, G}{\text{minimize}} \quad & \text{trace}\,(QP) + \text{trace}\left(K^\top R K P\right) \\
& + \rho \cdot \text{trace}\left(GPG^\top\right) \\
\text{subject to} \quad & X_1 GPG^\top X_1^\top - P + I \preceq 0 \\
& \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G
\end{aligned}
$$

Performance & robustness analysis

- **SNR** (signal-to-noise-ratio) $\dfrac{\sigma_{min}([X_0\ U_0])}{\sigma_{max}(D_0)}$

- **relative performance** metric

realized cost from regularized design with λ & ρ

if exact system matrices *A* and *B* were known

$$\frac{\{\text{regularized data-driven LQR performance}\} - \{\text{ground-truth performance}\}}{\{\text{ground-truth performance}\}}$$

certificate: optimal control problem is **always feasible & stabilizing** for suff. large SNR & **relative performance** $\sim \mathcal{O}\left(\text{SNR}^{-1}\right) + \text{const.} \cdot \rho$ robust reg.

proof bounds Lyapunov constraint $(X_1 - D_0)GPG^\top(X_1 - D_0)^\top - P + I \preceq 0$

FYI: another regularization promoting low-rank

- de-noising of data-matrices via **low-rank approximation**

$$\begin{aligned}
\underset{\hat{U}_0,\hat{X}_0,\hat{X}_1}{\text{minimize}} \quad & \left\| \begin{bmatrix} \hat{U}_0 \\ \hat{X}_0 \\ \hat{X}_1 \end{bmatrix} - \begin{bmatrix} U_0 \\ X_0 \\ X_1 \end{bmatrix} \right\| \\
\text{subject to} \quad & \operatorname{rank} \begin{bmatrix} \hat{U}_0 \\ \hat{X}_0 \\ \hat{X}_1 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \hat{U}_0 \\ \hat{X}_0 \end{bmatrix} = n + m
\end{aligned}$$

Let PE hold: $\operatorname{rank} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} = n+m$

The following are equivalent:

(i) $\operatorname{rank} \begin{bmatrix} U_0 \\ X_0 \\ X_1 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} = n+m$

(ii) $\exists$ unique B & A so that $X_1 = AX_0 + BU_0$

Proof: (ii) $\Rightarrow$ (i) follows since $X_1 = AX_0 + BU_0$ implies that X_1 is dependent

(i) $\Rightarrow$ (ii): n rows of $\begin{bmatrix} U_0 \\ X_0 \\ X_1 \end{bmatrix}$ are dependent

due to PE, the rows of $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}$ are independent

$\leadsto \exists \begin{bmatrix} B & A \end{bmatrix}$ so that $X_1 = \begin{bmatrix} B & A \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}$

$\leadsto$ uniqueness due to PE

Surrogate for low-rank pre-processing

$$\underset{P,\,K,\,G}{\text{minimize}} \quad \operatorname{trace}(QP) + \operatorname{trace}(K^T R K P)$$

$$A_{cl} P A_{cl}^T - P + I \preceq 0$$

$$\begin{bmatrix} K \\ I \\ 0 \end{bmatrix} = \begin{bmatrix} u_0 \\ x_0 \\ x_1 \end{bmatrix} G \quad \text{as } X_1 G = A + BK = A_{cl}$$

0. new constraint without loss of generality since rank of $\begin{bmatrix} u_0 \\ x_0 \\ x_1 \end{bmatrix} = n+m$ — number of non-zero entries of every column G_i of G is less than $n+m$

1. relax new constraint as $\|G_i\|_1 \leq \alpha_i$ for suitable α_i
2. relax as $\|G\|_1 \leq \max \alpha_i$
3. lift to cost function as a penalty $\|G\|_1$

l_1 regularization as low-rank surrogate

- de-noising of data-matrices via **low-rank approximation** (low rank is equivalent to uniqueness of (A, B) matrices)

$$\begin{aligned}
&\underset{\hat{U}_0,\hat{X}_0,\hat{X}_1}{\text{minimize}} && \left\| \begin{bmatrix} \hat{U}_0 \\ \hat{X}_0 \\ \hat{X}_1 \end{bmatrix} - \begin{bmatrix} U_0 \\ X_0 \\ X_1 \end{bmatrix} \right\| \\
&\text{subject to} && \text{rank} \begin{bmatrix} \hat{U}_0 \\ \hat{X}_0 \\ \hat{X}_1 \end{bmatrix} = \text{rank} \begin{bmatrix} \hat{U}_0 \\ \hat{X}_0 \end{bmatrix} = n + m
\end{aligned}$$

- ℓ_1 **regularizer** as surrogate of pre-processing by low-rank approximation: bias solution G towards sparsity $\leadsto$ low-rank

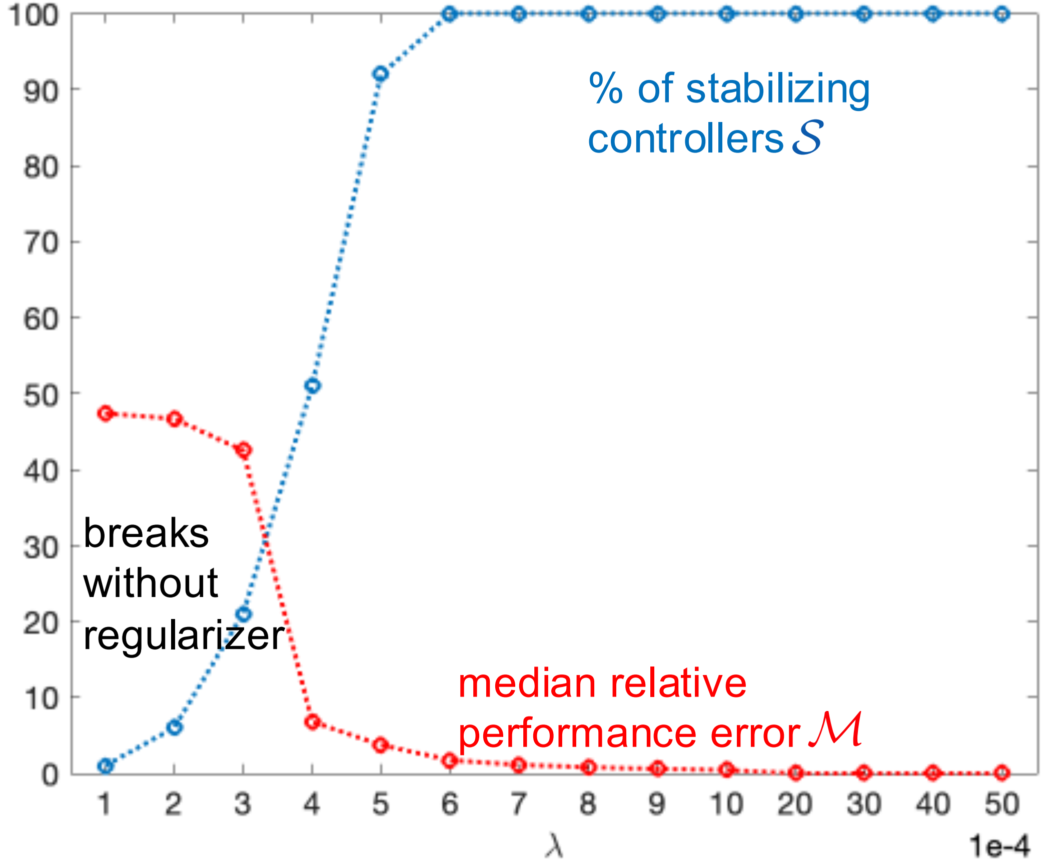
$$\begin{aligned}
&\underset{K,\, P \succeq I,\, G}{\text{minimize}} && \text{trace}\,(QP) + \text{trace}\left(K^\top R K P\right) + \lambda \|G\|_1 \\
&\text{subject to} && X_1 G P G^\top X_1^\top - P + I \preceq 0 \\
& && \begin{bmatrix} K \\ I \end{bmatrix} = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G
\end{aligned}$$

Numerical case study

- **case study** [Dean et al. '19]: discrete-time marginally unstable Laplacian system subject to noise of variance $\sigma^2 = 0.01$

$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B = I$$

- **take-home message** 1:
  regularization is needed !
  prior work without regularizer has no robustness margin



Numerical case study cont'd

- **take-home message 2**: different regularizers promote different features: robustness vs. certainty-equivalence (performance)

	$\sigma = 0.01$ (SNR > 15dB)	$\sigma = 0.1$ (SNR $\in [5, 10]$dB)	$\sigma = 0.3$ (SNR $\in [0, 5]$dB)	$\sigma = 0.7$ (SNR ≈ 0dB)	$\sigma = 1$ (SNR < -5dB)
Certainty-equivalence ($\lambda = 1, \rho = 0$)	$\mathcal{S} = 100\%$ $\mathcal{M} = 2.5599e\text{-}05$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0026$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0237$	$\mathcal{S} = 97\%$ $\mathcal{M} = 0.1366$	$\mathcal{S} = 84\%$ $\mathcal{M} = 0.2596$
Robust approach ($\lambda = 0, \rho = 1$)	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0035$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0074$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0369$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.2350$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.6270$

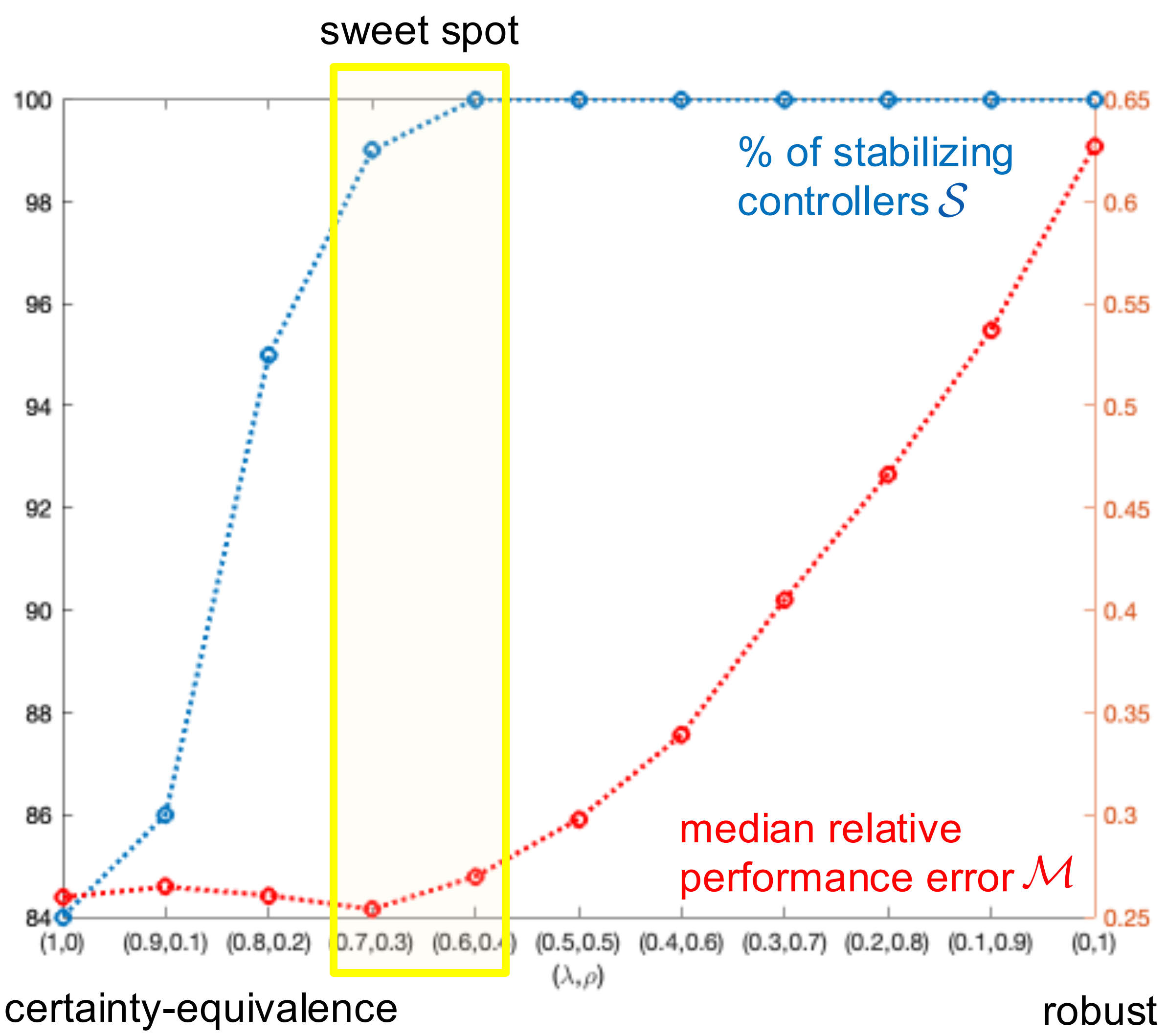
- **take-home message 3**: mixed regularization achieves best of both

Mixed regularization ($\lambda = \rho = 0.5$)	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0010$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0035$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.0235$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.1262$	$\mathcal{S} = 100\%$ $\mathcal{M} = 0.2978$

Intermediate conclusions… so far

- **interpolation** of different regularizers with high noise: $\sigma^2 = 1$ (SNR< -5db)
- **flexible multi-criteria formulation** trading off different objectives by regularizers (best of all is attainable)
- **classification direct vs. indirect** is less relevant: λ interpolates

$\rightarrow$ works… but lame: **learning is offline**



Contents

2. **data-enabled policy optimization** for online adaptation
→ story of a *model-free pipeline with model-based elements*

Data-enabled Policy Optimization for the Linear Quadratic Regulator

Feiran Zhao, Florian Dörfler, Keyou You

Abstract— **Policy optimization (PO), an essential approach of reinforcement learning for a broad range of system classes, requires significantly more system data than indirect (identification-followed-by-control) methods or behavioral-based direct methods even in the simplest linear quadratic regulator (LQR) problem. In this paper, we take an initial step towards bridging this gap by proposing the data-enabled policy optimization (DeePO) method, which requires only a finite number of sufficiently exciting data to iteratively solve the LQR problem via PO. Based on a data-driven closed-loop parameterization, we are able to directly compute the**

a considerable gap in the sample complexity between PO and indirect methods, which have proved themselves to be more sample-efficient [9], [10] for solving the LQR problem. This gap is due to the exploration or trial-and-error nature of RL, or more specifically, that the cost used for gradient estimate can only be evaluated *after* a whole trajectory is observed. Thus, the existing PO methods require numerous system trajectories to find an optimal policy, even in the simplest LQR setting.

with Alessandro Chiuso (Padova),

Feiran Zhao, Keyou You (Tsinghua),

& Linbin Huang (Zhezjiang)

Data-Enabled Policy Optimization for Direct Adaptive Learning of the LQR

Feiran Zhao, Florian Dörfler, Alessandro Chiuso, Keyou You

Abstract—**Direct data-driven design methods for the linear quadratic regulator (LQR) mainly use offline or episodic data batches, and their online adaptation has been acknowledged as an open problem. In this paper, we propose a direct adaptive method to learn the LQR from online closed-loop data. First, we propose a new policy parameterization based on the sample covariance to formulate a direct data-driven LQR problem, which is shown to be equivalent to the certainty-equivalence LQR with optimal non-asymptotic guarantees. Second, we design a novel data-enabled policy optimization (DeePO) method to directly update the policy, where the gradient is explicitly computed using only a batch of persistently exciting (PE) data. Third, we establish its global convergence via a projected gradient dominance property. Importantly, we efficiently use DeePO to adaptively learn the LQR by performing only one-step projected gradient descent per sample of the closed-loop system, which also leads to an explicit recursive update of the policy. Under PE inputs and for bounded noise, we show that the average regret of the LQR cost is upper-bounded by two terms signifying a sublinear decrease in time $\mathcal{O}(1/\sqrt{T})$ plus a bias scaling inversely with signal-to-**

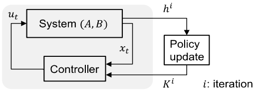


Fig. 1. An illustration of episodic approaches, where $h^i = (x_0, u_0, \ldots, x_{T^i})$ denotes the trajectory of the i-th episode.

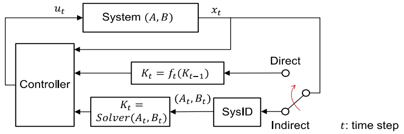


Fig. 2. An illustration of indirect and direct adaptive approaches in closed-loop, where f_t is some explicit function.

Online & adaptive solutions

- **shortcoming** of separating offline learning & online control
  → cannot improve policy **online** & cheaply / rapidly **adapt** to changes

Adaptive Control:
Towards a Complexity-Based General Theory*

G. ZAMES

“adaptive = improve over best control with a priori info”

- (elitist) **desired adaptive** solution: direct, online (non-episodic/non-batch) algorithms, with closed-loop data, & recursive algorithmic implementation

- “best” way to improve policy with new data → **go down the gradient !**

---

* disclaimer: a large part of the adaptive control community focuses on stability & not optimality

Ingredient 1: policy gradient methods

Annual Review of Control, Robotics, and Autonomous Systems

Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies

Bin Hu,[1] Kaiqing Zhang,[2,3] Na Li,[4] Mehran Mesbahi,[5] Maryam Fazel,[6] and Tamer Başar[1]

- LQR viewed as smooth program (many formulations)

$$\left.\begin{array}{ll} \underset{P \succeq I, K}{\text{minimize}} & \operatorname{trace}(QP) + \operatorname{trace}\left(K^\top RKP\right) \\ \text{subject to} & (A+BK)P(A+BK)^\top - P + I \preceq 0 \end{array}\right\}$$

after eliminating (unique) P**, denote this as** $J(K)$

- $J(K)$ is not convex …

  but on the set of stabilizing gains K, it's
  - coercive with compact sublevel sets,
  - smooth with bounded Hessian, &
  - degree-2 gradient dominated

$$J(K) - J^* \leq const. \cdot \|\nabla J(K)\|^2$$

Fact: policy gradient descent

$$K^+ = K - \eta\, \nabla J(K)$$

initialized from a stabilizing policy converges linearly to K^*.

Insights into the proof

- $J(K)$ is smooth with $\|\nabla^2 J(K)\| \leq L$: By Taylor or mean-value theorem:

$$J(K') \leq J(K) + \nabla J(K)^T (K' - K) + \frac{L}{2} \|K' - K\|_F^2 \quad (1)$$

- gradient dominance : $J(K) - J(K^*) \leq \frac{1}{2\mu} \|\nabla J(K)\|_F^2 \quad (2)$

- gradient descent : $K^+ = K - \eta \nabla J(K)$

$$\leadsto J(K^+) = J(K - \eta \nabla J(K)) \overset{(1)}{\leq} J(K) + \nabla J(K)^T (\underbrace{K - \eta \nabla J(K) - K}_{-\eta \nabla J(K)}) + \frac{L}{2}\eta^2 \|\nabla J(K)\|^2$$

$$= J(K) - \left(\eta - \frac{L\eta^2}{2}\right) \|\nabla J(K)\|^2$$

$$\overset{(2)}{\leq} J(K) - \left(\eta - \frac{L\eta^2}{2}\right) 2\mu \left(J(K) - J(K^*)\right) \Rightarrow \boxed{J(K^+) - J(K^*) \leq \left(1 - \eta - \frac{L\eta^2}{2}\right) 2\mu \left(J(K) - J(K^*)\right)}$$

Explicit formulae for model-based gradient

- For these results we need the equivalent LQR formulations (see beginning)

$J(K) = tr(PQ) + tr(K^T R K P)$ where $P \succ 0$ solves $(A+BK)P(A+BK)^T - P + X = 0$

$= tr(WX)$ where $W \succ 0$ solves $(A+BK)^T W (A+BK) - W + Q + K^T R K = 0$

where $X = x_0 x_0^T$ is the initial state covariance, though its particular value is irrelevant

- To calculate the gradient, we recognize $\nabla J(K) = \frac{\partial}{\partial K} tr(W(K) \cdot X)$

$= \left[\ddots \; tr\left(\frac{\partial W}{\partial K_{ij}} \cdot X\right) \ddots \right]$

... as you can see, the math for such derivatives can get cumbersome. For these reasons, we will work with differentials which will simplify the derivations.

The differential dx is the linear part (Jacobian) of the function $f(x+dx) - f(x)$

- mathematical preliminaries on differentials
  - $d\,Tr(A) = Tr(dA)$
  - $d(A \cdot B) = dA \cdot B + dB \cdot A$
  - $d(A^T) = dA^T$
  - Let J be a function of x. If $dJ = Tr(C \cdot dx)$, then $\nabla_x J = C^T$

this one is constant with zero differential

- Derivation of $\nabla_K J^{(K)}$: ① Since $dJ = d\,tr(W \cdot X) = tr(dW \cdot X)$

② to obtain dW, we evaluate :

$$\rightsquigarrow (A+BK)^T \partial W (A+BK) - \partial W + \overbrace{\partial K^T \left[(A+BK)^T W B + K^T R \right]}^{M} + \overbrace{(\dots \text{same term} \dots)^T}^{M^T}$$

$\rightsquigarrow$ this is a Lyapunov equation and thus

$$\frac{\partial W}{\partial K} = \sum_{t=0}^{\infty} \left((A+BK)^T \right)^t (M + M^T) (A+BK)^t$$

$\rightsquigarrow$ Hence, $\operatorname{tr}(\delta W \cdot X) = \operatorname{tr}\Big(2 M^T \underbrace{\sum_{t=0}^{\infty} (A+BK)^t X \big((A+BK)^t\big)^T}_{= P \text{ (controllability Gramian)}}\Big)$

$$= \operatorname{tr}\big(\delta K^T 2 \left(B^T W (A+BK) + RK \right) \cdot P\big)$$

$\rightsquigarrow$ Last, using that $dJ = \operatorname{Tr}(C \cdot dx) \Rightarrow \nabla_x J = C^T$, we obtain

$$\boxed{\nabla_K J(K) = 2 \left(B^T W (A+BK) + RK \right) \cdot P}$$

Model-free policy gradient methods

- **model-based setting**: explicit formulae for $\nabla J(K)$ based on closed-loop controllability + observability Gramians [Levine & Athans, '70]

- **model-free 0th order methods** constructing two-point gradient estimate

conceptual for a scalar function: $\nabla f(x) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon}\left(f(x+\varepsilon) - f(x-\varepsilon)\right) = \lim_{\eta \to 0} \mathbb{E}_{\eta \sim \text{uniform in } [-1,1]} \frac{\eta}{2\varepsilon} f(x+\eta\varepsilon)$

⇝ can be approximated sampling function, but scales very poorly for high dimension

from numerous & very long trajectories → extremely sample inefficient

relative performance gap	$\epsilon = 1$	$\epsilon = 0.1$	$\epsilon = 0.01$
# trajectories (100 samples)	1414	43850	142865

~ 10^7 samples

- IMO: policy gradient is a **potentially great** candidate for direct adaptive control **but sadly useless in practice**: sample-inefficient, episodic, …

Ingredient 2: sample covariance parameterization

$$U_0 = \begin{bmatrix} u(0) & u(1) & \cdots & u(t-1) \end{bmatrix} \longrightarrow \boxed{X_1 = AX_0 + BU_0} \begin{array}{l} \longrightarrow X_0 = \begin{bmatrix} x(0) & x(1) & \cdots & x(t-1) \end{bmatrix} \\ \longrightarrow X_1 = \begin{bmatrix} x(1) & x(2) & \cdots & x(t) \end{bmatrix} \end{array}$$

prior parameterization

- PE condition: full row rank $\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}$
- $A + BK = [B \;\; A] \begin{bmatrix} K \\ I \end{bmatrix} = [B \;\; A] \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} G = X_1 G$
- robustness: $G = \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top (\cdot) \leftrightarrow$ regularization
- dimension of all matrices grows with t

covariance parameterization

- sample covariance $\Lambda = \frac{1}{t} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \succ 0$
- $A + BK = [B \; A] \begin{bmatrix} K \\ I \end{bmatrix} = [B \; A] \Lambda V = \frac{1}{t} X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top V$
- robustness for free without regularization
- dimension of all matrices is constant

+ cheap rank-1 updates for online data

Covariance parameterization of the LQR

- state / input **sample covariance** $\Lambda = \frac{1}{t}\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top$ & $\overline{X}_1 = \frac{1}{t} X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top$

- **closed-loop matrix** $A + BK = \overline{X}_1 V$ with $\begin{bmatrix} K \\ \hline I \end{bmatrix} = \Lambda V = \begin{bmatrix} \overline{U}_0 \\ \hline \overline{X}_0 \end{bmatrix} V$

- **LQR covariance parameterization**

  after eliminating K with variable V,
  Lyapunov eqn (explicitly solvable),
  smooth cost $J(V)$ (after removing P),
  & linear parameterization constraint

$$\min_{V, P \succ 0} \; \operatorname{trace}(QP) + \operatorname{trace}\left(V^T \overline{U}_0^T R \overline{U}_0 V P\right)$$
$$\text{s.t. } P = I + \overline{X}_1 V P V^T \overline{X}_1^T, \quad I = \overline{X}_0 V$$

details are not important

Projected policy gradient with sample covariances

- **data-enabled policy optimization (DeePO)**

$$V^+ = V - \eta\, \Pi_{\overline{X}_0}(\nabla J(V))$$

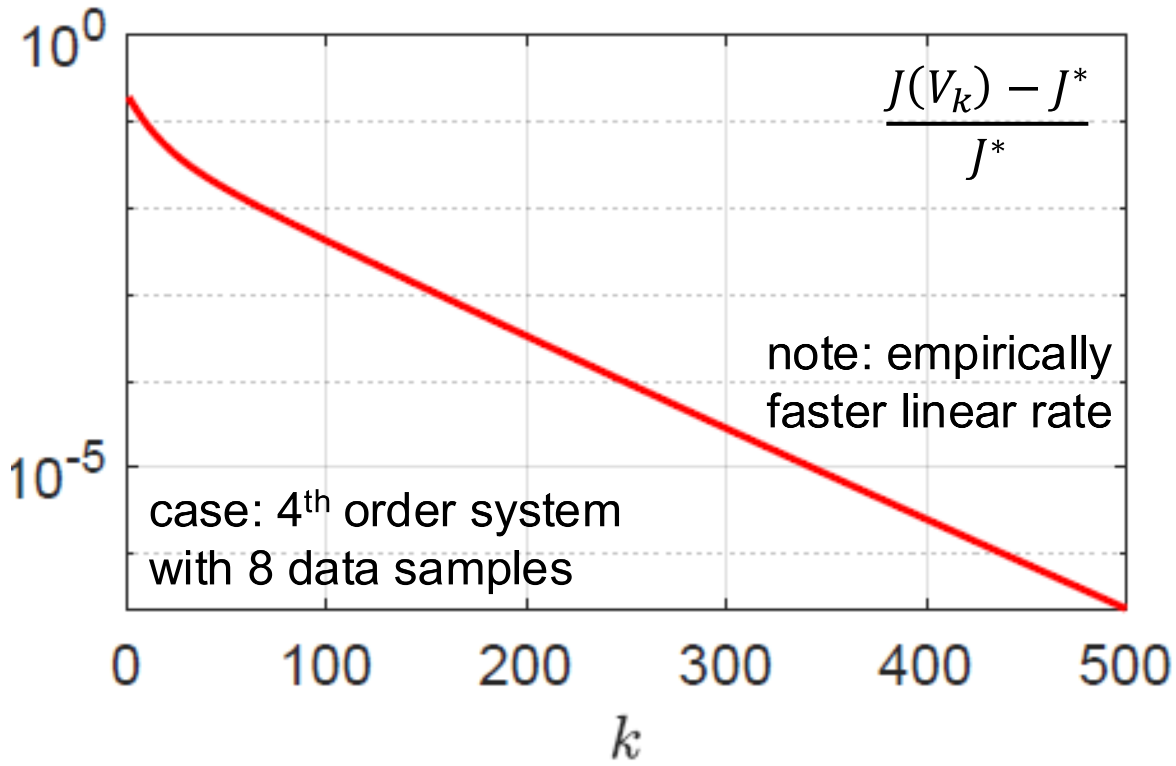
$\Pi_{\overline{X}_0}$ projects on parameterization constraint $I = \overline{X}_0 V$ & gradient $\nabla J(V)$ is computed from two Lyapunov equations with sample covariances

- **optimization landscape**: smooth, degree-1 proj. grad dominance

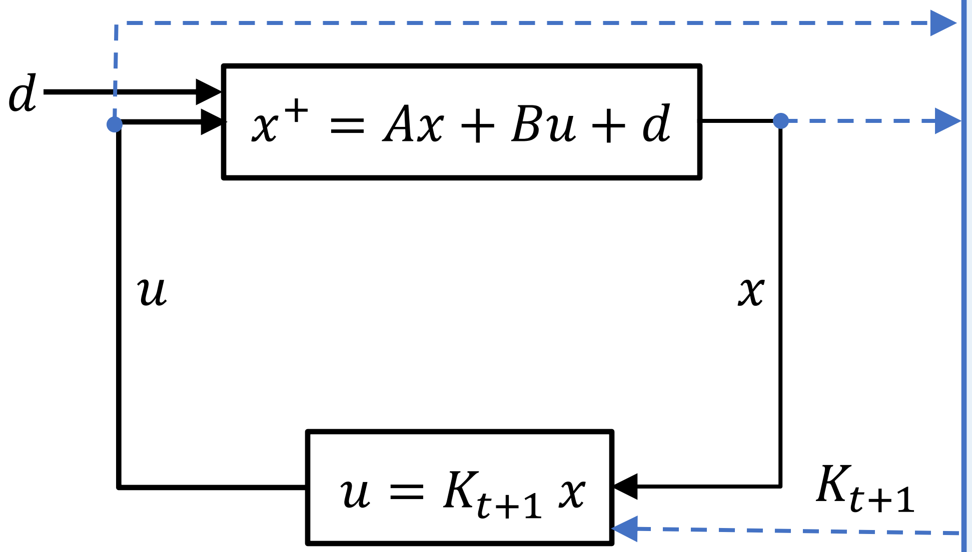
$$J(V) - J^* \leq const. \cdot \left\| \Pi_{\overline{X}_0}(\nabla J(V)) \right\|$$

- warm-up: offline data & no disturbance

Sublinear convergence for feasible initialization $J(V^k) - J^* \leq \mathcal{O}(1/k)$.



Online, adaptive, & closed-loop DeePO



DeePO policy update

Input: $(X_{0,t+1}, U_{0,t+1}, X_{1,t+1})$, K_t

1. update sample covariances: Λ_{t+1} & $\bar{X}_{0,t+1}$
2. update decision variable: $V_{t+1} = \Lambda_{t+1}^{-1} \begin{bmatrix} K_t \\ I_n \end{bmatrix}$
3. gradient descent: $V'_{t+1} = V_{t+1} - \eta \Pi_{\bar{X}_{0,t+1}}(\nabla J_{t+1}(V_{t+1}))$
4. update control gain: $K_{t+1} = \bar{U}_{0,t+1} V'_{t+1}$

Output: K_{t+1}

where $X_{0,t+1} = [x(0), x(1), \ldots x(t), x(t+1)]$ & similar for other matrices

- **cheap & recursive implementation:** rank-1 update of (inverse) sample covariances, cheap computation, & no memory needed to store old data

Underlying assumptions for theoretic certificates

- **initially stabilizing controller:** the LQR problem parameterized by offline data $\left(X_{0,t_0}, U_{0,t_0}, X_{1,t_0}\right)$ is feasible with stabilizing gain K_{t_0}.

- **persistency of excitation** due to process noise or probing:

$$\underline{\sigma}\left(\mathcal{H}_{n+1}\left(U_{0,t}\right)\right) \geq \gamma \cdot \sqrt{t} \quad \text{with Hankel matrix } \mathcal{H}_{n+1}\left(U_{0,t}\right)$$

- **bounded noise:** $\|d(t)\| \leq \delta \;\; \forall\, t \;\; \rightarrow$ **signal-to-noise** ratio $SNR := \gamma/\delta$

- **BIBO:** there are $\bar{u}, \bar{x}$ such that $\|u(t)\| \leq \bar{u}$ & $\|x(t)\| \leq \bar{x}$

  (∃ common Lyapunov function ?)

Bounded regret of DeePO in adaptive setting

- **average regret** performance metric $\mathrm{Regret}_T := \frac{1}{T}\sum_{t=t_0}^{t_0+T-1} \left(J(K_t) - J^* \right)$

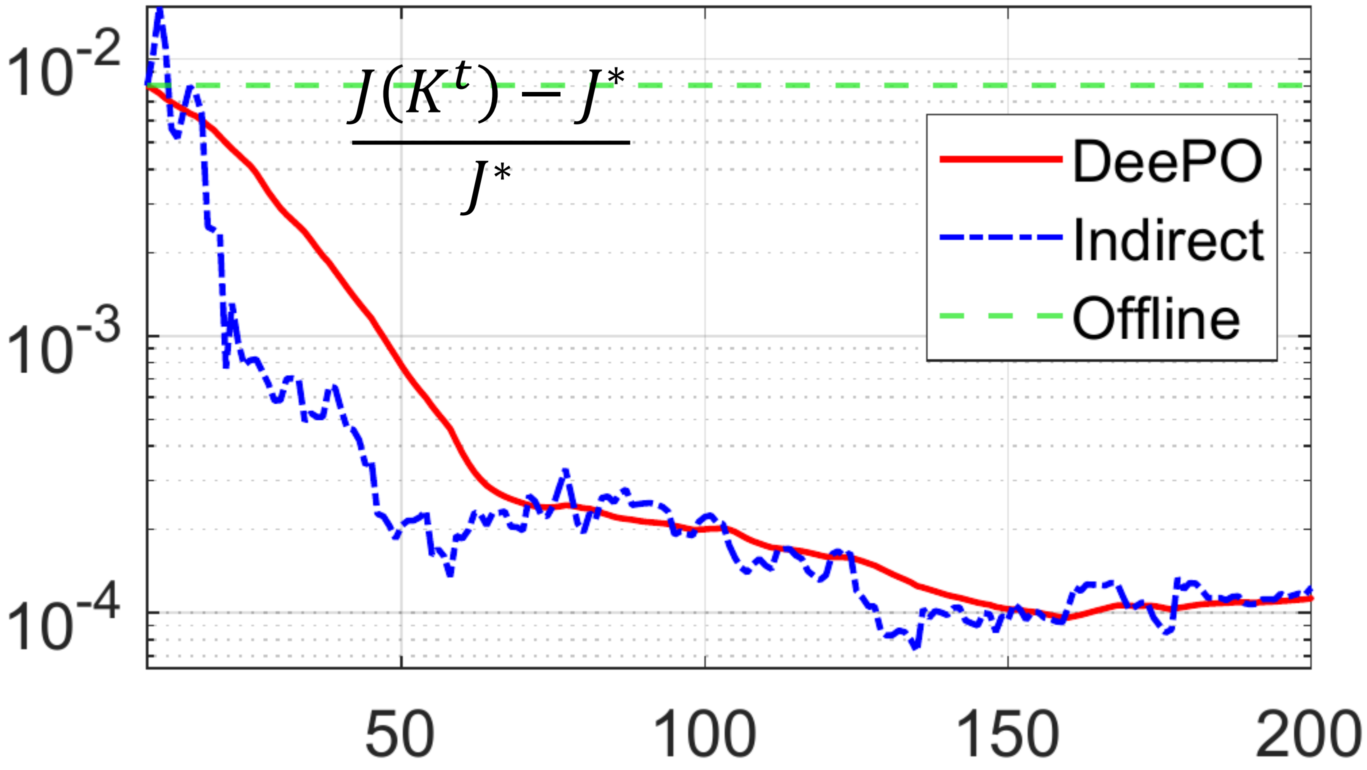
Sublinear regret: Under the assumptions, there are $\nu_1, \nu_2, \nu_3, \nu_4 > 0$ such that for $\eta \in (0, \nu_1]$ & $SNR \geq \nu_2$, it holds that $\{K_t\}$ is stabilizing &

$$\mathrm{Regret}_T \leq \frac{\nu_3}{\sqrt{T}} + \frac{\nu_4}{\sqrt{SNR}} .$$

- **comments** on the qualitatively expected result:
  - analysis is independent of the noise statistics & **consistent** $\mathrm{Regret}_{T\to\infty} \to 0$
  - **favorable sample complexity**: sublinear decrease term matches best rate $\mathcal{O}(1/\sqrt{T})$ of first-order methods in online convex optimization
  - empirically observe smaller **bias term**: $\mathcal{O}(1/SNR^2)$ & not $\mathcal{O}(1/\sqrt{SNR})$

Comparison case studies

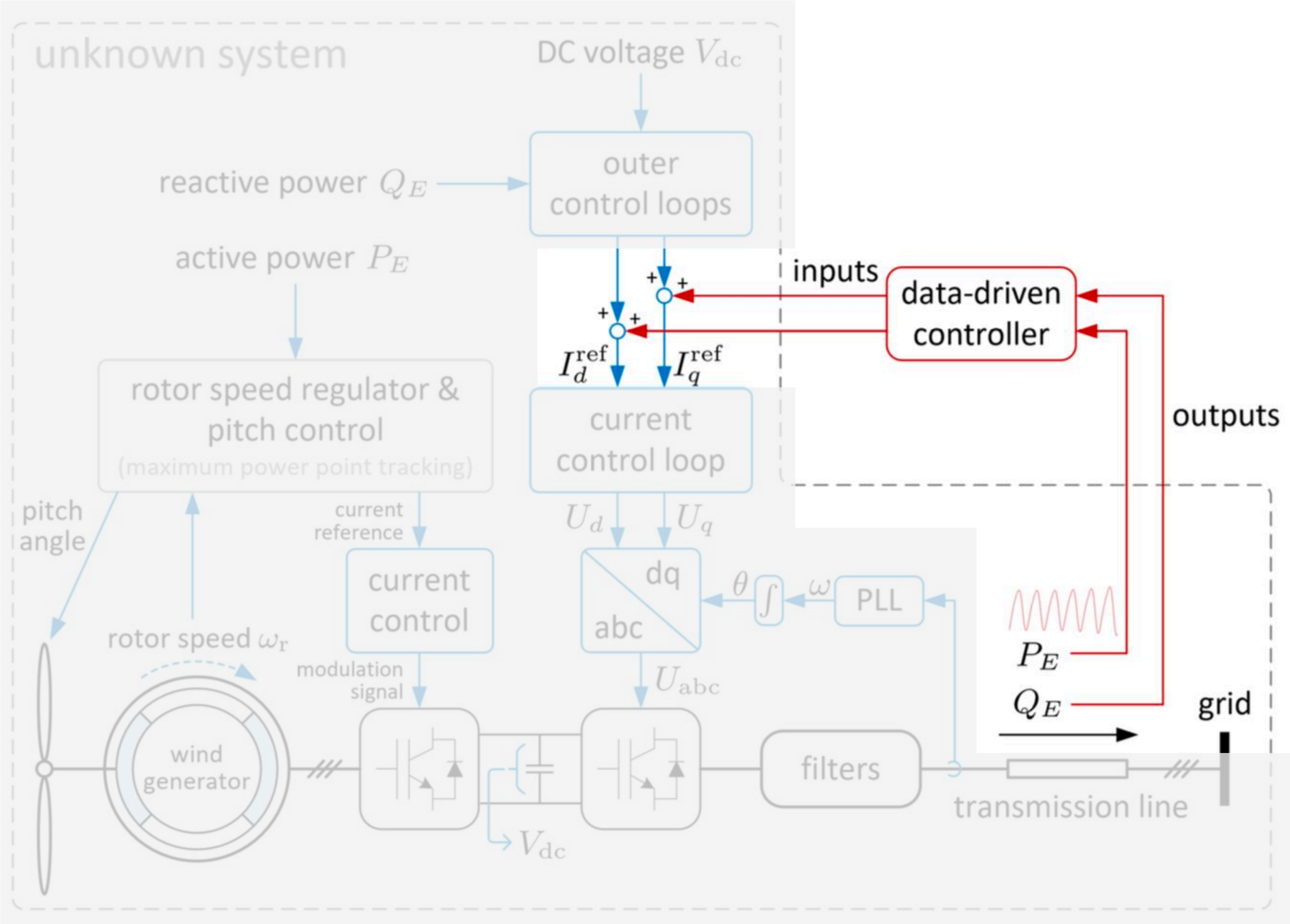
- **same case study** [Dean et al. '19]
- **case 1:** offline LQR
  vs direct adaptive DeePO
  vs indirect adaptive: rls + dlqr
  → **adaptive outperforms offline**
  → direct/indirect **rates matching**
  but **direct is much(!) cheaper**



- **case 2:** adaptive DeePO
  vs 0^{th} order methods
  → **significantly less data**

relative performance gap	$\epsilon = 1$	$\epsilon = 0.1$	$\epsilon = 0.01$
# long trajectories (100 samples) for 0^{th} order LQR	1414	43850	142865
DeePO (# I/O samples)	10	24	48

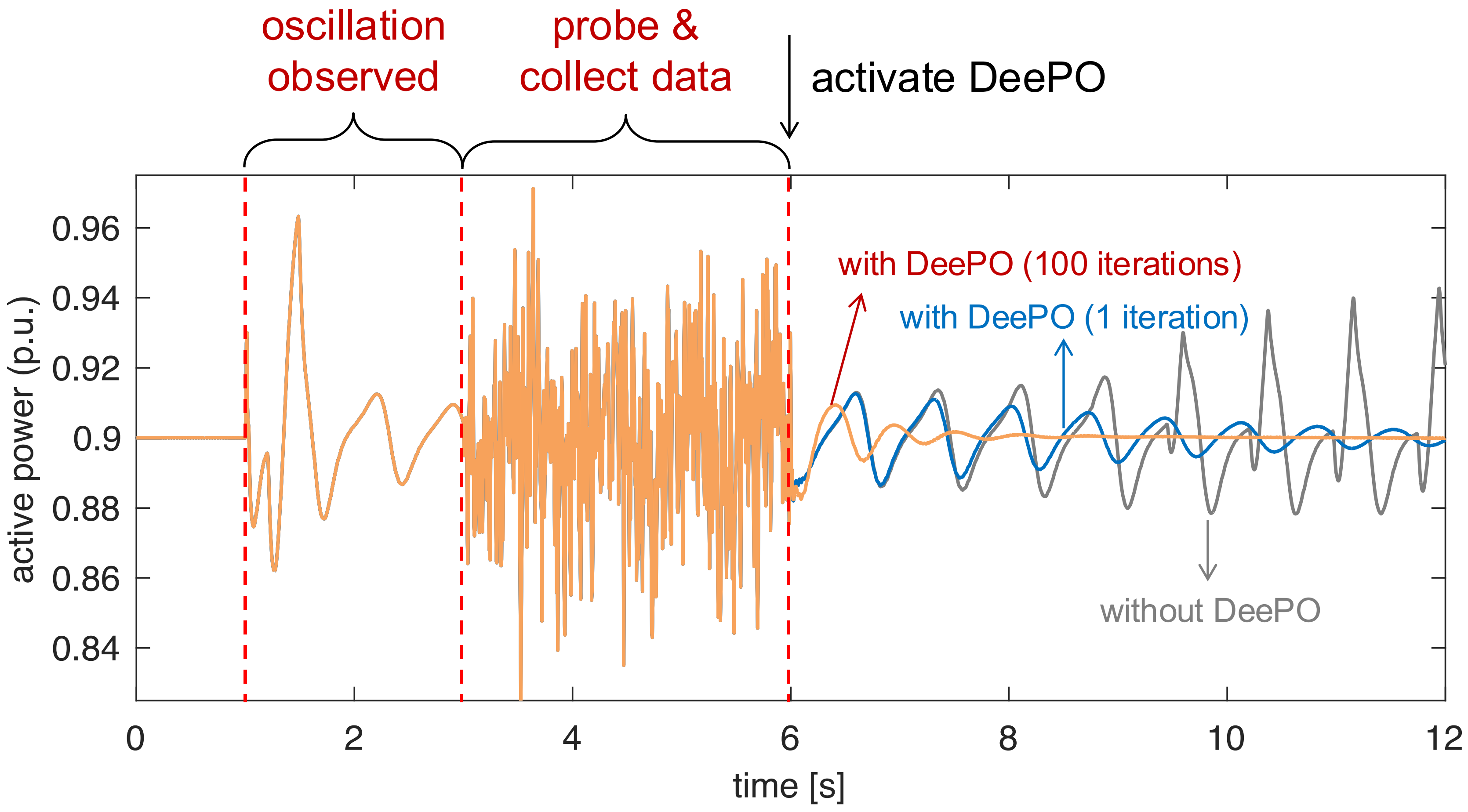
Power systems / electronics case study



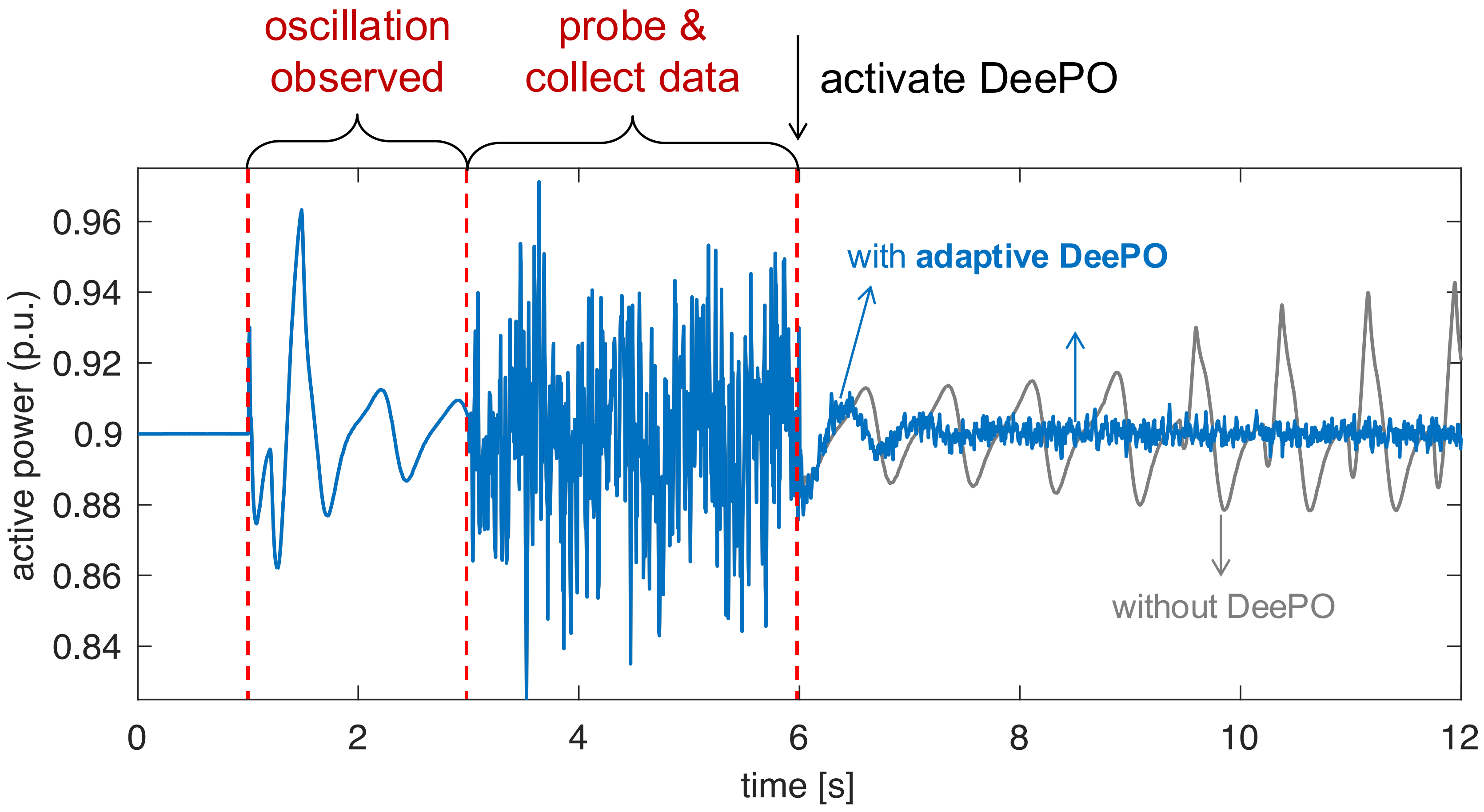
synchronous generator & full-scale converter

- wind turbine becomes **unstable** in weak grids with nonlinear oscillations
- converter, turbine, & grid are a **black box** for the commissioning engineer
- construct state from time shifts (5ms sampling) of $(y(t), u(t))$ & use **DeePO**

Power systems / electronics case study



… same in the adaptive setting with excitation



Conclusions

- **Summary**
  - model-based pipeline with model-free block: data-driven LQR parametrization
    → works well when regularized (note: further flexible regularizations available)
  - model-free pipeline with model-based block: policy gradient & sample covariance
    → DeePO is adaptive, online, with closed-loop data, & recursive implementation
  - academic case studies & can be made useful in power systems/electronics

- **Future work**
  - technicalities: weaken assumptions & improve rates
  - control: based on output feedback & for other objectives
  - further system classes: stochastic, time-varying, & nonlinear
  - open questions: online vs episodic? “best” batch size? triggered?

Papers

1. **model-based pipeline with model-free elements**

On the Role of Regularization in Direct Data-Driven LQR Control

Florian Dörfler, Pietro Tesi, and Claudio De Persis

Abstract— **The linear quadratic regulator (LQR) problem is a cornerstone of control theory and a widely studied benchmark problem. When a system model is not available, the conventional approach to LQR design is indirect, i.e., based on a model identified from data. Recently a suite of direct data-driven LQR design approaches has surfaced by-passing explicit system identification (SysID) and based on ideas from subspace methods and behavioral systems theory. In either approach, the data underlying the design can be taken at face value (certainty-**

problems when identifying models from data. They facilitate finding solutions to optimization problems by rendering them unique or speeding up algorithms. Aside from such numerical advantages, a Bayesian interpretation of regularizations is that they condition models on prior knowledge [26], and they robustify problems to uncertainty [27], [28].

An emergent approach to data-driven control is borne out of the intersection of behavioral systems theory and

2. **model-free pipeline with model-based elements**

Data-enabled Policy Optimization for the Linear Quadratic Regulator

Feiran Zhao, Florian Dörfler, Keyou You

Abstract— **Policy optimization (PO), an essential approach of reinforcement learning for a broad range of system classes, requires significantly more system data than indirect (identification-followed-by-control) methods or behavioral-based direct methods even in the simplest linear quadratic regulator (LQR) problem. In this paper, we take an initial step towards bridging this gap by proposing the data-enabled policy optimization (DeePO) method, which requires only a finite number of sufficiently exciting data to iteratively solve the LQR problem via PO. Based on a data-driven closed-loop parameterization, we are able to directly compute the**

a considerable gap in the sample complexity between PO and indirect methods, which have proved themselves to be more sample-efficient [9], [10] for solving the LQR problem. This gap is due to the exploration or trial-and-error nature of RL, or more specifically, that the cost used for gradient estimate can only be evaluated *after* a whole trajectory is observed. Thus, the existing PO methods require numerous system trajectories to find an optimal policy, even in the simplest LQR setting.

On the Certainty-Equivalence Approach to Direct Data-Driven LQR Design

Florian Dörfler, *Senior Member, IEEE*, Pietro Tesi, *Member, IEEE*, and Claudio De Persis, *Member, IEEE*

Abstract—**The linear quadratic regulator (LQR) problem is a cornerstone of automatic control, and it has been widely studied in the data-driven setting. The various data-driven approaches can be classified as indirect (i.e., based on an identified model) versus direct or as robust (i.e., taking uncertainty into account) versus certainty-equivalence. Here, we show how to bridge these different formulations and propose a novel, direct, and regularized formulation. We start from indirect certainty-equivalence LQR, i.e., least-square identification of state-space matrices followed by a nominal model-based design, formalized as a bilevel program. We show how to transform this problem into a single-level, regularized, and direct data-driven control formulation, where the regularizer accounts for the least-square data fitting criterion. For this novel formulation, we carry out a robustness and performance analysis in presence of noisy data. In a numerical case study, we compare regularizers promoting either robustness or certainty-equivalence, and we demonstrate the remarkable performance when blending both of them.**

methods [10], [11], [12], reinforcement learning [13], behavioral methods [14], and Riccati-based methods [15] in the certainty-equivalence setting as well as [16], [17], [18] in the robust setting. We remark that the world is not black and white: a multitude of approaches have successfully bridged the direct and indirect paradigms, such as identification for control [19], [20], dual control [21], [22], control-oriented identification [23], and regularized data-enabled predictive control [24]. In essence, these approaches all advocate that the identification and control objectives should be blended to regularize each other.

An emergent approach to data-driven control is borne out of the intersection of behavioral systems theory and subspace methods; see the recent survey [25]. In particular, a result termed the *Fundamental Lemma* [26] implies that the behavior of an LTI system can be characterized by the range space of a matrix containing raw time series data. This perspective gave rise to implicit formulations (notably data-enabled predictive control [24], [27], [28]) as well as the design of explicit feedback policies [14], [15], [16], [17]. Both of these are direct

Data-Enabled Policy Optimization for Direct Adaptive Learning of the LQR

Feiran Zhao, Florian Dörfler, Alessandro Chiuso, Keyou You

Abstract—**Direct data-driven design methods for the linear quadratic regulator (LQR) mainly use offline or episodic data batches, and their online adaptation has been acknowledged as an open problem. In this paper, we propose a direct adaptive method to learn the LQR from online closed-loop data. First, we propose a new policy parameterization based on the sample covariance to formulate a direct data-driven LQR problem, which is shown to be equivalent to the certainty-equivalence LQR with optimal non-asymptotic guarantees. Second, we design a novel data-enabled policy optimization (DeePO) method to directly update the policy, where the gradient is explicitly computed using only a batch of persistently exciting (PE) data. Third, we establish its global convergence via a projected gradient dominance property. Importantly, we efficiently use DeePO to adaptively learn the LQR by performing only one-step projected gradient descent per sample of the closed-loop system, which also leads to an explicit recursive update of the policy. Under PE inputs and for bounded noise, we show that the average regret of the LQR cost is upper-bounded by two terms signifying a sublinear decrease in time $\mathcal{O}(1/\sqrt{T})$ plus a bias scaling inversely with signal-to-**

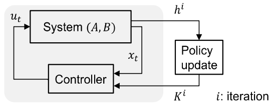


Fig. 1. An illustration of episodic approaches, where $h^i = (x_0, u_0, \ldots, x_{T^i})$ denotes the trajectory of the i-th episode.

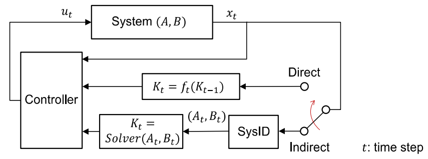


Fig. 2. An illustration of indirect and direct adaptive approaches in closed-loop, where f_t is some explicit function.

thanks