LQR Learning Pipelines

RantzerFest ECC 2024

Pietro Tesi (Florence)

Alessandro Chiuso (Padova)

Claudio de Persis (Groningen)

\+

Florian Dörfler

ETH zürich

\+

Feiran Zhao (Tsinghua)

Keyou You (Tsinghua)

Linbin Huang (Zhejiang)

Revisiting *old* problems with *old* tools in a *new* light

Linear Quadratic Dual Control

Andrea Bontz

paper posted on Arxiv as a docu-
ture at CDC2023. Some of the core
ed for publication at L4DC 2024.
r subject to (unknown) linear dynamics
atic objective is derived based on a "data-
driven Riccati equation". The main result quantifies closed loop
performance in terms of input excitation level and the degree of
the plant stabilizability.

Assuming that the system is stabilizable, the optimal value has the form $|x_0|_P^2$ where P can be obtained by solving the Riccati equation

$$|x|_P^2 = \min_u \left[|x|^2 + |u|^2 + |Ax + Bu|_P^2\right]. \qquad (1)$$

Define Q by $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right]^\top Q \left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] = |x|^2 + |u|^2 + |Ax + Bu|_P^2$. Then (1)

sample covariance parameterization

Harnessing the Final Control Error
Optimal Data-Driven Predictive Con

Alessandro Chiuso [a], Marco Fabris [a], Valentina Breschi [b], Simon

egret Analysis of Adaptive Linear Quadratic Control
with Model Misspecification

ANDRS.RANTZER@CONTROL.LTH.SE
NMATNI@SEAS.UPENN.EDU

and Systems Engineering, University of Pennsylvania
e Control, Lund University

sample complexity estimates

Formulas for Data-driven Control: Stabiliz
Optimality and Robustness

C. De Persis and P.

2 x plenary talks

ECC24
EUROPEAN CONTROL CONFERENCE

behavioral systems & subspace methods

Low-Rank and Low-

ns for Lo Identification

ers Rantzer

Grad ethods for Large-Scale and
D inear Quadratic Control

Conv mple complexity of gradient methods
for t ce linear quadratic regulator problem

Hesameddi hadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R. Jovanović

policy gradient

Annual Review of Control, Robotics, and Autonomous Systems

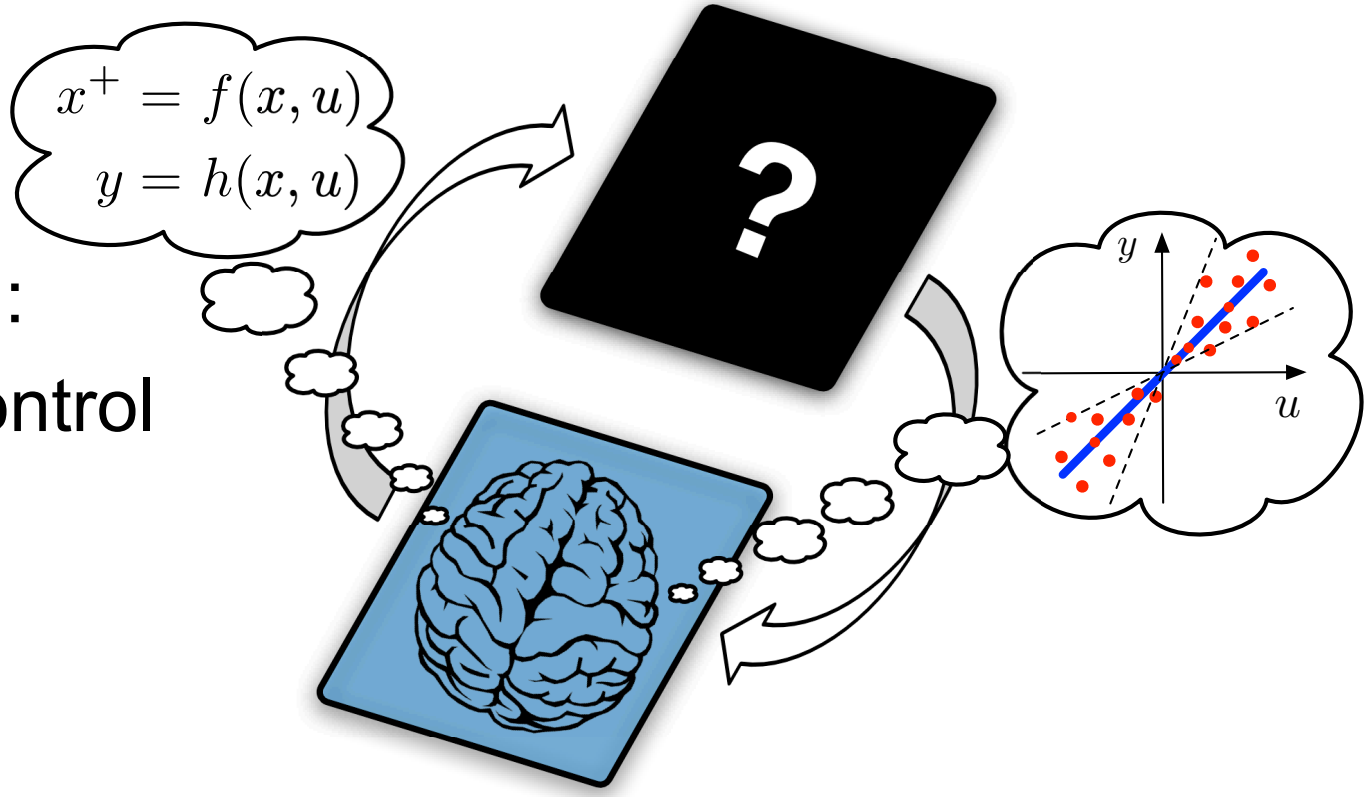
Toward a Theoretical
Foundation of Policy
Optimization for Learning
Control Policies

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

Data informativity: a new perspective on data-driven analysis and control

Henk J. van Waarde, Jaap Eising, Harry L. Trentelman, and Kanat Camlibel

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

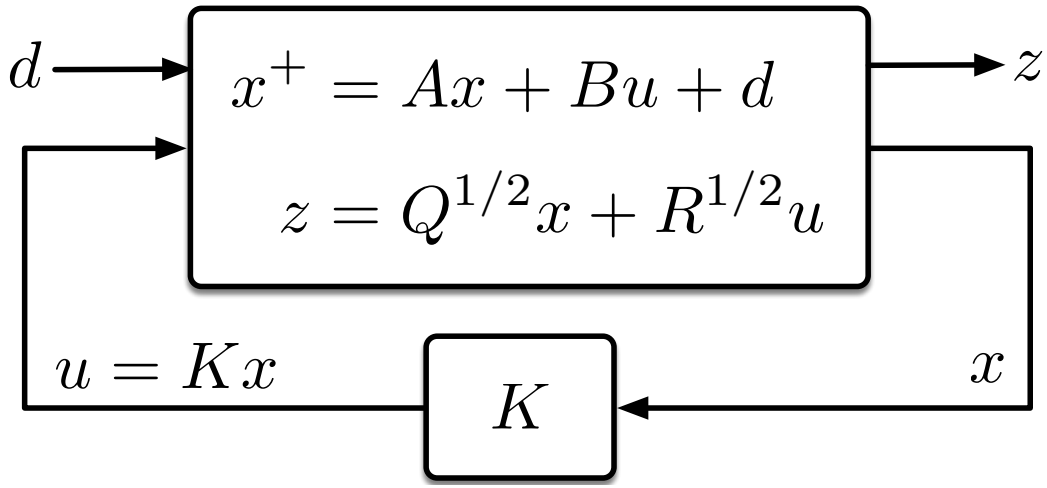
well-documented **trade-offs** concerning

- 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



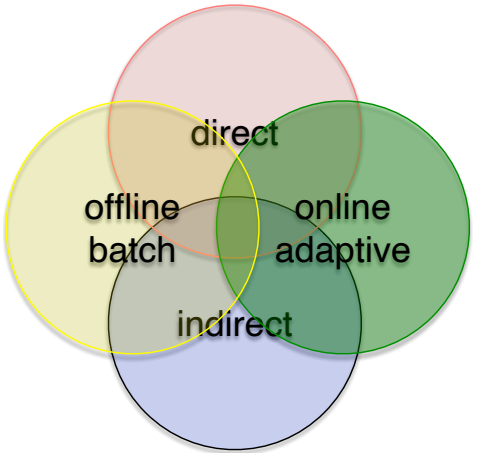
- $\mathcal{H}_2$ **parameterization**

  (can be posed as convex SDP,

  as differentiable program, as…)

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

- **<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

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)

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

certainty-equivalent LQR

least squares SysID

Direct approach from subspace relations in data

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

$$U_0 := \begin{bmatrix} u(0) & u(1) & \dots & u(T-1) \end{bmatrix} \quad D_0 := \begin{bmatrix} d(0) & d(1) & \dots & d(T-1) \end{bmatrix}$$

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

$$X_0 := \begin{bmatrix} x(0) & x(1) & \dots & x(T-1) \end{bmatrix} \quad X_1 := \begin{bmatrix} x(1) & x(2) & \dots & x(T) \end{bmatrix}$$

- **subspace relations**

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

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

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

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:

$$\left(X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\dagger \begin{bmatrix} K \\ I \end{bmatrix} \right) P \left(\cdots \right)^\top$$

$$-P + I \preceq 0$$

- **indirect** approach

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

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

Regularized, certainty-equivalent, & direct 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 (equivalent for λ large)

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

- λ **interpolates** between control & SysID … but may not be **robust (?)**

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

for robustness $GPG^\top$ should be small
→ forced by small $\|\Pi G\|$

Performance & robustness certificates

- **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 large λ — if exact system matrices *A* & *B* were known

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

Certificate for sufficiently large SNR: the optimal control problem is feasible (robustly stabilizing) with relative performance $\sim \mathcal{O}\,(1/SNR)$.

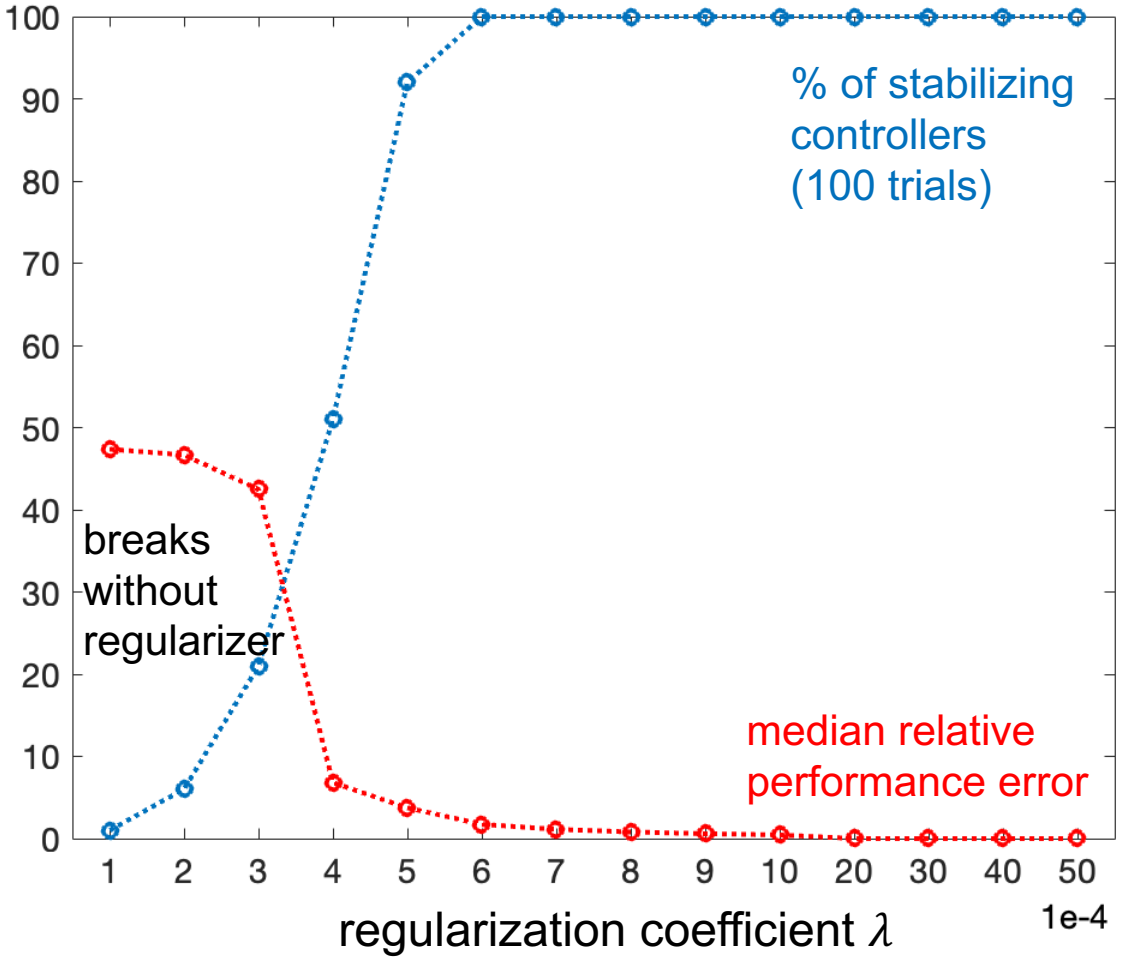
Numerical case study

- **case study** [Dean et al. '19]: discrete-time system with noise variance $\sigma^2 = 0.01$ & variable regularization coefficient λ

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

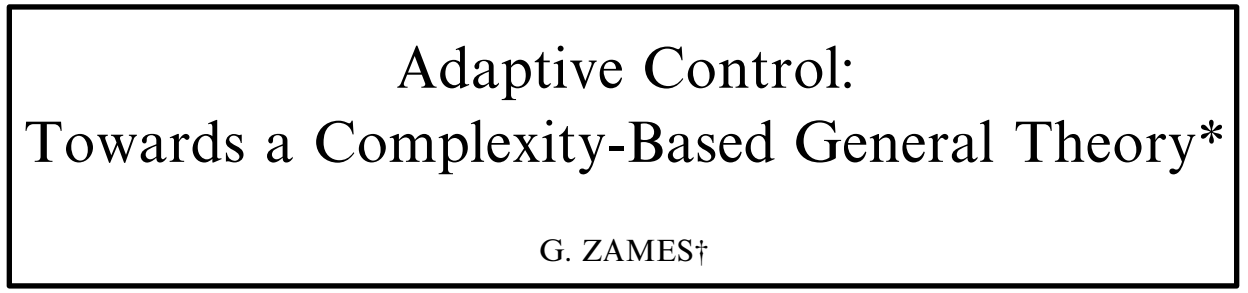
- **take-home message**: regularization is *needed* for robustness & performance

→ works… but lame: **learning is offline**



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)

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

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^*.

Model-free policy gradient methods

- policy gradient: $K^+ = K - \eta\, \nabla J(K)$ converges linearly to K^*

- **model-based setting**: explicit *Anderson-Moore formula* for $\nabla J(K)$ based on closed-loop controllability + observability Gramians

- **model-free 0$^{\text{th}}$ order methods** constructing two-point gradient estimate 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

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

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** $\Lambda = \frac{1}{t}\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \quad \& \quad \overline{X}_1 = \frac{1}{t} X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top$

- **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

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

details are not important

Projected policy gradient with sample covariances

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

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

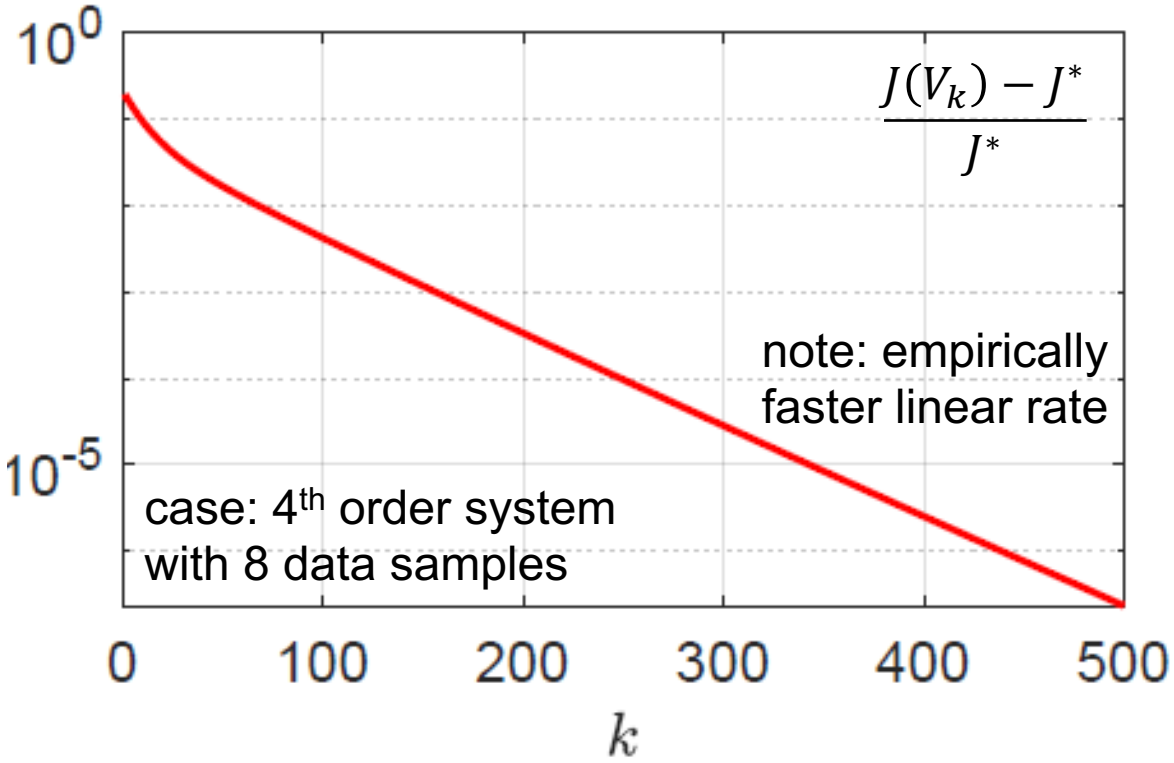
$\Pi_{\overline{X}_0}$ projects on parameterization constraint $I = \overline{X}_0 V$ & gradient $\nabla J(V)$ is computed from two Lyapunov equations with sample covariances

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

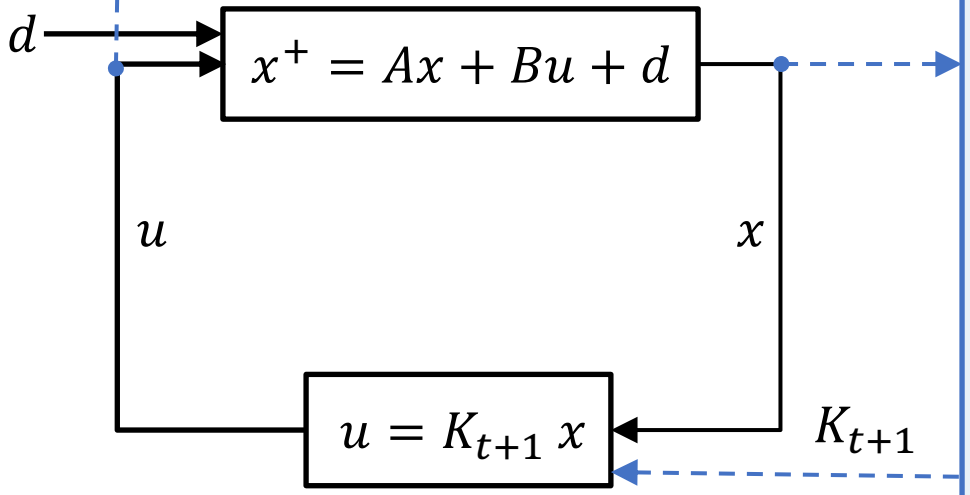
$$J(V) - J^* \leq const. \cdot \left\| \Pi_{\overline{X}_0}\left(\nabla J(V)\right) \right\|$$

- 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 $\|u(t)\| \leq \bar{u} \quad \& \quad \|x(t)\| \leq \bar{x}$

  (∃ 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 &

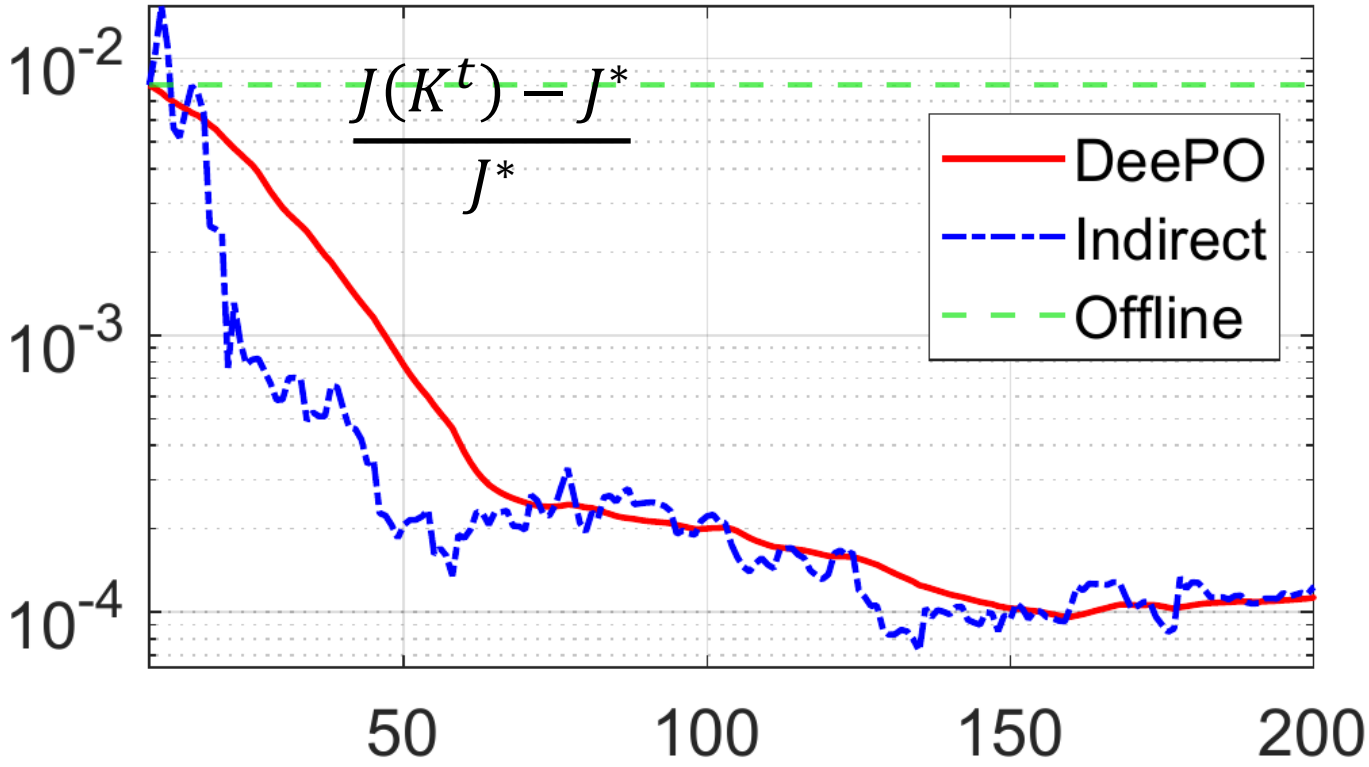
$$\mathrm{Regret}_T \leq \frac{\nu_3}{\sqrt{T}} + \frac{\nu_4}{\sqrt{SNR}} \ .$$

- **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]

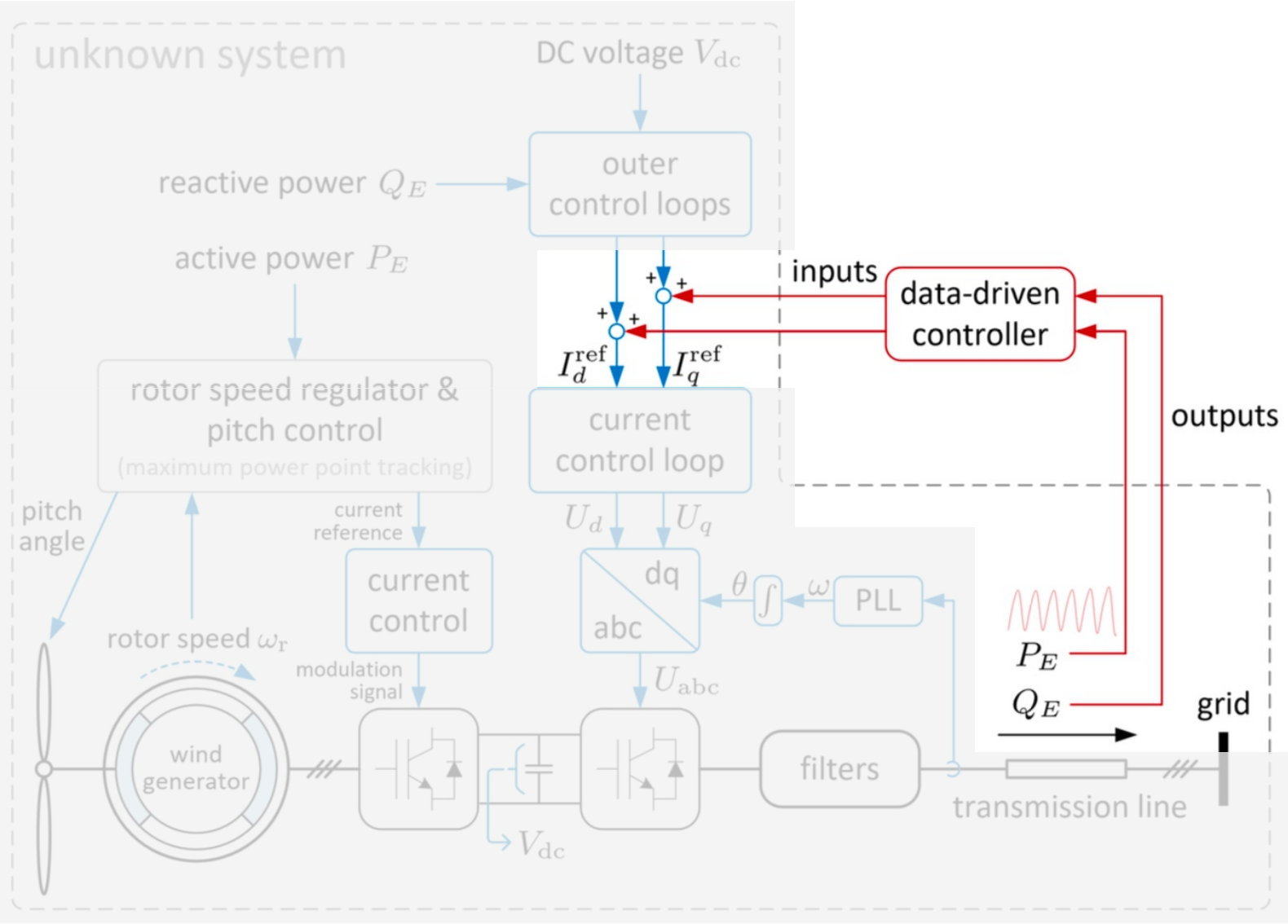
- **case 1:** offline LQR
  vs direct adaptive DeePO
  vs indirect adaptive: rls + dlqr
  → **adaptive outperforms offline**
  → direct/indirect **rates matching**
  but **direct is much(!) cheaper**



- **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

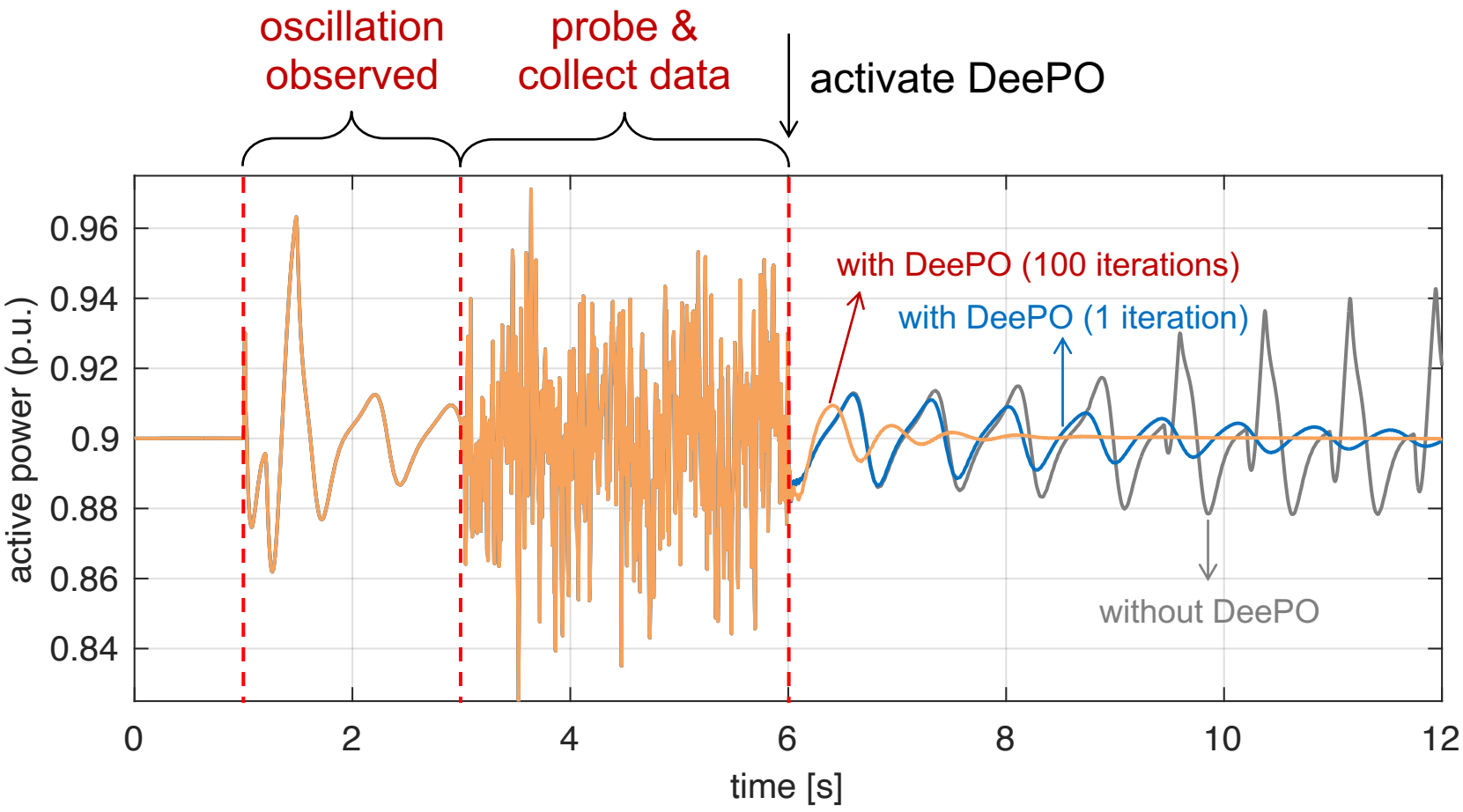
Power systems / electronics case study



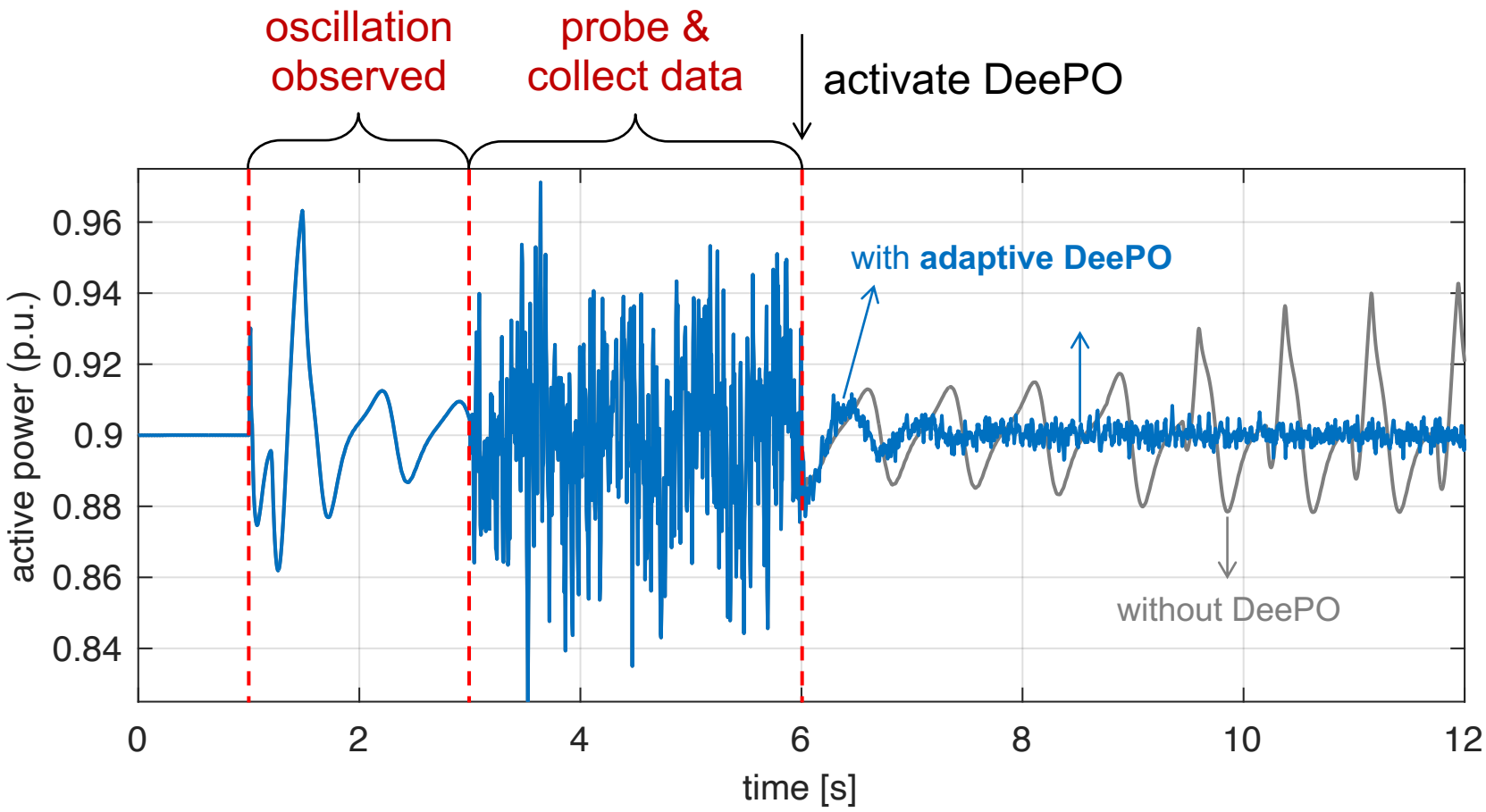
synchronous generator & full-scale converter

- 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

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

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.

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-**

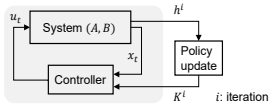


Fig. 1. An illustration of episodic approaches, where $h^i = (x_0, u_0, \ldots, x_{T^i})$ denotes the trajectory of the i-th episode.

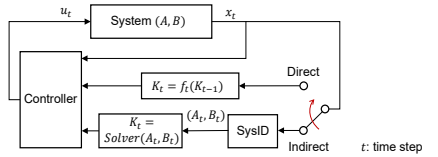


Fig. 2. An illustration of indirect and direct adaptive approaches in closed-loop, where f_t is some explicit function.