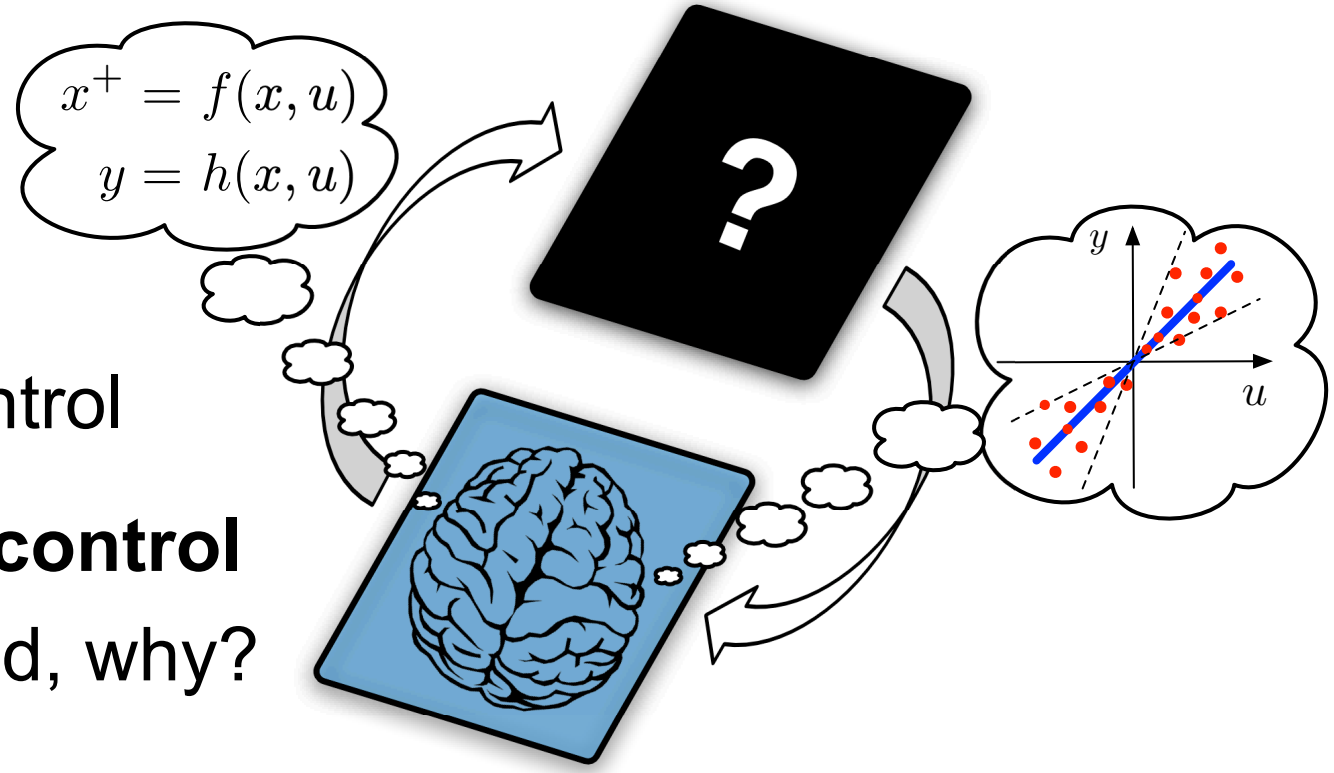
Direct Adaptive Learning of the LQR

Florian Dörfler

IFAC TC 1.2 Adaptive & Learning Systems Webinar, 2/5/2024

Data-driven control



- **indirect data-driven control**
  data $\xrightarrow{\text{ID}}$ model + uncertainty $\rightarrow$ control
- growing trend **direct data-driven control**
  by-passing models… (again) hyped, why?

The direct approach is a **viable alternative**

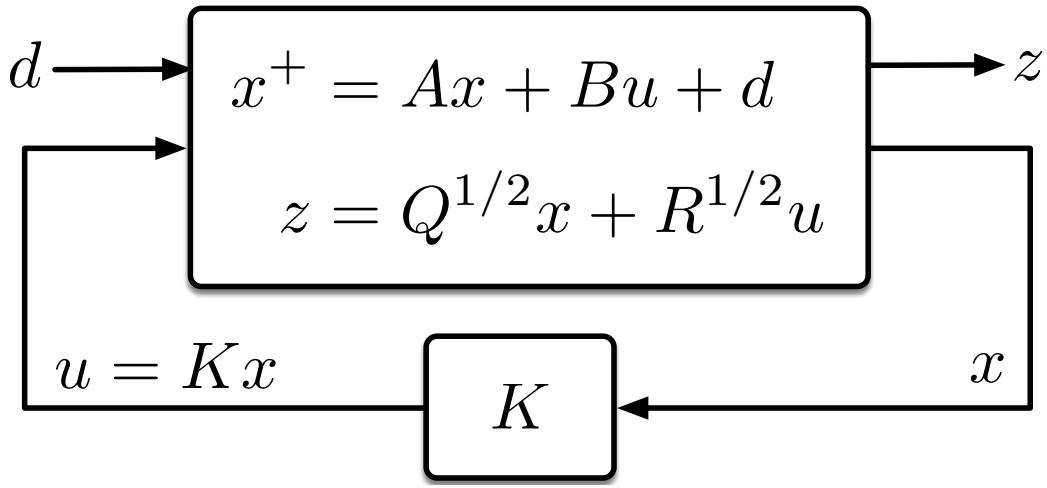
- for some **applications**: model-based approach is too complex to be useful → complex processes, sensing modalities, environment
- due to **shortcomings of ID** → cumbersome, models not identified for control, model selection, or incompatible uncertainty estimates
- when sufficient **brute force** data / compute / storage is available

- **trade-offs**
  - (non)modular
  - (in)tractable
  - (sub)optimal
  - data size
  - online adaptation

today: give explicit answers for **LQR**

LQR

- **cornerstone** of automatic control



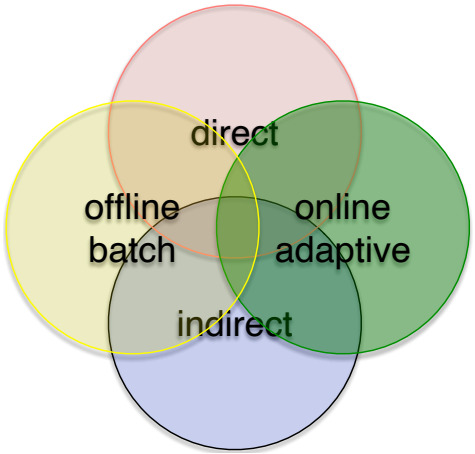
- $\mathcal{H}_2$ **parameterization**

  (can be posed as convex SDP,

  as differentiable program, as…)

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

- **<u>the</u> benchmark** for all data-driven control approaches in last decades(!)



Contents

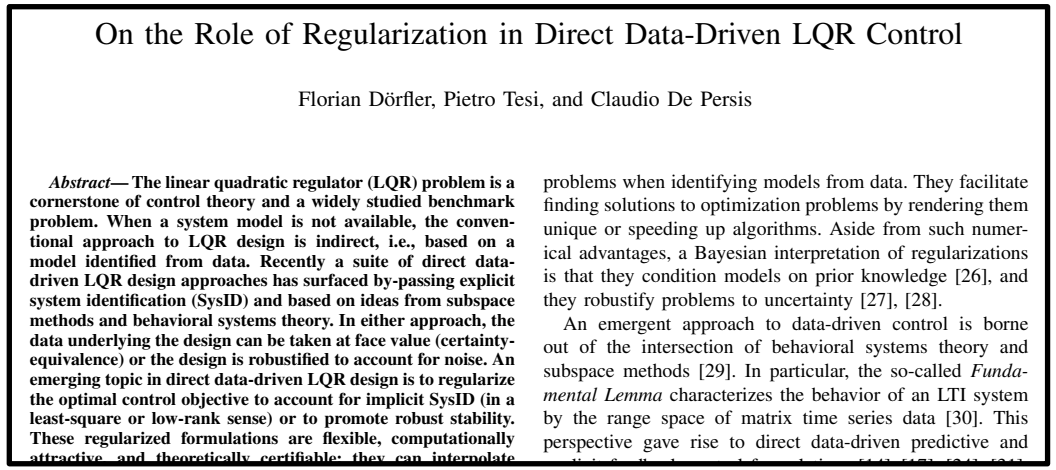
1. **regularizations** bridging direct & indirect data-driven LQR
   → story of a *model-based pipeline with model-free elements*

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

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*



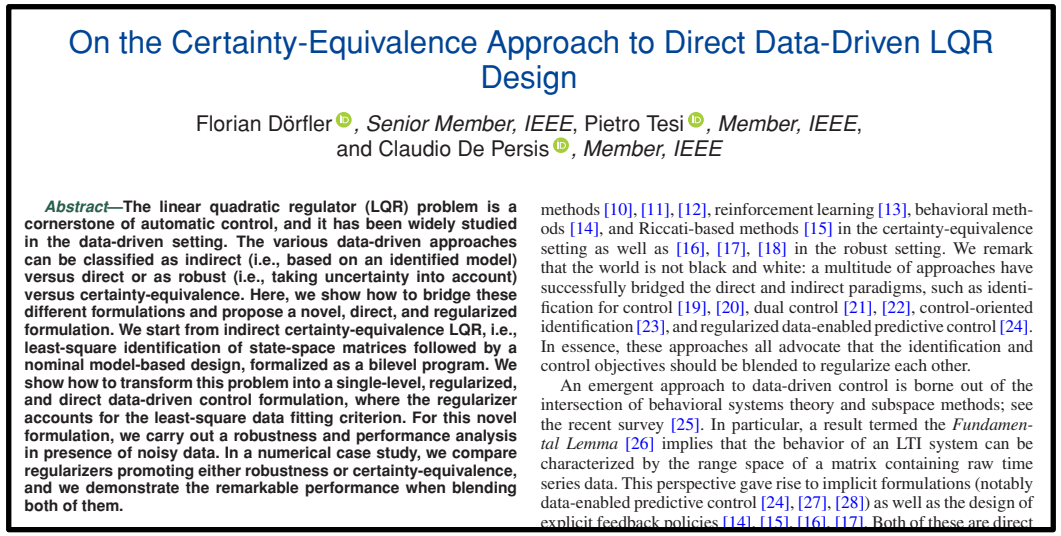
On the Role of Regularization in Direct Data-Driven LQR Control

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 explicit feedback policies [14], [15], [16], [17]. Both of these are direct

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)

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

certainty-equivalent LQR

least squares SysID

Direct approach from subspace relations in data

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

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

$$X_1 = AX_0 + BU_0 + D_0$$

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

- **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: $A + BK = X_1 G$

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 $\sim \mathcal{O}\,(1/SNR)$.

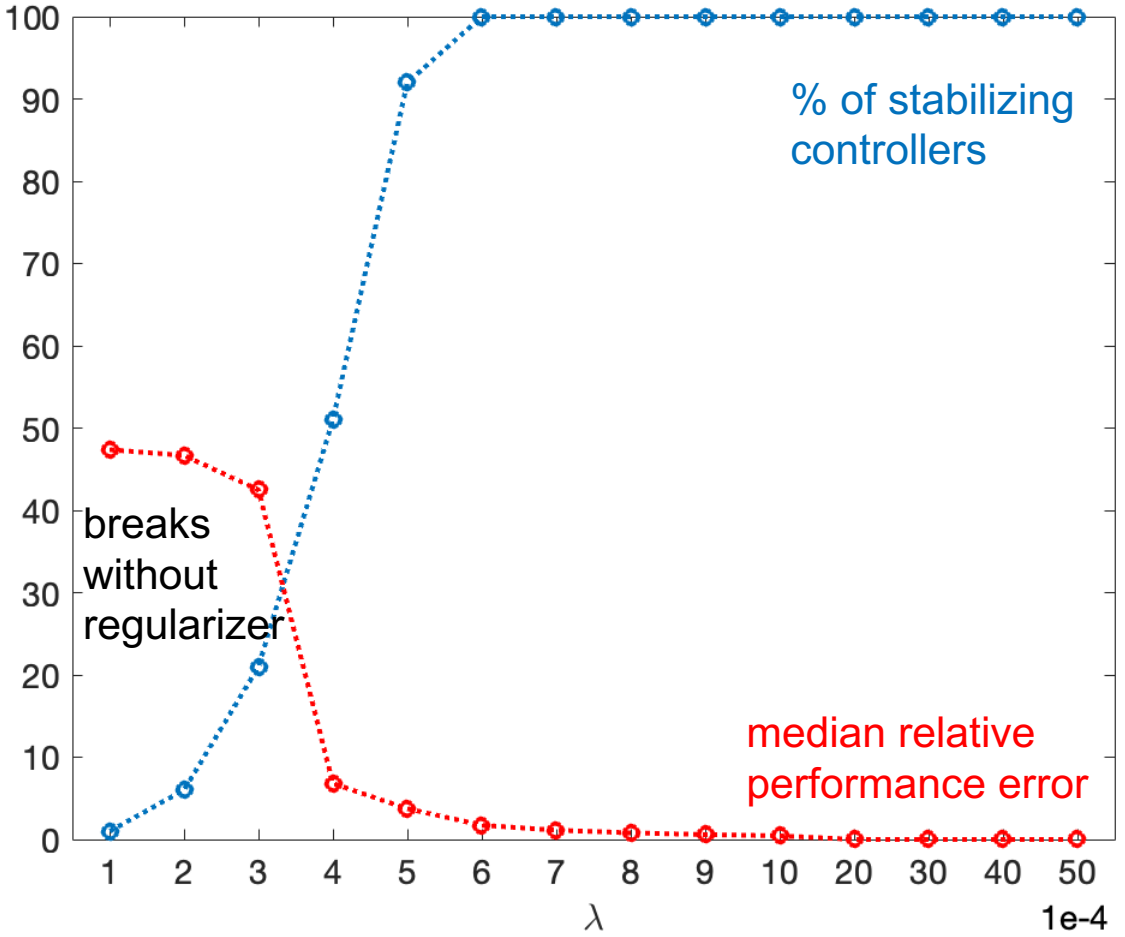
Numerical case study

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

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

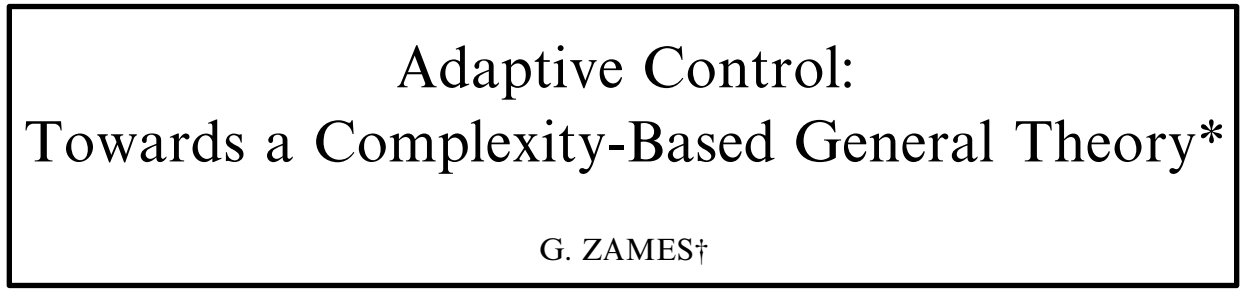
- **take-home message**: regularization is *needed* for robustness & performance

→ works well … but **learning is offline**



Why online, adaptive, & what does it mean anyways ?

- **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 / batch) algorithms, with closed-loop data, & recursive algorithmic implementation

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)

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-**

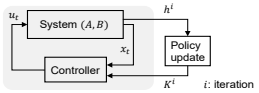


Fig. 1. An illustration of episodic approaches, where $h^i = (x_0, u_0, \ldots, x_{T^i})$ denotes the trajectory of the i-th episode.

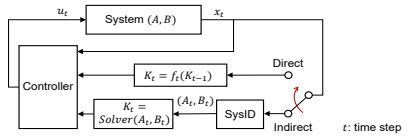


Fig. 2. An illustration of indirect and direct adaptive approaches in closed-loop, where f_t is some explicit function.

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

Fact: gradient descent $K^+ = K - \eta\, \nabla J(K)$ initialized from a stabilizing policy converges linearly to K^*.

- **model-based setting**: explicit *Anderson-Moore formula* for $\nabla J(K)$

- **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

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

Ingredient 2: 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 $\Lambda = \frac{1}{t} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix} \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top \succ 0$
- $A + BK = [B \; A] \begin{bmatrix} K \\ I \end{bmatrix} = [B \; A] \Lambda V = \underbrace{\frac{1}{t} X_1 \begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top}_{= \overline{X}_1} V$
- robustness for free without regularization
- dimension of all matrices is constant

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 dynamics** expressed with $\begin{bmatrix} K \\ I \end{bmatrix} = \Lambda V$ as $A + BK = \overline{X}_1 V$
- **covariance parameterization**

$$\min_{V,K,\Sigma \succ 0} \quad \text{trace}(Q\Sigma) + \text{trace}\left(K^T R K \Sigma\right)$$
$$\text{s.t. } \Sigma = I + \overline{X}_1 V \, \Sigma \, V^T \overline{X}_1^T \;, \begin{bmatrix} K \\ I \end{bmatrix} = \Lambda V$$

- with $\begin{bmatrix} K \\ I \end{bmatrix} = \frac{1}{t}\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}\begin{bmatrix} U_0 \\ X_0 \end{bmatrix}^\top V = \begin{bmatrix} \overline{U}_0 \\ \overline{X}_0 \end{bmatrix} V$

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

Policy gradient with covariance parameterization

- warm-up scenario: offline PE data (X_0, U_0, X_1) without disturbances $d(t)$

- parameterization:

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

after eliminating (unique) Σ, we denote blue part by $J(V)$

- **data-enabled policy optimization** (**DeePO**) via projected gradient

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

where $\Pi_{\overline{X}_0}$ projects on $I = \overline{X}_0 V$ & gradient $\nabla J(V) =$ is computed from data:

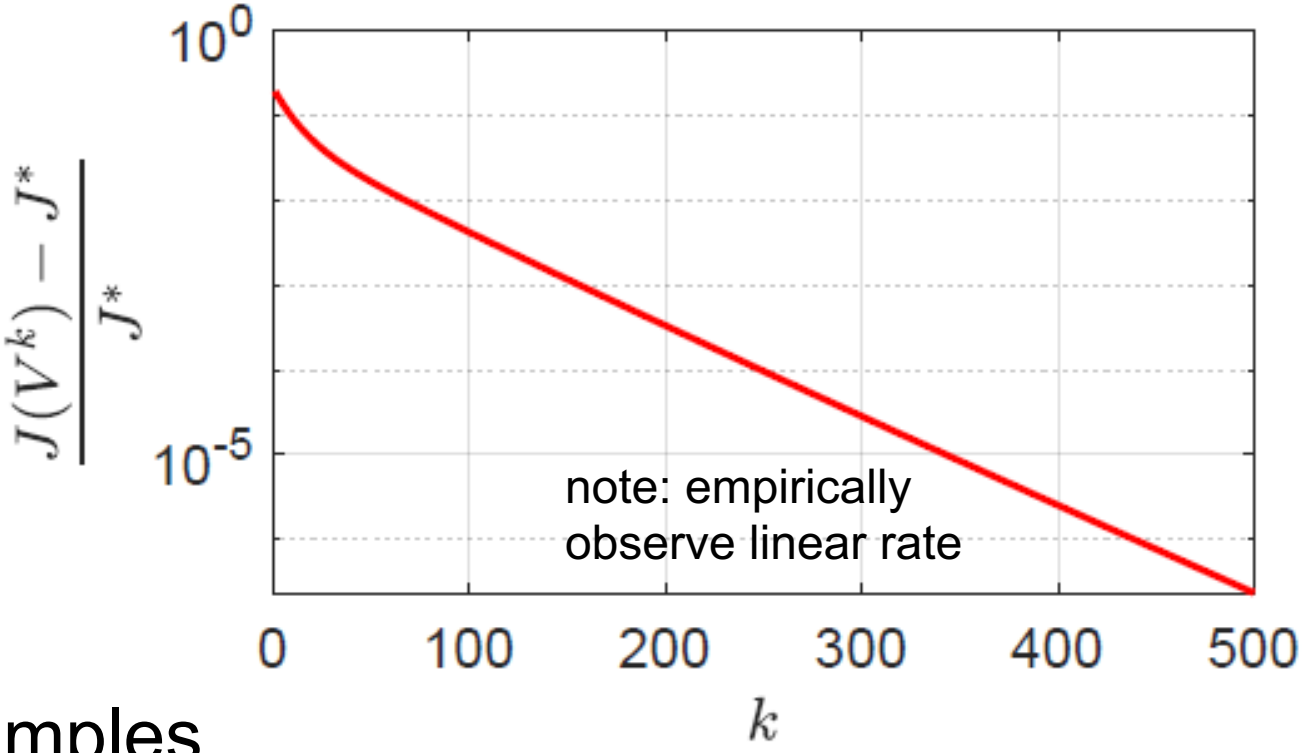
$$\nabla J(V) = 2\left(\overline{U}_0^T R\overline{U}_0 + \overline{X}_1^T P\overline{X}_1\right)V\,\Sigma \quad \text{with} \quad P = Q + V^T\overline{U}_0^T R\overline{U}_0 V + V^T\overline{X}_1^T P\overline{X}_1 V$$

Features of data-enabled policy optimization (DeePO)

- **optimization landscape**: for any feasible $V \in \mathcal{S}(a) = \{V \mid J(V) \leq a\}$
  - **projected gradient dominance** of degree 1: $J(V) - J^* \leq \mu(a) \left\| \Pi_{\overline{X}_0}(\nabla J(V)) \right\|$
  - smoothness with a **bounded Hessian**: $\|\nabla^2 J(V)\| \leq l(a)$

Sublinear convergence for a feasible initialization $V^0 \in \mathcal{S}(a)$ & step size $\eta \in (0, 1/l(a)]$. Then $\forall\, \epsilon > 0,\ J(V^k) - J^* \leq \epsilon,$ where

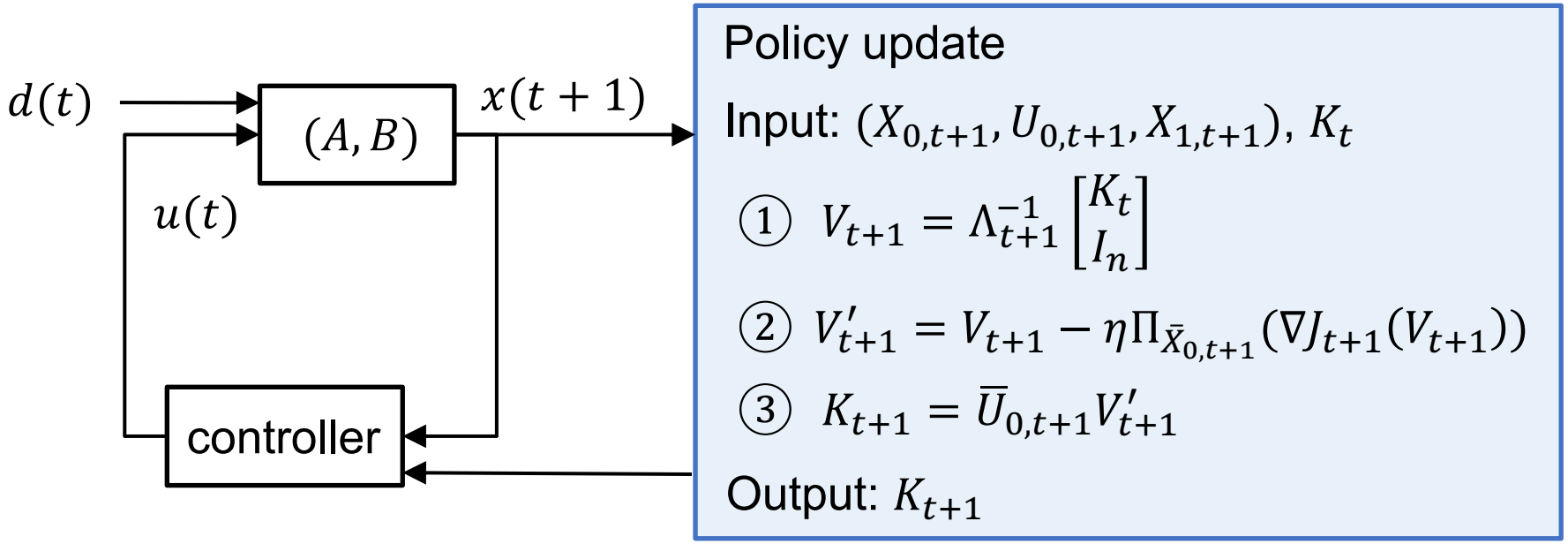
$$k \geq \frac{2\mu(a)^2}{\epsilon \cdot (2\eta - l(a) \cdot \eta^2)}\ .$$



- **simulation:** 4th order system, 8 samples

Online & adaptive DeePO

- **features**: direct, online, closed-loop data, & recursive implementation



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 sample covariances, cheap computation, & no memory needed for old data

Underlying assumption 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}.

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

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

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

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

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 & $\mathrm{Regret}_{T\to\infty} \to 0$
  - favorable sample complexity: sublinear decrease term matches best rate $\mathcal{O}(1/\sqrt{T})$ of first-order methods in online convex optimization

Numerical case study

- **setup:** random controllable & stable system of order 4 subject to uniform process noise with variance σ, 2 inputs, $Q = I_4$, & $R = I_2$

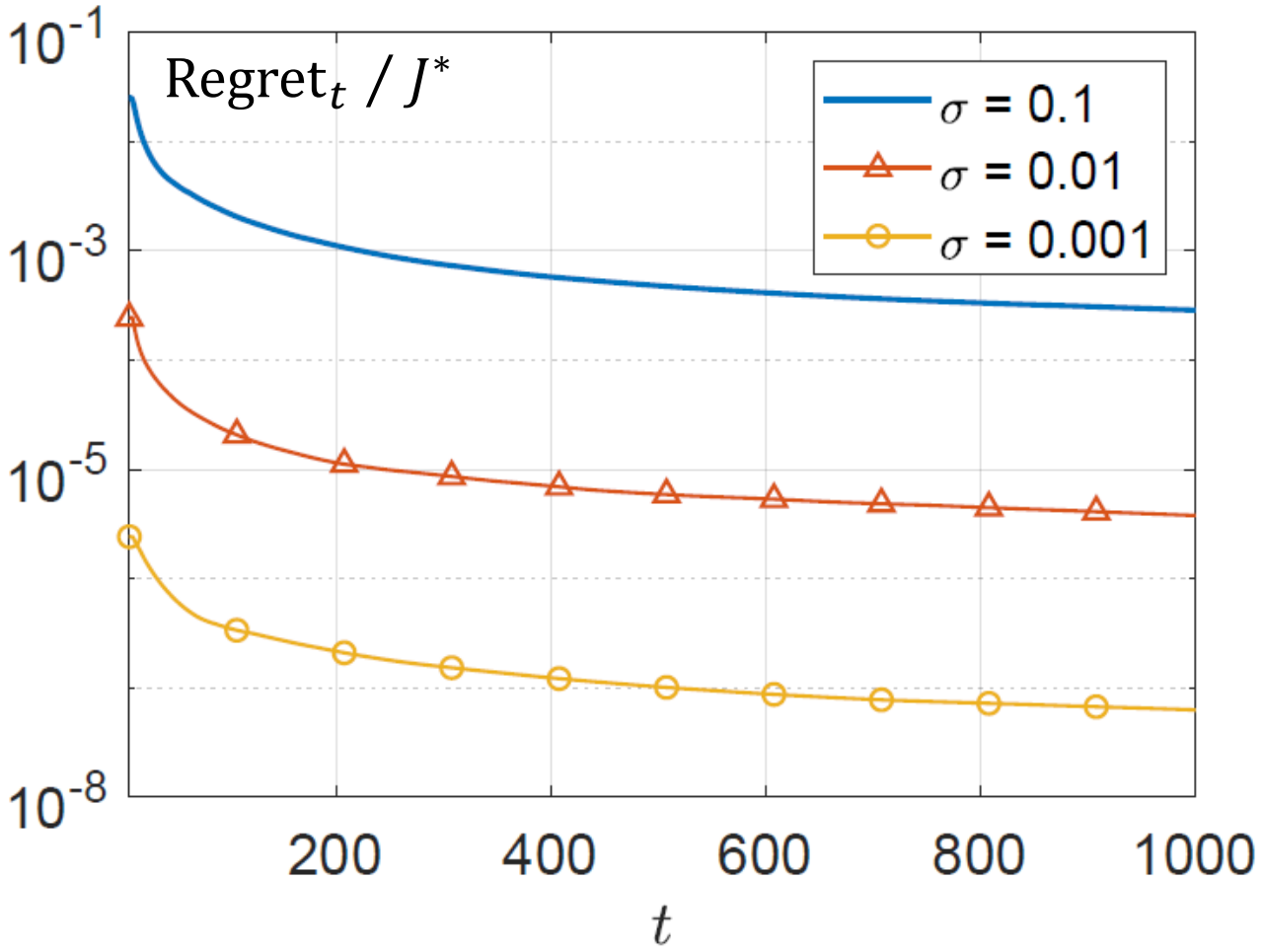
- **adaptive DeePO** implementation

- **theoretical certificate**

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

- **empirically** observe

$$\text{Regret}_T \leq \frac{\nu_3}{\sqrt{T}} + \frac{\nu_4}{SNR^2}$$

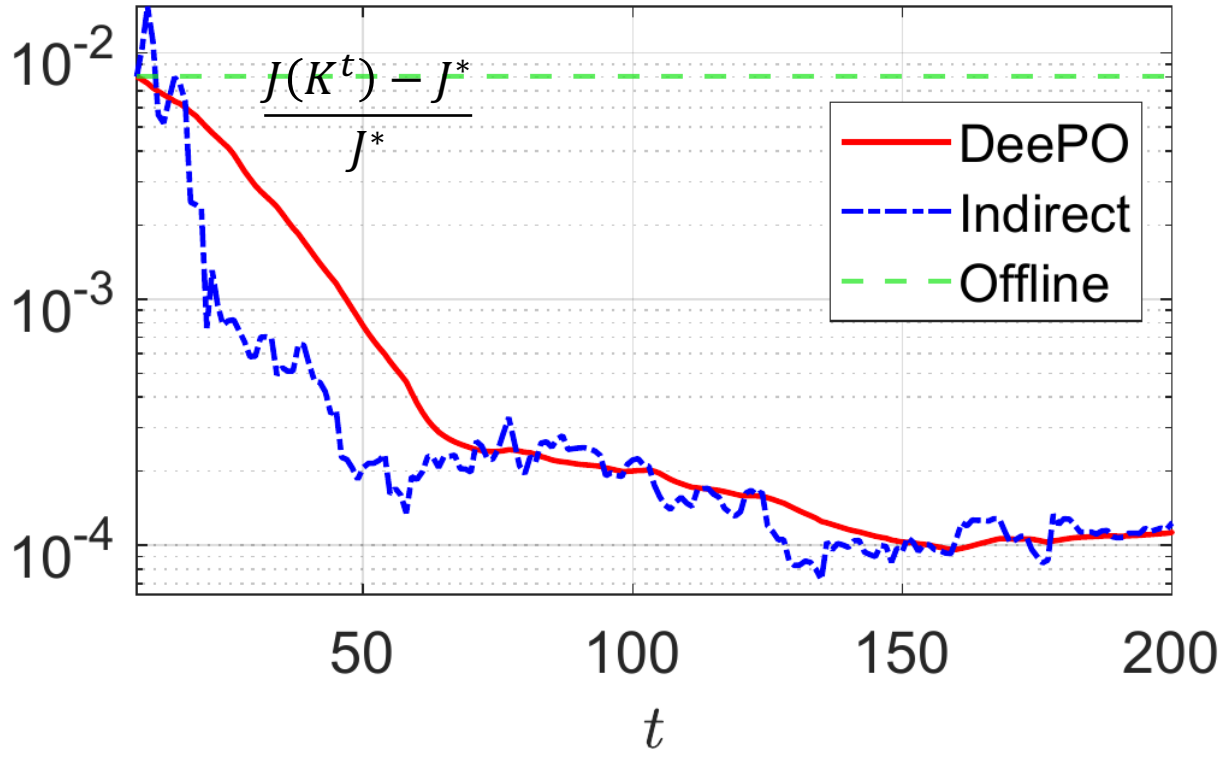


Comparison case study

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

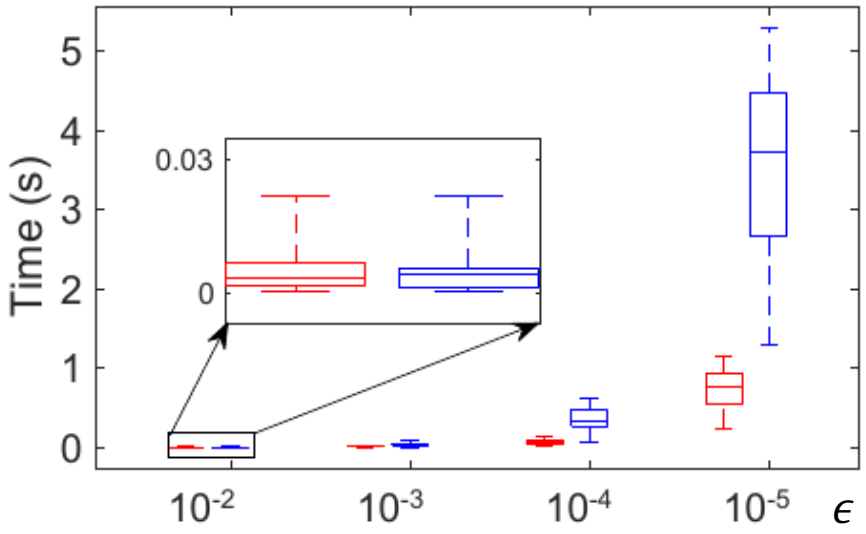
$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B = I$$

- **comparison**: DeePO vs offline LQR design vs indirect adaptive approach (rls + dlqr) [Wang et al. '21, Lu et al. '23]

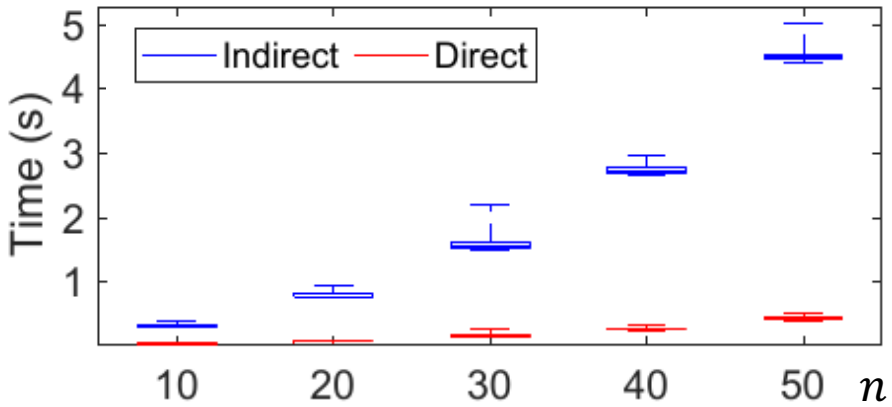


Comparison of computational & sample complexity

time to reach desired accuracy ϵ



time for increasing state dimension n

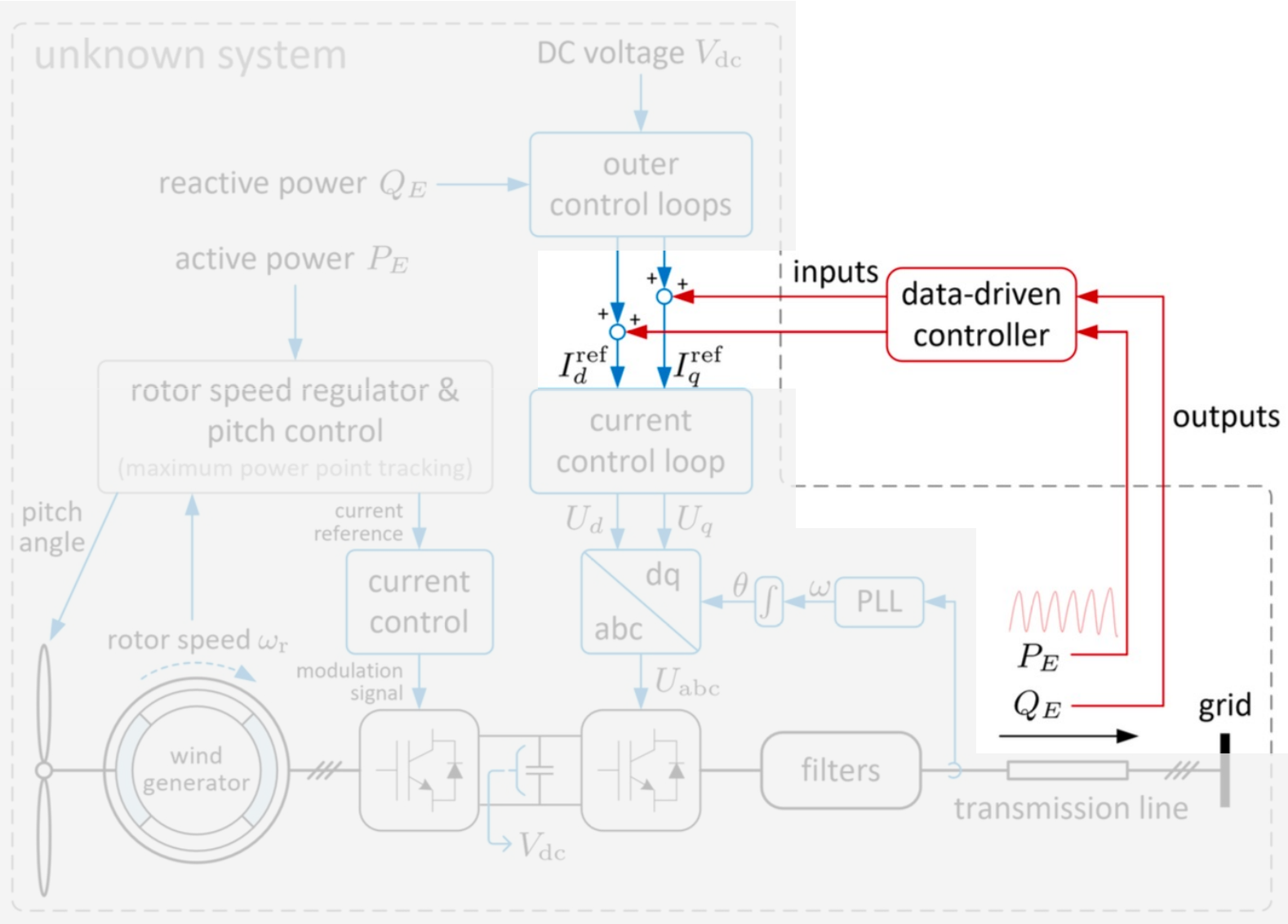


← direct DeePO significantly outperforms indirect adaptive design in **computational effort**

↓ DeePO requires **significantly fewer data** samples than model-free 0^{th} order gradient methods

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

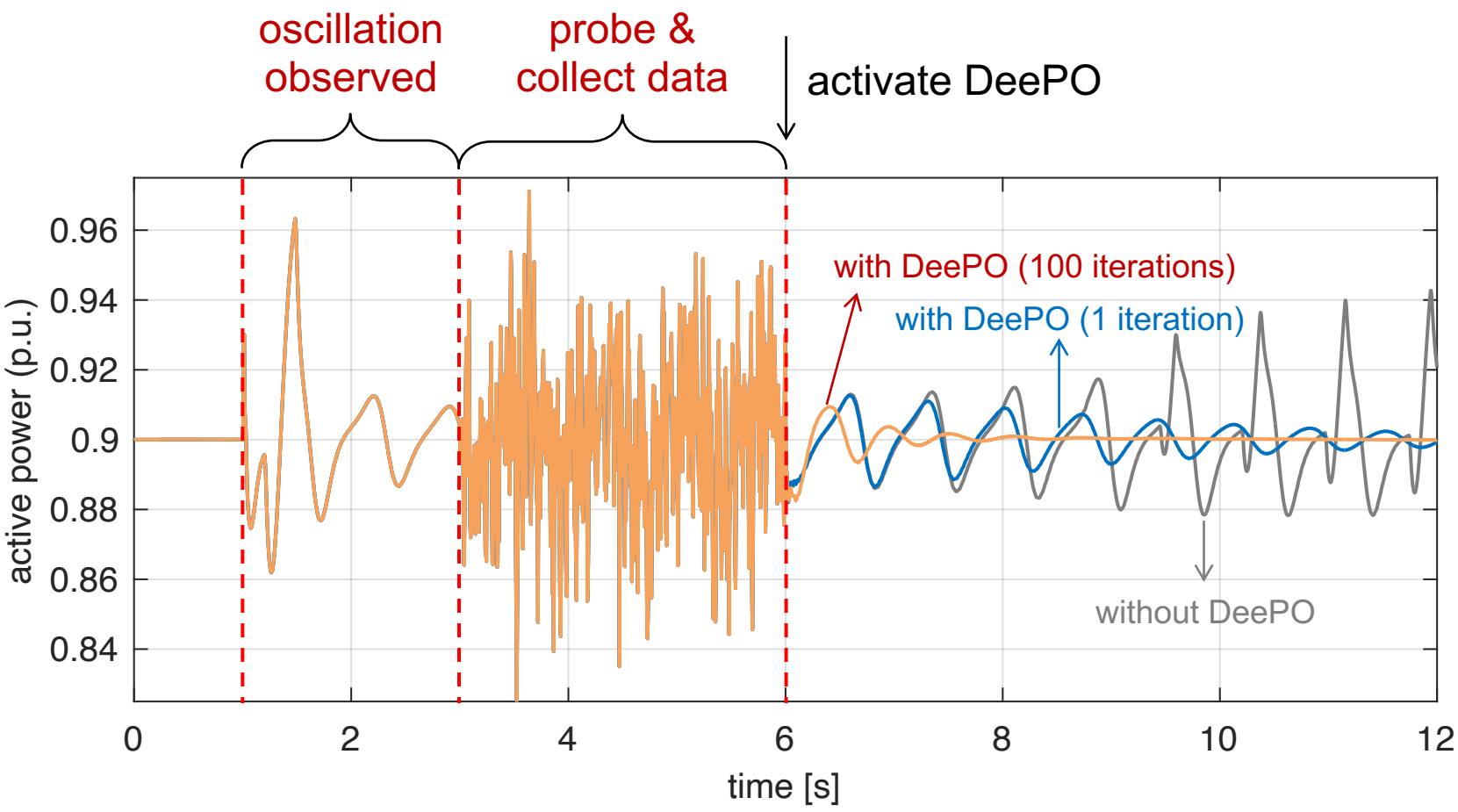
Power systems case study



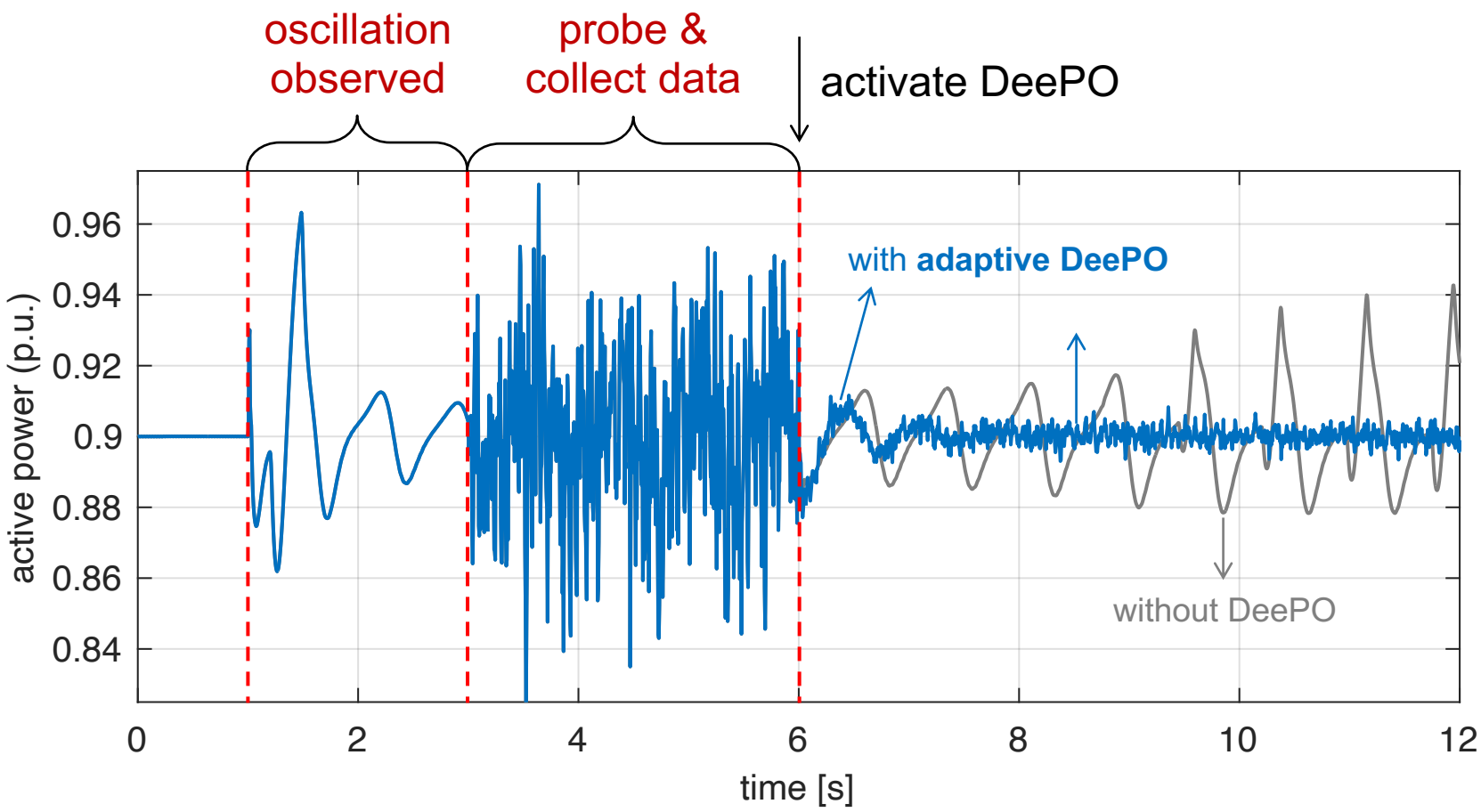
synchronous generator & full-scale converter

- 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 case study



Power systems case study



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 with sample covariance
    → DeePO is adaptive, online, with closed-loop data, & recursive implementation
  - academic case studies & can be made useful in power systems

- **Future work**
  - technicalities: weaken assumptions & improve rates
  - control: based on output feedback & for other objectives
  - adaptivity: sliding data window &/or forgetting factor
  - further system classes: stochastic & time-varying