A Modified Adaptive Data-Enabled Policy Optimization Control to Resolve State Perturbations

Mojtaba Kaheni, Niklas Persson, Vittorio De Iuliis, Costanzo Manes, and Alessandro V. Papadopoulos

The Data-enabled Policy Optimization (DeePO) algorithm [1], [2] is an adaptive, direct, data-driven method for computing Linear Quadratic Regulators (LQR) for controllable linear time-invariant (LTI) systems. DeePO incorporates an adaptation feature on top of direct data-driven LQR design. At each time step, newly measured input and state data are added to the previously stored dataset, and the control feedback gain is updated iteratively using a learning rate, steering the design towards reducing the objective function's cost based on the updated data. Consider an LTI discrete-time system, represented in state space form as:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \boldsymbol{\omega}_k, \\ \mathbf{z}_k &= \begin{bmatrix} \mathbf{Q}^{1/2} & \mathbf{0}_{n\times m} \\ \mathbf{0}_{m\times n} & \mathbf{R}^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}, \end{aligned} \tag{1}$$

where $k \in \mathbb{N}$ is the index for counting samples, $\mathbf{x} \in \mathbb{R}^n$ is the state, $\mathbf{u} \in \mathbb{R}^m$ represents the input, and $\boldsymbol{\omega}_k$ is noise. Furthermore, let $\mathbf{z}_k \in \mathbb{R}^{n+m}$ represent the performance signal. We assume that the pair $(\mathbf{A},\mathbf{B})$ is controllable, and that $(\mathbf{Q},\mathbf{R})$ are positive definite square matrices with compatible dimensions. The objective of the LQR design is to determine a state feedback controller, $\mathbf{K} \in \mathbb{R}^{m\times n}$, that minimizes the $\mathscr{H}_2$-norm of the transfer function $\mathscr{T}(\mathbf{K}) : \boldsymbol{\omega} \mapsto \mathbf{z}$ of:

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \mathbf{z}_k \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{A}+\mathbf{B}\mathbf{K} & \mathbf{I}_n \\ \hline \begin{bmatrix} \mathbf{Q}^{1/2} \\ \mathbf{R}^{1/2}\mathbf{K} \end{bmatrix} & \mathbf{0}_{(m+n)\times n} \end{array}\right] \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\omega}_k \end{bmatrix} \tag{2}$$

If $(\mathbf{A},\mathbf{B})$ are unknown, it may still be possible to find LQR. Suppose signals of length t of states, inputs, noises, and successor states, which do not necessarily need to be consecutive. These signals are defined as follows:

$$\begin{aligned} \mathbf{X}_0 &\triangleq \begin{bmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{t-1} \end{bmatrix}, \\ \mathbf{X}_1 &\triangleq \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_t \end{bmatrix}, \\ \mathbf{U}_0 &\triangleq \begin{bmatrix} \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_{t-1} \end{bmatrix}, \\ \mathbf{W}_0 &\triangleq \begin{bmatrix} \boldsymbol{\omega}_0 & \boldsymbol{\omega}_1 & \cdots & \boldsymbol{\omega}_{t-1} \end{bmatrix}. \end{aligned} \tag{3}$$

The input signal $\mathbf{U}_0$ must be *sufficiently rich* to effectively represent the dynamical system described by (1). This property is commonly referred to as *persistently exciting*.

Definition 1 ([3]): A signal $\mathbf{U}_0$ is said to be persistently exciting of order l when

$$\mathscr{U}_0 = \begin{bmatrix} \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_{t-l} \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{t-l+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_{l-1} & \mathbf{u}_l & \cdots & \mathbf{u}_{t-1} \end{bmatrix} \tag{4}$$

has full rank ml. ■

The following lemma is also useful for determining the persistent excitation of a system.

Lemma 1 ([3]): If the system (1) is controllable and $\mathbf{U}_0$ is persistently exciting of order $n+1$, then

$$\operatorname{rank}(\mathscr{D}) = n+m, \tag{5}$$

where

$$\mathscr{D} \triangleq \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{X}_0 \end{bmatrix}. \tag{6}$$

■

In [2], the authors introduce a policy parametrization based on the sample covariance of the data, defined as:

$$\Phi \triangleq \frac{1}{t}\mathscr{D}\mathscr{D}^\top = \begin{bmatrix} \mathbf{U}_0\mathscr{D}^\top/t \\ \mathbf{X}_0\mathscr{D}^\top/t \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{U}}_0 \\ \overline{\mathbf{X}}_0 \end{bmatrix}. \tag{7}$$

Defining $\mathbf{V} \in \mathbb{R}^{(n+m)\times n}$ as the solution to:

$$\begin{bmatrix} \mathbf{K} \\ \mathbf{I}_n \end{bmatrix} = \Phi\mathbf{V}, \tag{8}$$

then, the data-driven LQR optimization problem can be reformulated as:

$$\begin{aligned} \min_{\mathbf{V},\boldsymbol{\Sigma}_\mathbf{V}\succeq 0} \quad & C(\mathbf{V}) = \operatorname{Tr}\left(\left(\mathbf{Q}+\mathbf{V}^\top\overline{\mathbf{U}}_0^\top\mathbf{R}\overline{\mathbf{U}}_0\mathbf{V}\right)\boldsymbol{\Sigma}_\mathbf{V}\right) \\ \text{subject to} \quad & \boldsymbol{\Sigma}_\mathbf{V} = \mathbf{I}_n + \overline{\mathbf{X}}_1\mathbf{V}\boldsymbol{\Sigma}_\mathbf{V}\mathbf{V}^\top\overline{\mathbf{X}}_1^\top, \\ & \overline{\mathbf{X}}_0\mathbf{V} = \mathbf{I}_n. \end{aligned} \tag{9}$$

Since the dimension of $\mathbf{V}$ is independent of the number of samples, t, this formulation is particularly advantageous in adaptive design strategies where the sample size grows linearly. In the DeePO algorithm, starting from an initial feasible solution $\mathbf{K}_t$, the feedback gain evolves iteratively via a gradient descent approach to reach the optimal solution $\mathbf{K}^*$. Since both $\mathbf{x}_i \to 0$ and $\mathbf{K}_i \to \mathbf{K}$ may compromise the full rank condition of Φ, a probing noise $\mathbf{e}_i$ is added to the control input in DeePO, resulting in $\mathbf{u}_i = \mathbf{K}_i\mathbf{x}_i + \mathbf{e}_i$. In asymptotically stable LTI systems, the system naturally drives the state $\mathbf{x}_i$ towards equilibrium as time progresses. However, when probing noise is added to the input signal to maintain persistent excitation, the noise introduces high-frequency components into the control input. These high-frequency components can interact with the feedback dynamics, causing rapid oscillations or fluctuations in the control signal and, consequently, the system state. States perturbations is particularly problematic in practical implementations, as it can lead to actuator wear, increased energy consumption, and degraded overall system performance.

We propose Perturbation-Free DeePO (PFDeePO) to address the aforementioned drawbacks. The modification we propose to the original DeePO algorithm is summarized in the following.

Algorithm 1 Perturbations-free DeePO (PFDeePO)

Require: $\mathbf{U}_0$, $\mathbf{X}_0$, $\mathbf{X}_1$, $\mathbf{K}_t$, $\gamma > 0$, $\delta > 0$, and $\eta > 0$.
Start
$i = t$.
$\Delta\mathbf{K} = (\delta + 1) \cdot \mathbf{1}_{m \times n}$.
while the stop criterion is not satisfied, do:
if $\|\Delta\mathbf{K}\| > \delta$ **or** $\|\mathbf{x}_i\| \leq \gamma$
Apply $\mathbf{u}_i = \mathbf{K}_i \mathbf{x}_i$ and observe $\mathbf{x}_{i+1}$.
else
Find $\overline{\nu}$ and $\underline{\nu}$ that ensures stability.
Randomly select $\underline{\nu} \leq \nu_i \leq \overline{\nu}$.
Apply $\mathbf{u}_i = \nu_i \mathbf{K}_i \mathbf{x}_i$ and observe $\mathbf{x}_{i+1}$.
End if
if $\|\mathbf{x}_i\| > \gamma$
Update $\mathbf{X}_0$ by $\mathbf{X}_0 = [\mathbf{X}_0, \mathbf{x}_i]$.
Update $\mathbf{X}_1$ by $\mathbf{X}_1 = [\mathbf{X}_1, \mathbf{x}_{i+1}]$.
Update $\mathbf{U}_0$ by $\mathbf{U}_0 = [\mathbf{U}_0, \mathbf{u}_i]$.
$\mathbf{V}_{i+1} = \Phi_{i+1}^{-1} \begin{bmatrix} \mathbf{K}_i \\ \mathbf{I}_n \end{bmatrix}$.
$\mathbf{V}'_{i+1} = \mathbf{V}_{i+1} - \eta \Pi_{\overline{\mathbf{X}}_0} \widehat{\nabla C}$.
Update the control gain by $\mathbf{K}_{i+1} = \overline{\mathbf{U}}_0 \mathbf{V}'_{i+1}$.
else
Update the control gain by $\mathbf{K}_{i+1} = \mathbf{K}_i$.
End if
$\Delta\mathbf{K} = \mathbf{K}_{i+1} - \mathbf{K}_i$.
$i = i + 1$.
End while
End

The main idea behind PFDeePO is to prevent conditions that may compromise the full rank of Φ, as in the following Theorem.

Theorem 1: Let Φ_i be the matrix constructed at time step i during the execution of PFDeePO. By implementing Algorithm 1, the minimum singular value of Φ_i, denoted as $\underline{\sigma}(\Phi_i)$, satisfies:

$$\underline{\sigma}(\Phi_i) > 0.$$

As a result, the matrix Φ_i attains full rank, i.e., $\text{rank}(\Phi_i) = n + m$. □

The following Lemma and Theorem, show that there exist $\overline{\nu}$ and $\underline{\nu}$ that ensures stability.

Lemma 2: Consider matrices $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{K} \in \mathbb{R}^{m \times n}$, and symmetric positive definite matrices $\mathbf{Q} \in \mathbb{R}^{n \times n}$ and $\mathbf{R} \in \mathbb{R}^{m \times m}$. Then, there exists an interval $\mathscr{V} = [\underline{\nu}, \overline{\nu}]$, where $0 < \underline{\nu} < 1 < \overline{\nu}$, such that

$$\mathbf{Q} - \mathbf{K}^{\mathrm{T}} \big((\nu - 1)^2 \mathbf{B}^{\mathrm{T}} \mathbf{H} \mathbf{B} + (1 - 2\nu) \mathbf{R} \big) \mathbf{K} \geq 0, \quad \forall \nu \in [\underline{\nu}, \overline{\nu}]. \tag{10}$$

□

Theorem 2: Consider a system controlled by Algorithm 1. Let $\mathbf{Q} \in \mathbb{R}^{n \times n}$ and $\mathbf{R} \in \mathbb{R}^{m \times m}$ be given symmetric and positive definite matrices, and let $\beta \in (0, 1)$. Consider the interval $[\underline{\nu}, \overline{\nu}]$, as defined in Lemma 2, such that the inequality (10) holds. Then, the time-varying state feedback control law

$$\mathbf{u}_k = -\nu_k \mathbf{K} \mathbf{x}_k, \tag{11}$$

where ν_k is any sequence taking values in the interval $[\underline{\nu}, \overline{\nu}]$, ensures that the origin of the closed-loop system

$$\mathbf{x}_{k+1} = (\hat{\mathbf{A}} - \nu_k \hat{\mathbf{B}} \mathbf{K}) \mathbf{x}_k, \quad k = 0, 1, \ldots \tag{12}$$

is exponentially stable. □

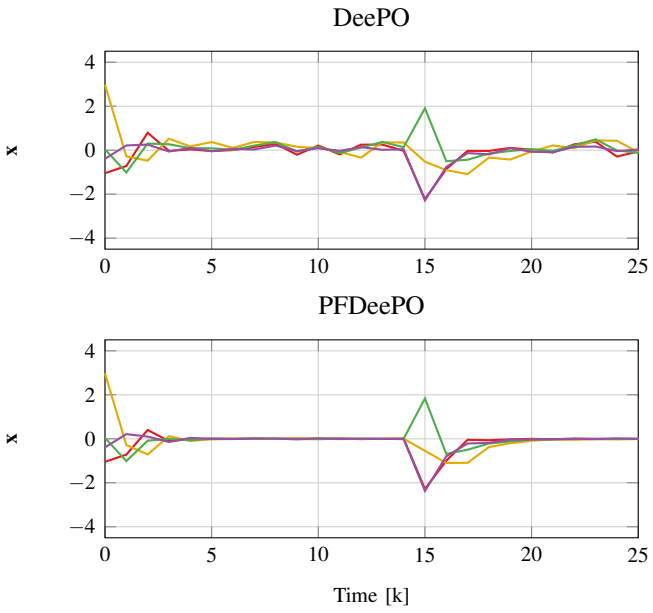

Fig. 1. Evolution of the states $\mathbf{x}$, using DeePO (top) and PFDeePO (bottom).

Figure 1 illustrates the evolution of the system states when the open-loop controllable LTI system from [2] is used in the numerical simulations, with

$$A = \begin{bmatrix} -0.13 & 0.14 & -0.29 & 0.28 \\ 0.48 & 0.09 & 0.41 & 0.30 \\ -0.01 & 0.04 & 0.17 & 0.43 \\ 0.14 & 0.31 & -0.29 & -0.10 \end{bmatrix}, \; B = \begin{bmatrix} 1.63 & 0.93 \\ 0.26 & 1.79 \\ 1.46 & 1.18 \\ 0.77 & 0.11 \end{bmatrix}. \tag{13}$$

At sample $k = 15$, a disturbance is induced in the states using a uniform random value. As observed, once the states reach equilibrium, PFDeePO does not introduce further perturbations. In contrast, the probing noise in DeePO continuously disturbs the states, inducing oscillations that increase control effort.

References

[1] F. Zhao, F. Dörfler, and K. You, "Data-enabled policy optimization for the linear quadratic regulator," in *62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 6160–6165.

[2] F. Zhao, F. Dörfler, A. Chiuso, and K. You, "Data-enabled policy optimization for direct adaptive learning of the LQR," *arXiv preprint arXiv:2401.14871*, 2024.

[3] J. C. Willems, P. Rapisarda, I. Markovsky, and B. L. De Moor, "A note on persistency of excitation," *Systems & Control Letters*, vol. 54, no. 4, pp. 325–329, 2005.