Resilient Direct Data-Driven Control Design under Poisoned Input–State Samples

Mojtaba Kaheni, Niklas Persson, and Alessandro V. Papadopoulos

***Abstract*— This article explores a resilient, data-driven control framework for unknown linear time-invariant (LTI) systems under the threat of poisoned data. Specifically, we consider scenarios in which up to f out of T input–state samples may be manipulated by an adversary aiming to degrade performance or destabilize the system. We begin with a simple example demonstrating that even a single carefully crafted sample can destabilize a closed-loop system, underscoring the vulnerability of direct data-driven control methods and motivating the need for resilience. To address this challenge, we first propose a resilient control design strategy for noise-free settings, based on majority voting across all possible subsets of the input–state dataset with a certain cardinality. We then extend this approach to noisy and disturbed environments, showing that the geometric median of the mentioned subsets provides a resilient solution. Finally, we validate the effectiveness of our framework through numerical simulations.**

I. INTRODUCTION

In classical control theory, designing a controller that achieves desired performance typically requires a model of the plant. Such a model can be obtained either through *first-principle modeling*, which relies on physical laws and domain-specific principles to describe system dynamics, or via *system identification*, which constructs mathematical models from measured data. These models are then used to predict system behavior under various inputs and to design control laws that satisfy performance objectives.

In contrast to classical approaches, the seminal work by Willems *et al.* [1] introduced a fundamentally different perspective. They demonstrated that a linear time-invariant (LTI) system can be fully characterized using a finite set of trajectories generated by a persistently exciting input. As a result, the complete system behavior can be inferred directly from historical input–state data, eliminating the need for explicit modeling or system identification.

This insight has sparked significant interest in the control community, giving rise to a data-driven paradigm that bypasses the traditional, often costly, modeling process. For example, several model-based control design techniques have been adapted to direct data-driven methods in [2]. Data-driven Linear Quadratic Gaussian (LQG) control [3] and Linear Quadratic Regulator (LQR) design [4]–[8] have been studied extensively. Adaptive features have been incorporated into direct data-driven approaches, such as Data-enabled Policy Optimization (DeePO) [9]–[11] and Perturbation-Free DeePO (PFDeePO) [12]. Furthermore, data-driven methods have been validated in practical applications, including control of power converter systems [13], as well as balancing an autonomous bicycle in both simulation and real-world experiments [14], [15].

The foundation of controller design in direct data-driven approaches lies in input-state measurements. A natural concern, therefore, is what happens when these measurements are not fully reliable, either because they can be maliciously manipulated by an adversary seeking to degrade closed-loop performance, or because technical imperfections may cause disruptions or significant inaccuracies. It is important to note that the class of data perturbation discussed here differs from noisy data, which is extensively studied in data-driven control literature [3], [16], [17]. This is because we do not impose restrictions on the stochastic properties of the perturbation or its upper bound. To the best of the authors' knowledge, this concern was first raised in [18], where the authors demonstrated that carefully crafted data-poisoning attacks can severely degrade or even destabilize data-driven control methods. They formulated the problem as a bi-level optimization and validated their analysis through theoretical insights and numerical experiments, though without discussing possible defenses against such attacks. This vulnerability was further confirmed experimentally in a building temperature-control testbed in [19]. Following this seminal work, subsequent research extended the analysis to the vulnerability of predictive [20] and optimal [21]–[23] direct data-driven approaches under poisoning attacks.

In light of the aforementioned studies highlighting the vulnerability of data-driven control designs to poisoning attacks, the need for research on designing *resilient* data-driven controllers against such threats is evident. To the best of the authors' knowledge, the only existing work addressing resiliency in data-driven control design is [24], where the authors propose a design that ensures stability under aperiodic denial-of-service (DoS) attacks. However, resiliency against poisoning attacks remains an open challenge for the research community.

We consider a highly sophisticated setting in which the adversary possesses complete knowledge of the system dynamics and the controller design algorithm, and can also eavesdrop on sensor measurements and control signals. The adversary is further assumed to inject its desired vectors in

This work was supported by the Knowledge Foundation (KKS) with grant "Mälardalen University Automation Research Center (MARC)", n. 20240011.

M. Kaheni is with the Power Consulting group, Hitachi Energy Sweden AB, Evenemangsgatan 17, 169 79 Solna, Stockholm, Sweden. (e-mail: mojtaba.kaheni@hitachienergy.com)

N. Persson and A.V. Papadopoulos are with the Division of Intelligent Future Technologies, Mälardalen University, 721 23 Västerås, Sweden. (e-mails: niklas.persson@mdu.se, alessandro.papadopoulos@mdu.se).

up to f input–state measurements, which are then used for controller design. Inspired by research on resilient distributed optimization [25]–[28] and consensus [29]–[31] in multi-agent systems, our objective is to develop a controller that is accepted by the majority of samples.

A. Statement of Contributions

The main contributions of this article are summarized as follows:

- While prior work has shown that data-driven control design approaches are vulnerable to poisoning attacks targeting *all* data samples, we extend this line of research by demonstrating that even a carefully crafted attack manipulating *a single data sample* can significantly degrade performance, or even destabilize, a closed-loop control system.
- We propose a resilient data-driven control design methodology that guarantees stability in noise- and disturbance-free systems, when a limited number of samples are manipulated.
- We further extend our methodology to systems subject to noise and disturbances by leveraging the well-known robustness properties of the geometric median.

B. Notation

Throughout this paper, unless clearly stated otherwise, the symbols $\mathbb{N}$ and $\mathbb{R}$ denote the sets of integers and real numbers, respectively. Scalars are denoted by lowercase letters, such as x, whereas sets are denoted by uppercase letters, such as X. $\mathbf{x}$ and $\mathbf{X}$ denote a (column) vector and a matrix, respectively. The notation $\mathbf{X} \prec 0$ ($\mathbf{X} \preceq 0$) and $\mathbf{X} \succ 0$ ($\mathbf{X} \succeq 0$) indicates that $\mathbf{X}$ is negative (semi-) definite and positive (semi-) definite, respectively. Additionally, $\mathbf{I}_i$ denotes the $i \times i$ identity matrix, and $\mathcal{N}(\mathbf{x}, \mathbf{X})$ represents a multivariate Gaussian distribution with a mean vector $\mathbf{x}$ and a covariance matrix $\mathbf{X}$. $\binom{a}{b}$ denotes the binomial coefficient. For a set X, we use $|X|$ to denote its cardinality. Finally, we denote by $\mathbf{E}_x \in \mathbb{R}^{m\times n}$ the matrix with all ones in its x-th row and zeros elsewhere, i.e.,

$$(\mathbf{E}_x)_{ij} = \begin{cases} 1, & \text{if } i = x, \\ 0, & \text{otherwise.} \end{cases}$$

II. Preliminaries

In this section, we outline some preliminary concepts to design controllers solely based on historical input–state data of the system. We consider a discrete-time LTI system, represented by

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k, \quad k \in \mathbb{N}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{u} \in \mathbb{R}^m$ denote the state and input vectors, respectively. Consider signals of length T corresponding to the states, inputs, and successor states, which are not required to be consecutive, are defined as follows:

$$\begin{aligned} \mathbf{X}_0 &\triangleq \begin{bmatrix} \mathbf{x}_0, & \mathbf{x}_1, & \cdots, & \mathbf{x}_{T-1} \end{bmatrix}, \\ \mathbf{X}_1 &\triangleq \begin{bmatrix} \mathbf{x}_1, & \mathbf{x}_2, & \cdots, & \mathbf{x}_{T} \end{bmatrix}, \\ \mathbf{U}_0 &\triangleq \begin{bmatrix} \mathbf{u}_0, & \mathbf{u}_1, & \cdots, & \mathbf{u}_{T-1} \end{bmatrix}. \end{aligned} \tag{2}$$

We recall the fundamental result of Willems et al. [1], which provides a necessary condition under which the T-length historical trajectories of $\mathbf{x}_k$ and $\mathbf{u}_k$ adequately characterize the dynamical system in (1). To proceed, we first recall the notion of persistently exciting inputs.

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

$$\mathcal{U}_0 = \begin{bmatrix} \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_{T-l} \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{T-l+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_{l-1} & \mathbf{u}_l & \cdots & \mathbf{u}_{T-1} \end{bmatrix},$$

has full rank ml.

■

The following lemma provides a useful condition for verifying the persistent excitation property of a system.

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

$$\text{rank}(\mathcal{D}) = n + m,$$

where

$$\mathcal{D} \triangleq \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{X}_0 \end{bmatrix}. \tag{3}$$

■

In [2], De Persis and Tesi transformed the state space representation in (1) into a form that exclusively relies on historical data.

Lemma 2 ([2]): If $\mathbf{U}_0$ is persistently exciting, then the system (1) with a state feedback $\mathbf{u} = \mathbf{K}\mathbf{x}$ can be represented by:

$$\mathbf{x}_{k+1} = \mathbf{X}_1 \mathcal{G}_{\mathbf{K}} \mathbf{x}_k, \tag{4}$$

where $\mathcal{G}_{\mathbf{K}}$ is a $T \times n$ matrix that satisfies

$$\begin{bmatrix} \mathbf{K} \\ \mathbf{I}_n \end{bmatrix} = \mathcal{D}\mathcal{G}_{\mathbf{K}}, \tag{5}$$

and as a result

$$\mathbf{u}_k = \mathbf{U}_0 \mathcal{G}_{\mathbf{K}} \mathbf{x}_k. \tag{6}$$

■

From (4), we observe that for the closed-loop discrete system (1) under state-feedback control $\mathbf{u} = \mathbf{K}\mathbf{x}$, the following relation holds:

$$\mathbf{A} + \mathbf{B}\mathbf{K} = \mathbf{X}_1 \mathcal{G}_{\mathbf{K}}. \tag{7}$$

Consequently, one can select a suitable $\mathcal{G}_{\mathbf{K}}$ such that $\mathbf{X}_1\mathcal{G}_{\mathbf{K}}$ satisfies the classical Lyapunov stability condition

$$\mathbf{X}_1 \mathcal{G}_{\mathbf{K}} \mathbf{P} \mathcal{G}_{\mathbf{K}}^\top \mathbf{X}_1^\top - \mathbf{P} \prec 0, \tag{8}$$

for some $\mathbf{P} \succ 0$, thereby ensuring the stability of the closed-loop system. The corresponding controller can then be obtained from (5) as

$$\mathbf{K} = \mathbf{U}_0 \mathcal{G}_{\mathbf{K}}. \tag{9}$$

III. Problem Statement

To formalize the threat model considered in this work, we make the following assumption:

Assumption 1: We consider a poisoning attack that affects at most f time steps in total, where at each attacked time step the adversary may manipulate the state measurements $\mathbf{x}_k$, the input signals $\mathbf{u}_k$, or both.

To preserve generality, we focus on a highly sophisticated adversarial poisoning scenario. Specifically, the adversary is assumed to possess full knowledge of the system dynamics and the control design algorithm, while also being able to eavesdrop on the input signals and state measurements. Thus, the system dynamics, state evolution, and control strategy are entirely transparent (white-box) to the adversary. Moreover, the adversary is free to inject poisoning attacks with arbitrary magnitudes.

To motivate this study and highlight its significance, we present the following numerical example, which demonstrates that even a poisoning attack affecting a single time step among the T samples can potentially invalidate the Lyapunov stability guarantees of the closed-loop system.

Example 1: Consider the discrete-time LTI system in (1) with

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 0.5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{10}$$

We assume that the system dynamics in (10) are unknown. Consequently, the direct data-driven control design based on the Lyapunov inequality in (8) is implemented using the following input-state measurements of the system:

$$\mathbf{U}_0^\top = \begin{bmatrix} -1.28 \\ 0.01 \\ -0.00 \\ -0.89 \\ -0.27 \\ -1.36 \\ -0.14 \\ -0.44 \\ -1.13 \\ -0.47 \\ -0.81 \\ 1.00 \end{bmatrix}, \mathbf{X}_0^\top = \begin{bmatrix} 0.00 & 1.00 \\ -1.28 & 0.50 \\ -1.27 & -1.03 \\ -1.28 & -1.79 \\ -2.17 & -2.17 \\ -2.44 & -3.25 \\ -3.79 & -4.06 \\ -3.94 & -5.83 \\ -4.37 & -6.85 \\ -5.50 & -7.79 \\ -5.97 & -9.40 \\ -6.79 & -10.67 \end{bmatrix}, \mathbf{X}_1^\top = \begin{bmatrix} -1.28 & 0.50 \\ -1.27 & -1.03 \\ -1.28 & -1.79 \\ -2.17 & -2.17 \\ -2.44 & -3.25 \\ -3.79 & -4.06 \\ -3.94 & -5.83 \\ -4.37 & -6.85 \\ -5.50 & -7.79 \\ -5.97 & -9.40 \\ -6.79 & -10.67 \\ -5.79 & -12.12 \end{bmatrix}. \tag{11}$$

Now, suppose that only the state measurement at time step 12 is corrupted, i.e.,

$$\tilde{\mathbf{X}}_1 = \mathbf{X}_1 + \begin{bmatrix} -4.35 \\ -2.90 \end{bmatrix} \mathbf{E}_{12}^\top, \tag{12}$$

while $\mathbf{X}_0$ and $\mathbf{U}_0$ remain unchanged. It should be noted that, in general, modifications in $\mathbf{X}_1$ affect $\mathbf{X}_0$ and vice versa. The only exceptions are when poisoning occurs in the last column of $\mathbf{X}_1$, and first column of $\mathbf{X}_0$.

It can be verified that the matrix

$$\mathcal{G}_{\mathbf{K}} = \begin{bmatrix} 0.00 & 0.00 \\ -0.62 & 0.40 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \\ -0.58 & -0.04 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \\ 0.34 & -0.05 \end{bmatrix}, \tag{13}$$

satisfies (8) with

$$\mathbf{P} = \begin{bmatrix} 1.53 & 0.34 \\ 0.34 & 1.70 \end{bmatrix}, \tag{14}$$

and

$$\tilde{\mathbf{X}}_1 \mathcal{G}_{\mathbf{K}} \mathbf{P} \mathcal{G}_{\mathbf{K}}^\top \tilde{\mathbf{X}}_1^\top - \mathbf{P} = -\mathbf{I}_2. \tag{15}$$

From (5), the corresponding feedback controller is

$$\mathbf{K} = \mathbf{U}_0 \mathcal{G}_{\mathbf{K}} = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{16}$$

However, this controller destabilizes the actual system because the closed-loop matrix

$$\mathbf{A}_{cl} = \mathbf{A} + \mathbf{B}\mathbf{K} = \begin{bmatrix} 2 & 0 \\ 1 & 0.5 \end{bmatrix}, \tag{17}$$

has an eigenvalue outside the unit circle.

IV. Resilient Direct Data-Driven Control Design

In this section, we propose a method for designing a resilient data-driven controller that can withstand the poisoning attack described in Assumption 1. The following assumption is required in our study.

Assumption 2: The outcome of the direct data-driven design is independent of the specific data used for the design. In other words, if two different datasets are employed to design a controller for an identical system, the resulting controller is the same.

The resilient design method presented in this section is independent of the particular control design technique, provided that Assumption 2 holds. For instance, direct data-driven LQR [2], [4], [5], [7], [8] fits naturally into our framework because the solution of the discrete algebraic Riccati equation (DARE) is unique, and therefore the corresponding state-feedback gain is uniquely determined. In contrast, approaches such as pole placement [32] or the Lyapunov-based method in [2] may yield multiple valid controllers for the same system, and are not appropriate for the presented algorithms in this article.

To design a resilient controller, first observe that in consecutive measurements, a manipulation in $\mathbf{x}_k$ corrupts two behavioral samples:

$$\{\mathbf{u}_k, \mathbf{x}_k, \mathbf{x}_{k+1}\} \quad \text{and} \quad \{\mathbf{u}_{k-1}, \mathbf{x}_{k-1}, \mathbf{x}_k\}. \tag{18}$$

Thus, if measurements are poisoned at f time steps, up to $2f$ behavior samples become unreliable. Consequently, among the total of T behavior samples, at least $T-2f$ are guaranteed to remain reliable.

To ensure that adversarially manipulated samples do not affect controller design, we employ the well-known *majority voting* technique [33], [34]. Let us define the set S of all T behavioral samples:

$$S = \{\, \{\mathbf{u}_0, \mathbf{x}_0, \mathbf{x}_1\}, \ldots, \{\mathbf{u}_{T-1}, \mathbf{x}_{T-1}, \mathbf{x}_T\} \,\}. \tag{19}$$

Let B denote the set of all subsets of S that have exactly b elements:

$$B = \{A \subseteq S \mid |A| = b\}. \tag{20}$$

Since S has T elements, the number of subsets of size b (i.e., the cardinality of B) is given by the binomial coefficient:

$$|B| = \binom{T}{b} = \frac{T!}{b!(T-b)!}. \tag{21}$$

While the technical aspects of our proposed approach can be adapted to scenarios where elements in B have input signal sequences that are not persistently exciting, since it is common practice to use white noise as the input in data-driven methods, to avoid unnecessary complexity and confusion, we make the following assumption:

Assumption 3: For all $\beta \in B$, the input signal in β is persistently exciting.

Assumption 3 ensures the feasibility of designing a data-driven controller for each element in B.

We partition the set B into two disjoint groups: B_r, consisting of elements that contain only reliable behavioral samples, and B_a, consisting of elements that include at least one manipulated sample. Clearly,

$$B = B_r \cup B_a. \tag{22}$$

From Assumption 3, we know that any element of B_r is sufficient to capture the system's behavior and, consequently, to design a valid controller.

Since at least $T - 2f$ elements of S are reliable, the cardinality of B_r satisfies

$$|B_r| \;\geq\; \binom{T-2f}{b} \;=\; \frac{(T-2f)!}{b!\,((T-2f)-b)!}. \tag{23}$$

If $|B_r| > |B_a|$, then the majority of controllers designed from subsets in B are based on reliable data, and adversarially manipulated subsets cannot dominate by colluding on a different feedback gain. Building on this observation, the inequality

$$\frac{|B_r|}{|B|} \;\geq\; \frac{\binom{T-2f}{b}}{\binom{T}{b}} \;\geq\; 0.5 \tag{24}$$

ensures that *more than half* of all subsets in B consist entirely of reliable data. Therefore, under Assumption 2, designing a controller from each subset in B and aggregating the resulting controllers through majority voting guarantees that the final controller is based on reliable data. Consequently, the obtained controller is resilient to adversarial manipulations and capable of stabilizing the system.

Algorithm 1 Resilient Feedback Gain Design- Noise and Disturbance Free Dynamics

Require: T (number of samples), f (maximum manipulated samples), S (input–state dataset)

1: Find an integer $b \geq m + n$ such that
$$\frac{\binom{T-2f}{b}}{\binom{T}{b}} \;\geq\; 0.5.$$
2: Construct the set
$$B = \{A \subseteq S \mid |A| = b\}.$$
3: **for** $i = 1$ to $|B|$ **do**
4: $\quad \mathbf{K}_i \leftarrow \textsc{DataDrivenDesign}(B_i)$
5: **end for**
6: $\mathbf{K}^* \leftarrow \textsc{MajorityVoting}(\{K_i\}_{i=1}^{|B|})$
7: **return** K^*

Algorithm 1 provides a step-by-step procedure for computing a resilient feedback gain against up to f manipulated measurements among the total T samples, in the noise-free LTI system defined in (1). In Algorithm 1, DATADRIVENDESIGN$(\cdot)$ denotes any data-driven design algorithm that satisfies the condition specified in Assumption 2, while MAJORITYVOTING$(\cdot)$ refers to the standard majority voting function.

V. Resilient Data-Driven Control Design under Noise and Disturbance

In this section, we address resilient control design in a more realistic setting where both the state evolution and the control input are affected by uncertainties. Specifically, the measurements are corrupted by noise, and the input channel is subject to external disturbances. Consequently, the nominal dynamics in (1) are modified as follows:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k, \qquad k \in \mathbb{N}, \\ \mathbf{u}_k &= \mathbf{u}_k^c + \mathbf{d}_k, \end{aligned} \tag{25}$$

where $\mathbf{w}_k \in \mathbb{R}^n$ denotes the process noise capturing unmodeled dynamics and random perturbations in the state evolution, and $\mathbf{d}_k \in \mathbb{R}^m$ represents the input disturbance accounting for actuator faults, uncertainties, or exogenous inputs affecting the control signal.

The main idea in resilient data-driven control design under noise and disturbances, as modeled in (25), is first to construct the set S as defined in (19), using parameters b and f that satisfy (24). However, in contrast to the noise-free dynamics in (1), when deriving direct data-driven controllers from each element of S, even in attack-free environments and under Assumption 2, the resulting controllers may differ. Consequently, majority voting cannot be applied directly. In this section, we exploit the well-known robustness property of the geometric median to design a data-driven controller that is resilient to the poisoning attacks described in Assumption 1. Theorem 1 states that, when controllers designed from the elements of B are aggregated using the geometric

median, adversarial poisoning cannot *arbitrarily shift* the geometric median achievable in the absence of attacks.

Theorem 1: Let Assumption 2 holds, and $\mathbf{K}_1, \dots, \mathbf{K}_{|B|} \in \mathbb{R}^{m\times n}$ be the controllers designed using each element of B, and let

$$\mathbf{K} = \arg\min_{\mathbf{Z}\in\mathbb{R}^{m\times n}} \sum_{i=1}^{|B|} \|\mathbf{K}_i - \mathbf{Z}\|_F, \tag{26}$$

be their geometric median with respect to the Frobenius norm. Let $\mathbf{K}^*$ be the geometric median when all elements of B are reliable. Now consider the case that f behavioral samples might be poisoned. Partition the indices into a set C_r of "reliable" controllers achieved based on B_r elements and a set C_a of "manipulated" controllers, derived by elements of B_a, with $|C_r| = |B_r|$ and $|C_a| = |B_a|$, where $|B_r| > |B_a|$ and let

$$\mathbf{K}' = \arg\min_{\mathbf{Z}\in\mathbb{R}^{m\times n}} \left(\sum_{\mathbf{K}_i\in C_r} \|\mathbf{K}_i - \mathbf{Z}\|_F + \sum_{\mathbf{K}_i\in C_a} \|\mathbf{K}_i - \mathbf{Z}\|_F \right), \tag{27}$$

be the geometric median with the poisoned data. Define

$$R := \max_{\mathbf{K}_i\in C_r} \|\mathbf{K}_i - \mathbf{K}^*\|_F. \tag{28}$$

If (24) holds, then

$$\|\mathbf{K}' - \mathbf{K}^*\|_F \;\le\; \frac{|B|}{|B| - 2|B_a|}\, R. \tag{29}$$

Proof: If $\mathbf{K}' = \mathbf{K}^*$ the result is trivial. Assume $\mathbf{K}' \neq \mathbf{K}^*$ and define the normalized displacement

$$\mathbf{D} := \frac{\mathbf{K}' - \mathbf{K}^*}{\|\mathbf{K}' - \mathbf{K}^*\|_F} \in \mathbb{R}^{m\times n}.$$

Noting that the optimization problem is convex, the necessary and sufficient condition for optimality at $\mathbf{K}'$ in (27) is that the gradient be 0 at $\mathbf{K}'$. This yields

$$\sum_{\mathbf{K}_i\in C_r} \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} = -\sum_{\mathbf{K}_i\in C_a} \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} \tag{30}$$

For each $\mathbf{K}_i \in C_r$, consider the Frobenius inner product

$$\Big\langle \mathbf{D}, \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} \Big\rangle = \frac{\langle \mathbf{D}, \mathbf{K}' - \mathbf{K}_i\rangle}{\|\mathbf{K}' - \mathbf{K}_i\|_F}. \tag{31}$$

Note that

$$\begin{aligned} \langle \mathbf{D}, \mathbf{K}' - \mathbf{K}_i\rangle &= \langle \mathbf{D}, \mathbf{K}' - \mathbf{K}^*\rangle + \langle \mathbf{D}, \mathbf{K}^* - \mathbf{K}_i\rangle \\ &= \|\mathbf{K}' - \mathbf{K}^*\|_F + \langle \mathbf{D}, \mathbf{K}^* - \mathbf{K}_i\rangle. \end{aligned} \tag{32}$$

By Cauchy–Schwarz,

$$\langle \mathbf{D}, \mathbf{K}^* - \mathbf{K}_i\rangle \ge -\|\mathbf{K}^* - \mathbf{K}_i\|_F \ge -R. \tag{33}$$

Also,

$$\begin{aligned} \|\mathbf{K}' - \mathbf{K}_i\|_F &\le \|\mathbf{K}' - \mathbf{K}^*\|_F + \|\mathbf{K}^* - \mathbf{K}_i\|_F \\ &\le \|\mathbf{K}' - \mathbf{K}^*\|_F + R. \end{aligned} \tag{34}$$

Therefore,

$$\Big\langle \mathbf{D}, \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} \Big\rangle \ge \frac{\|\mathbf{K}' - \mathbf{K}^*\|_F - R}{\|\mathbf{K}' - \mathbf{K}^*\|_F + R}. \tag{35}$$

Algorithm 2 Resilient Feedback Gain Design- Under Noise and Disturbance

Require: T (number of samples), f (maximum manipulated samples), S (input–state dataset)

1: Find an integer $b \ge m + n$ such that

$$\frac{\binom{T-2f}{b}}{\binom{T}{b}} \;\ge\; 0.5.$$

2: Construct the set

$$B = \{A \subseteq S \mid |A| = b\}.$$

3: **for** $i = 1$ to $|B|$ **do**
4: $\quad \mathbf{K}_i \leftarrow \textsc{DataDrivenDesign}(B_i)$
5: **end for**
6: $\mathbf{K}^* \leftarrow \textsc{GeometricMedian}(\{K_i\}_{i=1}^{|B|})$
7: **return** K^*

Summing over $\mathbf{K}_i \in C_r$ and recalling $|C_r| = |B_r|$ gives,

$$\Big\langle \mathbf{D}, \sum_{\mathbf{K}_i\in C_r} \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} \Big\rangle \ge |B_r| \cdot \frac{\|\mathbf{K}' - \mathbf{K}^*\|_F - R}{\|\mathbf{K}' - \mathbf{K}^*\|_F + R}. \tag{36}$$

Since $\|\mathbf{D}\|_F = 1$, and considering (30) we have

$$\begin{aligned} \Big|\Big\langle \mathbf{D}, \sum_{\mathbf{K}_i\in C_r} \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} \Big\rangle\Big| &\le \Big\| \sum_{\mathbf{K}_i\in C_r} \frac{\mathbf{K}' - \mathbf{K}_i}{\|\mathbf{K}' - \mathbf{K}_i\|_F} \Big\|_F \\ &\le |B_a|. \end{aligned} \tag{37}$$

Thus, (36) and (37) yield

$$|B_r| \cdot \frac{\|\mathbf{K}' - \mathbf{K}^*\|_F - R}{\|\mathbf{K}' - \mathbf{K}^*\|_F + R} \le |B_a|. \tag{38}$$

Therefore

$$|B_r|\|\mathbf{K}' - \mathbf{K}^*\|_F - |B_r|R \le |B_a|\|\mathbf{K}' - \mathbf{K}^*\|_F + |B_a|R, \tag{39}$$

which gives

$$\|\mathbf{K}' - \mathbf{K}^*\|_F \;\le\; \frac{|B|}{|B| - 2|B_a|}\, R. \tag{40}$$

■

Remark 1: In the absence of adversarial interference, controllers designed from different selections of behavioral samples are expected to be similar and close to each other. Consequently, the reference controller $\mathbf{K}^*$ remains close to all such designs. This implies that the value of R in (28) is relatively small. In contrast, under an adversarial environment, Theorem 1 shows that controllers obtained from poisoned behavioral samples cannot arbitrarily move the geometric median. The maximum possible displacement is not a function of the adversarial manipulation, but instead scales proportionally with R and $|B_a|$. Finally, it should be emphasized that the bound established in Theorem 1 is conservative; in practice, the actual displacement is expected to be significantly smaller. ■

Algorithm 2 presents the modifications to Algorithm 1 required to ensure resilience against poisoning attacks in systems with noise and disturbances.

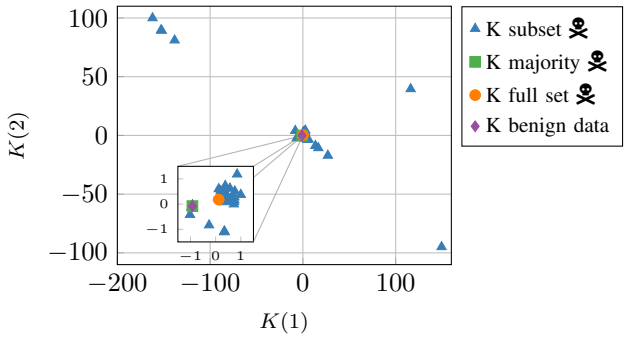

Fig. 1. Direct data-driven LQR gain obtained from noise- and disturbance-free data using the method introduced in [2]. The skull symbol indicates that the dataset is poisoned.

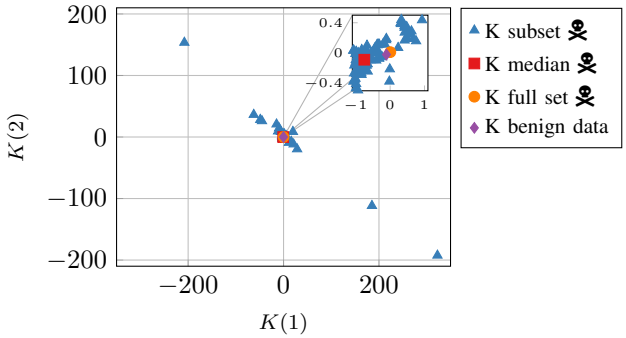

Fig. 2. Direct data-driven LQR gain obtained from noisy and disturbed data using the method introduced in [7]. The skull symbol indicates that the dataset is poisoned.

Remark 2: Both Algorithm 1 and Algorithm 2 necessitate many controller designs, and the number of required designs does not linearly scale with the number of samples. Generally, since the design process is conducted offline, this may be a negligible issue. However, it is advisable to select T not too large, provided it satisfies (23). ■

VI. Numerical Examples

In this section, we present numerical examples to illustrate the effectiveness of both Algorithm 1 and Algorithm 2.

A. Resilient Data-Driven Feedback Gain Design in Noise- and Disturbance-Free Systems

We first consider the problem of designing a resilient data-driven controller using Algorithm 1 for the system and its corresponding poisoning attack described in Example 1, based on the direct data-driven LQR design proposed in [2], with $\mathbf{R} = \mathbf{I}_m$ and $\mathbf{Q} = \mathbf{I}_n$. Without poisoning data, the resulting LQR gain is $\mathbf{K}^{\text{true}} = [-0.9238 - 0.0789]$, while the resulting feedback gain obtained via Algorithm 1 is

$$\mathbf{K} = \begin{bmatrix} -0.9238 & -0.0789 \end{bmatrix}, \tag{41}$$

when poisoning data is considered. Thus, it can be verified that $\mathbf{K}$ coincides with the true feedback gain that would be obtained from uncorrupted data, i.e., before the poisoning attack, as highlighted in Fig. 1. This demonstrates that Algorithm 1 can successfully mitigate the impact of data poisoning in data-driven control design.

By contrast, if the direct LQR design method of [2] is applied to the poisoned dataset, the resulting control gain is

$$\mathbf{K}^{\text{attacked}} = \begin{bmatrix} 0.1199 & 0.1798 \end{bmatrix}, \tag{42}$$

which significantly deviates from $\mathbf{K}^{\text{true}}$ and destabilizes the system.

B. Resilient Data-Driven Feedback Gain Design under Noise and Disturbance

Next, we consider the case where the state and input measurements are corrupted by noise, as in (25), with $d_k = \mathcal{N}(0, 0.05)$ and $\mathbf{w}_k = \mathcal{N}(0, 0.1)$. In the presence of both the introduced poisoning attack and measurement noise, applying Algorithm 2 in conjunction with the direct data-driven LQR method from [7], using $\alpha = 0.01$, yields

$$\mathbf{K} = \begin{bmatrix} -0.7520 & -0.0892 \end{bmatrix}. \tag{43}$$

The actual LQR gain achievable from the noisy but benign data is

$$\mathbf{K}^{\text{true}} = \begin{bmatrix} -0.1074 & -0.0204 \end{bmatrix}. \tag{44}$$

For comparison, the direct implementation of the method in [7] using the poisoned data produces

$$\mathbf{K}^{\text{attacked}} = \begin{bmatrix} 0.0056 & 0.0112 \end{bmatrix}, \tag{45}$$

which is again substantially different from the true gain obtainable from benign data, and destabilizes the system in (10). This emphasizes the critical role of resiliency in data-driven controller design. Fig. 2 depicts a comparison of the achieved LQR gains under noise and disturbance.

Remark 3: By comparing $\mathbf{K}^{\text{true}}$ in the noise- and disturbance-free case with that in their presence, it is evident that noise and disturbances can significantly alter the optimal LQR, even in the absence of attacks. However, the outcome of Algorithm 2 remains much closer to $\mathbf{K}^{\text{true}}$ obtained from noise- and disturbance-free data. This suggests that, in general, Algorithm 2 provides a valuable approach for designing robust data-driven controllers that are resilient to noise and disturbances, even in attack-free scenarios.

VII. Conclusion

Data-driven control approaches have gained significant attention in the control community due to the ease they bring to the control design process. However, poisoning attacks, even over a limited number of samples, can severely compromise the design and may even destabilize the closed-loop system.

In this work, we proposed two algorithms to mitigate the effects of such poisoning attacks. Our simulations further indicate that, even in attack-free scenarios, Algorithm 2 can substantially improve the accuracy of data-driven LQR introduced in [7], in the presence of noise and disturbances.

Future research may extend the results of this paper to other resilient aggregation methods. Additionally, the design

of structured attacks under a limited number of poisoned samples represents an interesting direction for further investigation.

REFERENCES

[1] 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.

[2] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 909–924, 2020.

[3] W. Liu, G. Wang, J. Sun, F. Bullo, and J. Chen, "Learning robust data-based LQG controllers from noisy data," *IEEE Transactions on Automatic Control*, vol. 69, no. 12, pp. 8526–8538, 2024.

[4] F. Dörfler, P. Tesi, and C. De Persis, "On the role of regularization in direct data-driven LQR control," in *IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 1091–1098.

[5] V. G. Lopez, M. Alsalti, and M. A. Müller, "Efficient off-policy Q-learning for data-based discrete-time LQR problems," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 2922–2933, 2023.

[6] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, June 2019, pp. 1300–1309.

[7] C. De Persis and P. Tesi, "Low-complexity learning of linear quadratic regulators from noisy data," *Automatica*, vol. 128, p. 109548, 2021.

[8] F. Dörfler, P. Tesi, and C. De Persis, "On the certainty-equivalence approach to direct data-driven LQR design," *IEEE Transactions on Automatic Control*, vol. 68, no. 12, pp. 7989–7996, 2023.

[9] F. Zhao, F. Dörfler, A. Chiuso, and K. You, "Data-enabled policy optimization for direct adaptive learning of the lqr," *IEEE Transactions on Automatic Control*, vol. 70, no. 11, pp. 7217–7232, 2025.

[10] 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.

[11] X. Wang, F. Zhao, A. Jürisson, F. Dörfler, and R. S. Smith, "Unified aeroelastic flutter and loads control via data-enabled policy optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, no. 5, pp. 11 437–11 449, 2025.

[12] M. Kaheni, N. Persson, V. De Iuliis, C. Manes, and A. V. Papadopoulos, "A modified adaptive data-enabled policy optimization control to resolve state perturbations," in *IEEE 64th Conference on Decision and Control (CDC)*, 2025.

[13] F. Zhao, R. Leng, L. Huang, H. Xin, K. You, and F. Dörfler, "Direct adaptive control of grid-connected power converters via output-feedback data-enabled policy optimization," in *2025 European Control Conference (ECC)*, 2025, pp. 2563–2568.

[14] N. Persson, M. Kaheni, and A. V. Papadopoulos, "A direct data-driven control design for autonomous bicycles," in *IEEE 20th International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 114–120.

[15] N. Persson, F. Zhao, M. Kaheni, F. Dörfler, and A. V. Papadopoulos, "An adaptive data-enabled policy optimization approach for autonomous bicycle control," *IEEE Transactions on Control Systems Technology*, pp. 1–8, 2026.

[16] J. Bongard, J. Berberich, J. Köhler, and F. Allgöwer, "Robust stability analysis of a simple data-driven model predictive control approach," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 2625–2637, 2023.

[17] D. d. S. Madeira and W. B. Correia, "Data-driven saturated state feedback design for polynomial systems using noisy data," *IEEE Transactions on Automatic Control*, vol. 69, no. 11, pp. 7932–7939, 2024.

[18] A. Russo and A. Proutiere, "Poisoning attacks against data-driven control methods," in *2021 American Control Conference (ACC)*, 2021, pp. 3234–3241.

[19] A. Russo, M. Molinari, and A. Proutiere, "Data-driven control and data-poisoning attacks in buildings: the kth live-in lab case study," in *2021 29th Mediterranean Conference on Control and Automation (MED)*, 2021, pp. 53–58.

[20] Y. Yu, R. Zhao, S. Chinchali, and U. Topcu, "Poisoning attacks against data-driven predictive control," in *2023 American Control Conference (ACC)*, 2023, pp. 545–550.

[21] F. Fotiadis, A. Kanellopoulos, K. G. Vamvoudakis, and J. Hugues, "Poisoning actuation attacks against the learning of an optimal controller," in *2024 American Control Conference (ACC)*, 2024, pp. 4838–4843.

[22] H. Sasahara, "Adversarial attacks to direct data-driven control for destabilization," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 7094–7099.

[23] S. C. Anand, "Analysis and mitigation of data injection attacks against data-driven control," *arXiv preprint arXiv:2504.17347*, 2025.

[24] S. Hu, D. Yue, Z. Jiang, X. Xie, and J. Zhang, "Data-driven security controller design for unknown networked systems," *Automatica*, vol. 171, p. 111843, 2025.

[25] M. Kaheni, E. Usai, and M. Franceschelli, "Resilient and privacy-preserving multi-agent optimization and control of a network of battery energy storage systems under attack," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 4, pp. 5320–5332, 2024.

[26] M. Kaheni, M. Lippi, A. Gasparri, and M. Franceschelli, "Selective trimmed average: A resilient federated learning algorithm with deterministic guarantees on the optimality approximation," *IEEE Transactions on Cybernetics*, vol. 54, no. 8, pp. 4402–4415, 2024.

[27] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1063–1076, 2019.

[28] M. Kaheni, E. Usai, and M. Franceschelli, "Resilient constrained optimization in multi-agent systems with improved guarantee on approximation bounds," *IEEE Control Systems Letters*, vol. 6, pp. 2659–2664, 2022.

[29] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 766–781, 2013.

[30] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90–104, 2012.

[31] M. Franceschelli, A. Giua, and A. Pisano, "Finite-time consensus on the median value with robustness properties," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1652–1667, 2017.

[32] G. Bianchin, "Data-driven exact pole placement for linear systems," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 685–690.

[33] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "Intrusion detection in scada based power grids: Recursive feature elimination model with majority vote ensemble algorithm," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2559–2574, 2021.

[34] A. T. J. R. Cobbenhagen, A. Carè, M. C. Campi, F. A. Ramponi, D. J. Antunes, and W. P. M. H. Heemels, "Novel bounds on the probability of misclassification in majority voting: Leveraging the majority size," *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1513–1518, 2021.