Proceedings of Machine Learning Research vol 211:1–13, 2023 5th Annual Conference on Learning for Dynamics and Control

Analysis and Detectability of Offline Data Poisoning Attacks on Linear Dynamical Systems

Alessio Russo ALESSIOR@KTH.SE
Division of Decision and Control Systems, KTH Royal Institute of Technology

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

In recent years, there has been a growing interest in the effects of data poisoning attacks on data-driven control methods. Poisoning attacks are well-known to the Machine Learning community, which, however, make use of assumptions, such as cross-sample independence, that in general do not hold for linear dynamical systems. Consequently, these systems require different attack and detection methods than those developed for supervised learning problems in the i.i.d. setting. Since most data-driven control algorithms make use of the least-squares estimator, we study how poisoning impacts the least-squares estimate through the lens of statistical testing, and question in what way data poisoning attacks can be detected. We establish under which conditions the set of models compatible with the data includes the true model of the system, and we analyze different poisoning strategies for the attacker. On the basis of the arguments hereby presented, we propose a stealthy data poisoning attack on the least-squares estimator that can escape classical statistical tests, and conclude by showing the efficiency of the proposed attack. The code can be found here https://github.com/rssalessio/data-poisoning-linear-systems.

Keywords: data poisoning; data corruption; data-driven control; linear systems.

1. Introduction

Over the past few decades, the rise in computational power, data accessibility, technological progress, and successful results have fueled research in data-driven methods. Nonetheless, these methods may be vulnerable to data poisoning attacks, which aim to degrade performance by altering the training data Biggio et al. (2012); Barreno et al. (2006, 2010). The concept of poisoning was originally introduced for anomaly detection Barreno et al. (2006); Kloft and Laskov (2010); Rubinstein et al. (2009) and attacks against SVM models Biggio et al. (2012). Thereafter, various machine learning models have been shown to be susceptible to poisoning attacks Jagielski et al. (2018); Xiao et al. (2015); Zhang et al. (2020); Shafahi et al. (2018) (see also Tian et al. (2022) for a survey). In the field of systems control, data-driven methods are used due to a variety of reasons, such as decreased modeling complexity and/or reduced costs. In fact, data-driven control methods allow the user to formulate a control law directly from the data, thus bypassing the need of modeling the dynamical systems. There are a variety of these methods to use: techniques based on Willem's et al. lemma Willems et al. (2005); De Persis and Tesi (2019); Coulson et al. (2019), Virtual Reference Feedback Tuning (VRFT) Campi et al. (2002), Iterative Feedback Tuning Hjalmarsson (2002), Correlation-based Tuning Karimi et al. (2004), etc. However, data-driven control methods may also be affected by poisoning attacks. In Russo and Proutiere (2021), the authors formulate a poisoning attack against the VRFT method. Similarly, in Yu et al. (2022) the authors propose a poisoning attack against data-driven predictive control methods. In Showkatbakhsh et al. (2016); Feng and Lavaei (2021), the authors consider how to recover the underlying model of the system from poisoned data, while Chekan and Langbort (2020) considers enlarging the confidence set of the poisoned parameters to improve the performance of online LQR. The detection and analysis of these attacks, however, have received limited attention. Through first principles thinking, we investigate which models are compatible with the poisoned data. We focus on the least-squares (LS) estimator and study how poisoning affects the performance of this estimator. Through the lens of statistical tests, we examine how data can be poisoned and how these attacks can be detected. Our last contribution is to propose an attack that can impact the LS estimator while being stealthy to classical statistical tests, including residual and correlation tests typically used for anomaly detection. We provide examples and numerical simulations to accompany our results.

2. Related work

Data poisoning attacks can be categorized into two classes: *untargeted poisoning attacks* and *targeted poisoning attacks* Tian et al. (2022). Attacks in the former class lead to some form of denial-of-service and try to hinder the convergence of the target model. On the other hand, targeted attacks change the data so that the trained model behaves according to the goal of the attacker Liu et al. (2017); Shafahi et al. (2018). Countermeasures include preventive measures (e.g., encryption) or reactive measures (e.g., detection). In Nguyen and Tran (2013); Bhatia et al. (2017), the authors propose different techniques to recover a linear model from the data for oblivious adversaries, while for adaptive adversaries recovery is possible under some stringent assumptions Bhatia et al. (2015). As mentioned in Bhatia et al. (2017), it seems unlikely that consistent estimators are even possible in face of a fully adaptive adversary.However, to the best of our knowledge, most of these techniques are not directly applicable to control problems, due to the underlying dynamics of the system. In Alfeld et al. (2016), the authors study a targeted attack that poisons the forecasting of an autoregressive model. In Showkatbakhsh et al. (2016), the authors consider the problem of identifying a system whose output measurements have been corrupted by an adversary. They consider an omniscient adversary and given a bound on the number of attacked sensors, and some observability conditions, it is possible to derive a model that is useful for stabilizing the original system. A similar problem is studied in Feng and Lavaei (2021): the authors consider a linear system affected by an unknown sparse adversarial disturbance d_t, and study how to recover the original model of the system. A different problem is studied in Chekan and Langbort (2020), where the authors consider online poisoning of the adaptive LQR method Abbasi-Yadkori and Szepesvári (2011), and, to compensate for the attack, they enlarge the confidence set of the estimator. In Russo and Proutiere (2021) they formulate a bi-level optimization problem to compute poisoning attacks against data-driven control methods. Their attack is then applied to the digital twin of a building to demonstrate the potential of their attack Russo et al. (2021). Lastly, in Yu et al. (2022) they extend the bi-level attack problem to attack data-driven predictive control methods.

3. Preliminaries

Model. We consider a discrete-time LTI system affected by process noise:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \tag{1}$$

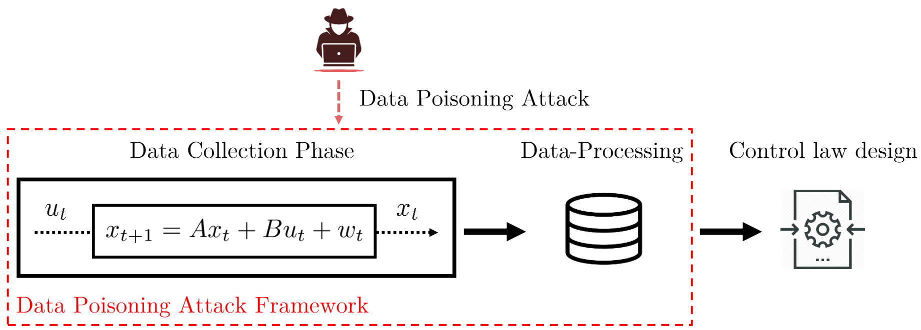

Figure 1: Data poisoning framework for data-driven control methods (figure adapted from Russo and Proutiere (2021)).

where $t \in \mathbb{Z}$ is the discrete time variable, $x_t \in \mathbb{R}^n$ is the state of the system, $u_t \in \mathbb{R}^m$ is the control signal, $A_\star \in \mathbb{R}^{n\times n}, B_\star \in \mathbb{R}^{n\times m}$ are the unknown system matrices, and $w_t \in \mathcal{W} \subseteq \mathbb{R}^n$ is an unmeasured disturbance belonging to some convex set $\mathcal{W}$ (which can be the entire $\mathbb{R}^n$). For a sequence of input-state measurements $\{u_k\}_{k=0}^{T-1}, \{x_k\}_{k=0}^{T}$, we define $\psi_t = (x_t^\top, u_t^\top)^\top \in \mathbb{R}^{n+m}$ and the following data matrices:

$$X_+ := \begin{bmatrix} x_1 & \dots & x_T \end{bmatrix}, \quad X_- := \begin{bmatrix} x_0 & \dots & x_{T-1} \end{bmatrix}, \quad U_- := \begin{bmatrix} u_0 & \dots & u_{T-1} \end{bmatrix},$$

and $X = \begin{bmatrix} x_0 & \dots & x_T \end{bmatrix}, \Psi_- = \begin{bmatrix} \psi_0 & \dots & \psi_{T-1} \end{bmatrix}$. We make the assumption that the user has access to one trajectory of the system, used for identification or data-driven control.

Assumption 1 *The data $\mathcal{D} = (U_-, X)$ available to the user consists of one input-state trajectory of length T. Furthermore, the data satisfy the rank condition* $\operatorname{rank} \Psi_- = n + m$.

This is a standard assumption in data-driven control, and it can be guaranteed for noise-free systems by choosing a persistently exciting input signal of order $n+1$ Willems et al. (2005).

Data poisoning attacks. We denote by $\Delta\mathcal{D} := (\Delta U_-, \Delta X)$ the poisoning signals on the input-state measurements, so that $\Delta U_- = \begin{bmatrix} \Delta u_0 & \dots \Delta u_{T-1} \end{bmatrix} \in \mathbb{R}^{m\times T}$, $\Delta X = \begin{bmatrix} \Delta x_0 & \dots \Delta x_T \end{bmatrix} \in \mathbb{R}^{n\times T+1}$. We let $\tilde{U}_- = U_- + \Delta U_-$, $\tilde{X} = X + \Delta X$ be the resulting poisoned signals. Similarly, we denote the poisoned dataset by $\tilde{\mathcal{D}} := (\tilde{U}_-, \tilde{X})$ and let $\tilde{\psi}_t = \psi_t + \Delta\psi_t$, where $\Delta\psi_t = (\Delta x_t^\top, \Delta u_t^\top)^\top$ (sim. we define $\tilde{\Psi}_-$, and $\Delta\Psi_-$). Attacks in the literature are generally classified as *targeted attacks* or *untargeted attacks*. *Untargeted attacks* just try to alter the performance of the data-driven control scheme, causing a denial-of-service. On the other hand, *targeted attacks* are usually carried out by the attacker to achieve some specific goals, e.g., making the closed-loop system unstable, maximizing the energy used by the system, making the system uncontrollable, etc. The attacker's goal is formulated as a bi-level optimization problem (see also Russo and Proutiere (2021))

$$\max_{(\Delta U_-, \Delta X)\in \mathcal{U}\times\mathcal{X}} \mathcal{A}(\mathcal{D}, K) \text{ s.t. } K \in \arg\min_{K'} \mathcal{L}(U_- + \Delta U_-, X + \Delta X, K'), \tag{2}$$

where $(\mathcal{U}, \mathcal{X})$ is a convex set, K is the closed-loop controller, $\mathcal{A}$ represents the objective function of the malicious agent, and $\mathcal{L}$ represents the function used by the victim to compute the control law K according to the poisoned data $(\tilde{U}_-, \tilde{X})$.

4. Attacks and Detection Strategies

In this section we examine what is the set of pairs (A, B) that are compatible with the poisoned data $\tilde{\mathcal{D}}$, and establish a sufficient and necessary condition for $(A_\star, B_\star)$ to be compatible with $\tilde{\mathcal{D}}$. We investigate how least-squares estimate changes under poisoning, examine attack detection, and propose a stealthy untargeted attack on the least-squares estimate.

4.1. The set of compatible models under data poisoning

Most data-driven control methods assume that there exists a linear system (A, B) that is consistent with the data. Ignoring the noise term w_t in eq. (1), consistency amounts to finding all $(A, B) \in \mathbb{R}^{n\times(n+m)}$ that satisfy the equation $X_+ = AX_- + BU_-$ (*i.e.*, all the pairs consistent with the dataset). In presence of a noise signal w_t, consistency is derived from the following relationship

$$\begin{bmatrix} I_n & -A_\star & -B_\star \end{bmatrix} \begin{bmatrix} X_+ \\ \Psi_- \end{bmatrix} = W_-, \tag{3}$$

where $W_- = \begin{bmatrix} w_0 & w_1 & \dots & w_{T-1} \end{bmatrix} \in \mathcal{W}^T$ (where $\mathcal{W}^T := \bigtimes_{n=1}^{T} \mathcal{W}$). Then, given a poisoned dataset $\tilde{\mathcal{D}} = (\tilde{U}_-, \tilde{X})$, we wonder for which $(A, B) \in \mathbb{R}^{n\times(n+m)}$ and $\tilde{W} \in \mathcal{W}^T$ the following relationship holds

$$\begin{bmatrix} I_n & -A & -B \end{bmatrix} \begin{bmatrix} \tilde{X}_+ \\ \tilde{\Psi}_- \end{bmatrix} = \tilde{W}_-. \tag{4}$$

To answer this question, we seek the set of noise sequences $\mathcal{W}_{\tilde{\mathcal{D}}}$ that are compatible with the data $\tilde{\mathcal{D}}$. Following a similar approach as in Koch et al. (2020), we note that the compatible noise terms $\{\tilde{w}_t\}_t$ belong to the image of $\tilde{G} := \begin{bmatrix} \tilde{X}_+^\top & \tilde{\Psi}_-^\top \end{bmatrix}^\top$. Straightforwardly, if $\tilde{W}_-\tilde{G}^\perp = 0$, then $\tilde{W}_-$ is compatible with the data (where $\tilde{G}^\perp$ denotes a basis of the kernel of $\tilde{G}$), and $\mathcal{W}_{\tilde{\mathcal{D}}}$ is

$$\mathcal{W}_{\tilde{\mathcal{D}}} = \left\{ \tilde{W}_- \in \mathcal{W}^T : \tilde{W}_-\tilde{G}^\perp = 0 \right\}. \tag{5}$$

Therefore, the following result characterizes in which cases $(A_\star, B_\star) \in \Sigma_{\tilde{\mathcal{D}}}$.

Lemma 1 [1] *Consider a poisoned dataset $\tilde{\mathcal{D}}$. The set of all pairs (A, B) that are consistent with the data $\tilde{\mathcal{D}}$ is $\Sigma_{\tilde{\mathcal{D}}} := \left\{ (A, B) \in \mathbb{R}^{n\times n+m} : \exists \tilde{W}_- \in \mathcal{W}_{\tilde{\mathcal{D}}} : (4) \text{ holds for } (A, B, \tilde{W}_-) \right\}$. Let $\tilde{W}_\star = W_- + \Delta W_-$, where $\Delta W_- = \Delta X_+ - A_\star \Delta X_- - B_\star \Delta U_-$. Then $(A_\star, B_\star) \in \Sigma_{\tilde{\mathcal{D}}} \iff \tilde{W}_\star \in \mathcal{W}^T$.*

In most applications, $\mathcal{W}$ is considered to be $\mathbb{R}^n$ itself, which essentially guarantees that $(A_\star, B_\star) \in \Sigma_{\tilde{\mathcal{D}}}$. However, if the disturbance $\{w_t\}_{t\geq 0}$ is generated according to some probability measure $\mathbb{P}$, then the likelihood of $\tilde{W}_\star$ may be small under $\mathbb{P}$ depending on the attack. In other applications, $\mathcal{W}$ is a known bounded convex set, and may not contain the sequence $\tilde{W}_\star$. In this case, it is necessary to enlarge $\mathcal{W}$ to be able to recover the original model. Lastly, in some scenarios the user may know in advance an over-approximate $\hat{\Sigma}_{\mathcal{D}}$ of $\Sigma_{\mathcal{D}}$, which can be used to infer if the data has been poisoned in case $\hat{\Sigma}_{\mathcal{D}}$ and $\Sigma_{\tilde{\mathcal{D}}}$ are too different (using, for example, Bayesian hypothesis testing).

This result can also be interpreted in the following way: if $\mathcal{W}$ is a bounded convex set, an attacker may try to bound ΔX and ΔU_- to bound ΔW_-, and, consequently, act on the compatibility

1. Refer to the technical report https://arxiv.org/abs/2211.08804 for all the proofs and numerical details.

of the true system matrices. Obviously, upper bounding the norm of the poisoning signals seems like a good way to make sure that is less detectable. However, just bounding the norm of the poisoning signals, as we see in the forthcoming sections, is not enough to achieve undetectability of an attack.

4.2. Attack strategies for the least-squares estimator

To better understand how to formulate attack strategies, and analyze the problem of detectability, we now turn our attention to the least-squares (LS) estimator. As pointed out in De Persis and Tesi (2019), this LS estimate is used in the formulation of data-driven controllers. The unpoisoned LS estimate can be compactly written as $(A_{\text{LS}}, B_{\text{LS}}) = X_+\Psi_-^\dagger$ ($\Psi_-^\dagger$ is the right inverse of Ψ_-). We also denote by $(A_{\text{LS}}, B_{\text{LS}})$ and $(\tilde{A}_{\text{LS}}, \tilde{B}_{\text{LS}})$ the LS estimates when, respectively, $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are used. Furthermore, let $(\Delta\tilde{A}_{\text{LS}}, \Delta\tilde{B}_{\text{LS}}) = (\tilde{A}_{\text{LS}} - A_\star, \tilde{B}_{\text{LS}} - B_\star)$ the difference between the LS estimate and the true parameter, and indicate by $\Delta\tilde{\theta}_{\text{LS}} = \text{vec}(\Delta\tilde{A}_{\text{LS}}, \Delta\tilde{B}_{\text{LS}})$ its vectorization. We further assume that $\text{rank}\,\tilde{\Psi}_- = n + m$. In presence of a poisoning attack $(\Delta X, \Delta U_-)$ we obtain the following straightforward result, which is used to discuss possible attack strategies.

Lemma 2 *The LS error is given by* $(\Delta\tilde{A}_{LS}, \Delta\tilde{B}_{LS}) = \tilde{W}_\star\tilde{\Psi}_-^\dagger$*, where* $\tilde{W}_\star$ *is as in lemma 1. In addition to that, we have* $\sigma_{min}(W_-) + \sigma_{min}(\Delta W_-) \leq \|\tilde{\Psi}_-\|_{\text{F}}\|\Delta\tilde{\theta}_{LS}\|_2 \leq \sigma_{max}(W_-) + \sigma_{max}(\Delta W_-)$*, where* $\|\cdot\|_{\text{F}}$ *indicates the Frobenius norm and* $\sigma_{min}, \sigma_{max}$ *the minimum and maximum singular values.*

Relationship between poisoning and exploration. This result provides a way to formulate possible attack strategies and to analyze their impact. The adversary clearly needs to minimize the amount of exploration, quantified by the term $\|\tilde{\Psi}_-\|_{\text{F}} = \sum_{t=0}^{T-1}\|\tilde{\psi}_t\|_2$ to maximize the error. Fundamentally, any attack wishing to maximize the error needs to change the data as to minimize the exploration performed by the victim. As an informal argument, define $\tilde{C}_T = \sum_{t=0}^{T-1}(\tilde{\psi}_t \otimes I_n)(\tilde{\psi}_t \otimes I_n)^\top$ and introduce the *unexcitation subspace* $\tilde{\mathcal{U}} = \left\{\theta \in \mathbb{R}^{n(n+m)} : \limsup_{T\to\infty} \theta^\top \tilde{C}_T \theta < \infty\right\}$ (see Bittanti et al. (1992)), and let $\tilde{\mathcal{E}}$ be its orthogonal complement (the data generation process is undefined on purpose, since it is just an illustrative argument). Denote by $\Delta\tilde{\theta}_{\text{LS}}^{\tilde{\mathcal{E}}}$ and $\Delta\tilde{\theta}_{\text{LS}}^{\tilde{\mathcal{U}}}$ the orthogonal projections of $\Delta\tilde{\theta}_{\text{LS}}$ on these two subspaces, so that $\Delta\tilde{\theta}_{\text{LS}} = \Delta\tilde{\theta}_{\text{LS}}^{\tilde{\mathcal{E}}} + \Delta\tilde{\theta}_{\text{LS}}^{\tilde{\mathcal{U}}}$. Then, under some simple assumptions, it is possible to show that asymptotically $\|\Delta\tilde{\theta}_{\text{LS}}\|_2 = \|\Delta\tilde{\theta}_{\text{LS}}^{\tilde{\mathcal{U}}}\|_2$. Hence, maximizing $\|\Delta\tilde{\theta}_{\text{LS}}\|$ amounts to changing the unpoisoned regressor ψ_t so that, the unexcitation subspace becomes "*larger*", which implies that the amount of exploration is lowered. To formalize the concept, let $\{v_i\}_{i=1}^{n(n+m)}$ be an orthonormal basis of $\mathbb{R}^{n(n+m)}$ with $v_1 = \theta_{\text{LS}}/\|\theta_{\text{LS}}\|_2$, where $\theta_{\text{LS}} = \text{vec}(A_{\text{LS}}, B_{\text{LS}})$ is the true LS-estimate for an unpoisoned dataset. Then, the estimation error in the direction of v_k is given by $|v_k^\top \Delta\tilde{\theta}_{\text{LS}}|$, which is lower bounded as follows.

Corollary 3 *For any* $k = 1, \ldots, n(n+m)$*, and a poisoned dataset* $\tilde{\mathcal{D}}$

$$\sqrt{(v_k^\top \Delta\tilde{\theta}_{LS})^2} \geq \frac{|\cos(\alpha_k)|(\sigma_{min}(W_-) + \sigma_{min}(\Delta W_-))}{\|V_k\tilde{\Psi}_-\|_{\text{F}}}, \tag{6}$$

where $V_k = \text{vec}_{n,n+m}^{-1}(v_k)$[2]*, and* α_k *is the angle between* $\text{vec}(V_k\tilde{\Psi}_-)$ *and* $\text{vec}(\tilde{W}_\star)$*.*

The term $\|V_k\tilde{\Psi}_-\|_{\text{F}}$ can be interpreted as the total amount of exploration in the direction of v_k. Minimizing the exploration in the direction of the true estimate θ_{LS} implies a larger value of $|v_1^\top \Delta\tilde{\theta}_{\text{LS}}|$,

2. $\text{vec}_{ab}^{-1}(x)$ reshapes a vector $x \in \mathbb{R}^{ab}$ into a matrix of size $a \times b$ by arranging the elements of x column-wise.

which is larger when the estimator error $\Delta\tilde{\theta}_{\text{LS}}$ is parallel to θ_{LS}, from which we deduce that asymptotically the unexcitation subspace includes the unpoisoned estimate θ_{LS}.

These results not only shed a light on the mechanics of poisoning, but also help us define a possible attack. The attacker can compute some poisoning signals $(\Delta X, \Delta U_-)$ by solving the convex problem $\min_{(\Delta X, \Delta U_-)\in\mathcal{C}} \|\tilde{\Psi}_-\|_{\text{F}}$ (or $\min_{(\Delta X, \Delta U_-)\in\mathcal{C}} \|V_1\tilde{\Psi}_-\|_{\text{F}}$), where $\mathcal{C}$ is some convex set. Nevertheless, this simple attack ignores other terms, such as ΔW_-, and therefore may not be enough to significantly impact the LS estimate. As we discuss in the next sections, maximizing the norm of the LS residuals fills this gap. Furthermore, as we see, an attacker needs to impose additional constraints on the optimization problem to make an attack stealthy. To that aim, we begin by discussing how statistical hypothesis testing can help to detect poisoning attacks.

4.3. Detection analysis for the LS estimator

The detection of any attack should be based on two important ingredients: (1) prior knowledge of the system and its signals; (2) independent statistical tests. Prior knowledge is useful to detect possible wrongdoings, however, that knowledge may be biased. Therefore, it is important to complement tests based on prior knowledge with tests that are independent of that knowledge. In the following, we relate poisoning attacks to classical statistical tests. We begin our study by considering attacks on the input signal, and then consider attacks on the state signal as well.

4.3.1. DETECTION OF ATTACKS ON THE INPUT SIGNAL

The statistical properties of the input signal are usually assumed to be known. In fact, in most experiments, the input is usually chosen as a white noise signal, as to excite the dynamics of the system. Assuming $\{u_t\}_t$ is a sequence of i.i.d. random variables with distribution P, whiteness tests Box et al. (2015); Drouiche (2000) can be used to deduce if the samples in $\{\tilde{u}_t\}_t$ are white, while one-sample tests (such as the Kolmogorov-Smirnov test Massey Jr (1951), or the Anderson-Darling test Nelson (1998)) can be used to assess whether the samples in $\{\tilde{u}_t\}_t$ are distributed according to P. More simply, if $u_t \sim \mathcal{N}(0, I_m)$, then $\|U_-\|_{\text{F}}^2 \sim \chi^2(Tm)$ is a Chi-squared distribution with Tm degrees of freedom. Consequently, for a small $\delta \geq 0$ we see the constraint $\|\Delta U_-\|_{\text{F}} \leq \delta\|U_-\|_{\text{F}}$ as a way to constraint the Chi-squared statistics. Along this reasoning, an important class of input attacks can be derived when $(u_t, \tilde{u}_t)$ are indistinguishable, i.e., statistically equivalent.

Definition 4 *Suppose that u_t is i.i.d., distributed according to P for every $t \geq 0$. Then, $(u_t, \tilde{u}_t)$ are indistinguishable if $\tilde{u}_t$ is i.i.d. and distributed according to P for every $t \geq 0$.*

To illustrate the attack, consider the following example.

Example 1 *Consider the system $x_{t+1} = a_\star x_t + b_\star u_t + w_t$, with $(a_\star, b_\star) = (0.7, 0.5)$ and $w_t \sim \mathcal{N}(0, 1)$. In fig. 2 are shown the confidence intervals for the LS estimate when the input has been poisoned according to the indistinguishable attack $\tilde{u}_t$, where $(u_t, \tilde{u}_t)$ are independent, i.i.d., and distributed according to $\mathcal{N}(0, 1)$. Under poisoning, the estimate of $b_\star$ converges to 0 as $T \to \infty$.*

This insight leads us to the following result.

Lemma 5 *Let $\{u_t\}_{t=0}^{T-1}$ be an i.i.d. sequence distributed according to $\mathcal{N}(0, \Sigma_u)$. Assume $\Delta u_t = -u_t + a_t$ to be an indistinguishable attack, with $a_t \sim \mathcal{N}(0, \Sigma_u)$ independent of u_t, and $\Delta x_t = 0, t \geq 0$. Then, if A is stable, $\Delta\tilde{B}_{LS} \to -B_\star$ w.p. 1 as $T \to \infty$.*

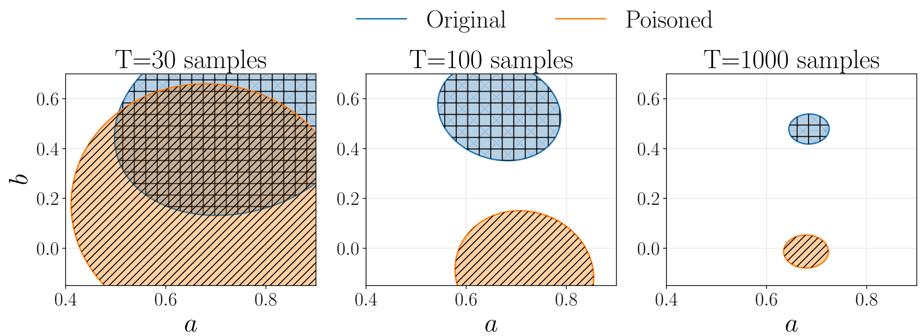

Figure 2: Example of input poisoning (see example 1). The colored ellipses depict the 95% confidence interval of the LS estimates of the true parameters $(a_\star, b_\star)$. When the input data is poisoned, we obtain an F-statistic of $Z_{\tilde{\mathcal{D}}} \approx (0.45, 0.62, 0.15)$ respectively for $T = (15, 100, 1000)$ samples. Otherwise, we obtain $Z_{\mathcal{D}} \approx (7.35, 29.05, 233.35)$.

Intuitively, if the poisoned input is completely uncorrelated from the data, then the best estimate of $B_\star$ is 0. Simply, as the number of samples grows larger, $B_\star + \Delta\tilde{B}_{\text{LS}} \to 0$.

To detect this type of attack, we propose to test the explanatory power of the input data. Using classical partial F-tests Kleinbaum et al. (2013), we test the hypothesis $H_0 : \|B_\star\| = 0$ against $H_1 : \|B_\star\| \neq 0$. Assume the underlying system is affected by some process noise $w_t \sim \mathcal{N}(0, \Sigma_w)$. Denote by $\tilde{A}_{\text{LS}}^{(1)}$ the LS estimate of $A_\star$ when the input signal $\tilde{u}_t$ is not used by the LS estimator. Similarly, denote by $(\tilde{A}_{\text{LS}}^{(2)}, \tilde{B}_{\text{LS}}^{(2)})$ the LS estimate when $\tilde{u}_t$ is considered in the estimation process. Consider the LS residuals, and define the statistic

$$Z_{\tilde{\mathcal{D}}} := \frac{\left\|X_+ - \tilde{A}_{\text{LS}}^{(1)} X_-\right\|_{\text{F}}^2 - \left\|X_+ - \tilde{A}_{\text{LS}}^{(2)} X_- + \tilde{B}_{\text{LS}}^{(2)} \tilde{U}_-\right\|_{\text{F}}^2}{(nm/(T - n(n+m) - 1)) \left\|X_+ - \tilde{A}_{\text{LS}}^{(2)} X_- + \tilde{B}_{\text{LS}}^{(2)} \tilde{U}_-\right\|_{\text{F}}^2}. \tag{7}$$

Under H_0 it can be shown that the statistics $Z_{\tilde{\mathcal{D}}}$ follows an F distribution with $(nm, T - n(n + m) - 1)$ degrees of freedom (follows from an application of (Ljung, 1998, Lemma II.4)). Using this partial F-test, we reject H_0 if and only if $Z_{\tilde{\mathcal{D}}} > f^\alpha_{nm, T-n(n+m)-1}$, where $f^\alpha_{a,b}$ is the upper α-point of an F distribution with (a, b) degrees of freedom. In conclusion, not rejecting H_0 may indicate that the input data has been poisoned. Clearly, for more complex cases, we need to resort to other tools, such as the analysis of the residuals, as explained in the following section.

4.3.2. RESIDUAL ANALYSIS

We claim that any attacker that wishes to remain stealthy needs to make sure that the residuals of the LS procedure satisfy certain statistical conditions. We begin by deriving the following bound on the residuals of the LS estimate for generic attacks that are independent of the noise signal $\{w_t\}_t$ (this includes the class of oblivious attacks). Let the residual of $(\tilde{A}_{\text{LS}}, \tilde{B}_{\text{LS}})$ at time t be $\tilde{R}_t = \tilde{x}_{t+1} - \tilde{A}_{\text{LS}}\tilde{x}_t - \tilde{B}_{\text{LS}}\tilde{u}_t$. In matrix notation, we write $\tilde{R} = \begin{bmatrix}\tilde{R}_0 & \tilde{R}_1 & \dots & \tilde{R}_{T-1}\end{bmatrix} = \tilde{X}_+ - \begin{bmatrix}\tilde{A}_{\text{LS}} & \tilde{B}_{\text{LS}}\end{bmatrix}\tilde{\Psi}_-$ (similarly, we denote by R the residuals in absence of poisoning).

Lemma 6 *Assume $\{(\Delta x_t, \Delta u_t)\}_t$ to be independent of the i.i.d. noise sequence $\{w_t\}_t$, with $w_t \sim \mathcal{N}(0, \Sigma_w)$. Then, the MSE $\mathbb{E}\left[\|\tilde{R}\|_{\rm F}^2\right]$ satisfies $0 \leq \mathbb{E}\left[\|\tilde{R}\|_{\rm F}^2 - \|R\|_{\rm F}^2\right] \leq \mathbb{E}[\sum_i \sigma_i^2(\Delta W_-)]$, where σ_i is the i-th singular value. Furthermore, $\|R\|_{\rm F}^2$ is a quadratic form of a normal random vector, distributed according to $\sum_{i=1}^n \lambda_i \chi^2(T-n-m)$, with λ_i being the i-th eigenvalue of Σ_w.*

In addition, observe the following lemma on the sensitivity of the residuals.

Lemma 7 (Sensitivity) *For any fixed attack satisfying the rank condition $\operatorname{rank} \tilde{\Psi}_- = n+m$, we obtain the following sensitivity on the residuals $\frac{\|\tilde{R}-R\|_{\rm F}}{\|R\|_{\rm F}} \leq \|\Delta\Psi_-\|_2 \|\tilde{\Psi}_-\|_{\rm F}^{-1}$.*

Lemma 6 and 7 link the problem of maximizing the LS error to that of minimizing the amount of exploration $\|\tilde{\Psi}_-\|_{\rm F}$ (as discussed in sec. 4.2) as well as maximizing the singular values of ΔW_-. Furthermore, Lemma 6 can be used to formulate a possible detection test on the variance of the residuals. In fact, we note that any attack independent of the noise will necessarily increase the variance of the residuals. This observation provides us a hint to test the variance of the residuals. It is possible to derive a two-tail test on the variance of the residuals (as long as the user knows has some knowledge on the covariance of the noise) to verify that the data has not been poisoned. The user can test whether $\|\tilde{R}\|_{\rm F}^2$ belongs to the range $Q_{(1-\alpha)/2, T-n-M}(\lambda_1, \dots, \lambda_n) \leq \|\tilde{R}\|_{\rm F}^2 \leq Q_{\alpha/2, T-n-M}(\lambda_1, \dots, \lambda_n)$, where $Q_{x,d}(\lambda_1, \dots)$ is the critical value of the distribution $\sum_{i=1}^n \lambda_i \chi^2(d)$ with significance $x \in (0,1)$. Before we continue with an example, consider that the assumption of independence between $(\Delta U_-, \Delta X)$ and W_- may not be always satisfied: if the attacker has access to the dataset $\mathcal{D}$, then it is likely that she uses X, which depends on W_-, to compute the attack vector $(\Delta U_-, \Delta X)$. In other cases, for example, when the attacker has limited capabilities on the dataset and/or the poisoned sensors, the assumption of independence is more likely to hold. Similarly, the assumption holds whenever the attacker is executing an attack that has been computed on a different set of data. Moreover, from simulations, it seems that this test is still valid to detect a possible adaptive attack.

Example 2 (Untargeted attack) *Consider an attacker that maximizes the norm of the residuals $\|\tilde{R}\|_F$ as a proxy to maximize $\left\|\begin{bmatrix}\Delta\tilde{A}_{LS} & \Delta\tilde{B}_{LS}\end{bmatrix}\right\|_2$. Let $\delta \geq 0$ be a parameter that limits the amplitude of the poisoning signals, and define $\Delta\tilde{W}_- := \Delta X_+ - A_{LS}\Delta X_- - B_{LS}\Delta U_-$. Then, as detailed in the appendix, the adversary can solve the following concave problem to compute an attack:*

$$\max_{\Delta U_-, \Delta X} \quad Tr\left((\Delta\tilde{W}_- + 2R)\Delta\tilde{W}_-^\top\right) \quad s.t. \quad \|\Delta X\|_{\rm F} \leq \delta\|X\|_{\rm F}, \|\Delta U_-\|_{\rm F} \leq \delta\|U_-\|_{\rm F}. \tag{8}$$

Consider the 4-dimensional system used in Russo and Proutiere (2021), affected by white noise with standard deviation $\sigma_w = 0.1$. The victim has collected $T = 200$ samples using $u_t \sim \mathcal{N}(0,1)$. In fig. 3 are shown the results when (8) is solved using (1) convex-concave programming (CCP), (2) the cross-entropy method De Boer et al. (2005) and (3) random sampling from a Gaussian distribution (check the appendix for details). Note that the attacks can be easily detected for small values of δ using the test on the residuals (central plot). On the right plot, observe that, as discussed after corollary 3, the vectors θ_{LS} and $\Delta\tilde{\theta}_{LS}$ tend to align with each other when the attack is impactful.

This last example indicates that minimizing the norm of the poisoning signal is not enough to minimize detectability. Even though the poisoned signal and the unpoisoned one are similar, maximizing the MSE greatly affects the distribution of the residuals. Thereby, it may be more beneficial for the

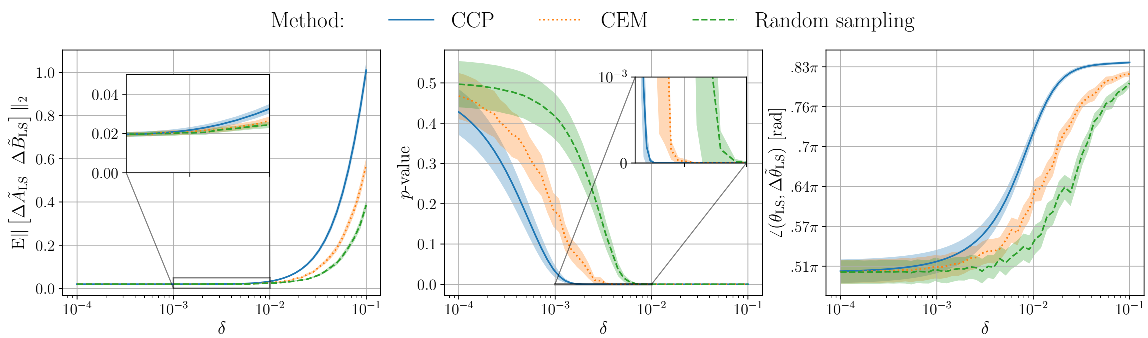

Figure 3: Untargeted attack that maximizes the MSE. The curves indicate an average and its 95% confidence interval over 100 runs. From left to right : (1) impact on the LS estimate; (2) the p-value (right-tail) under the assumption of white noise with $\sigma_w = 0.1$; (3) the angle between the unpoisoned estimate θ_{LS} and the error of the poisoned estimate $\Delta\tilde{\theta}_{\text{LS}}$.

adversary to directly maximize $\| \begin{bmatrix} \Delta\tilde{A}_{\text{LS}} & \Delta\tilde{B}_{\text{LS}} \end{bmatrix} \|_2$ (which is a non-convex problem), while constraining the residuals of the models, to decrease detectability. A hint comes from the fact that the noise term at time t is $\tilde{w}_t = w_t + \Delta w_t = w_t + \Delta x_{t+1} - A_\star \Delta x_t - B_\star \Delta u_t$. Since the noise $\tilde{w}_t$ depends on Δx_{t+1}, the victim can expect to observe a large value in the correlation of the residuals at lag 1. This last observation suggests that an adversary may consider constraining the correlation of the residuals to reduce the detectability.

Correlation tests. Consider white process noise $w_t \sim \mathcal{N}(0, \Sigma_w)$, and let $\tilde{C}_\tau = \frac{1}{T}\sum_{t=0}^{T-\tau-1} \tilde{R}_t \tilde{R}_{t+\tau}^\top$ be the sample correlation of the residuals at lag τ. Under the null hypothesis that the data has not been poisoned, asymptotically we obtain $\sqrt{T}\,\mathrm{vec}\, C_\tau \sim \mathcal{N}(0, \Sigma_w \otimes \Sigma_w)$ (from an application of (Ljung, 1998, Lemma 9.A1)), from which we derive the statistics $T\|C_\tau C_0^{-1}\|_{\text{F}}^2 \sim \chi^2(n^2)$. Similarly, following a similar approach as in Hosking (1980), it is possible to derive the asymptotic Portmanteau statistics to test the whiteness of the residuals $T\sum_{\tau=1}^{T} \|C_\tau C_0^{-1}\|_{\text{F}}^2 \sim \chi^2(n^2(T-n-m))$. Using these statistics, it is possible to formulate a stealthy attack, as explained in the next section.

4.4. Stealthy untargeted attack

On the basis of the previous findings, we argue that the main quantities of interest to make a poisoning attack stealthy are (1) the norm of the poisoning signals; (2) $\|\tilde{R}\|_{\text{F}}^2$ the norm of the residuals; (3) $\|\tilde{C}_\tau \tilde{C}_0^{-1}\|_{\text{F}}^2, \tau = 1, \dots, T-1$, the norm of the self-normalized correlation terms. Consequently, we propose the following optimization problem to compute poisoning stealthy untargeted attacks:

$$\max_{\Delta U_-, \Delta X} \quad \left\| \begin{bmatrix} \Delta\tilde{A}_{\text{LS}} & \Delta\tilde{B}_{\text{LS}} \end{bmatrix} \right\|_2 \text{ s.t. } g_i(\mathcal{D}, \Delta U, \Delta X) \le \delta_i, \quad i = 0, 2, \dots, s+3, \tag{9}$$

with $g_0 = \|\Delta X\|_{\text{F}}/\|X\|_{\text{F}}$, $g_1 = \|\Delta U\|_{\text{F}}/\|U\|_{\text{F}}$, $g_2 = \left|1 - \|\tilde{R}\|_{\text{F}}^2/\|R\|_{\text{F}}^2\right|$, $g_3 = |1 - Z_{\tilde{\mathcal{D}}}/Z_{\mathcal{D}}|$ and $g_{3+\tau} = \left|1 - \|\tilde{C}_\tau \tilde{C}_0^{-1}\|_{\text{F}}^2/\|C_\tau C_0^{-1}\|_{\text{F}}^2\right|, \tau = 1, \dots, s$, for some $s < T-1$. Intuitively, the constraints limit the relative change of each quantity.

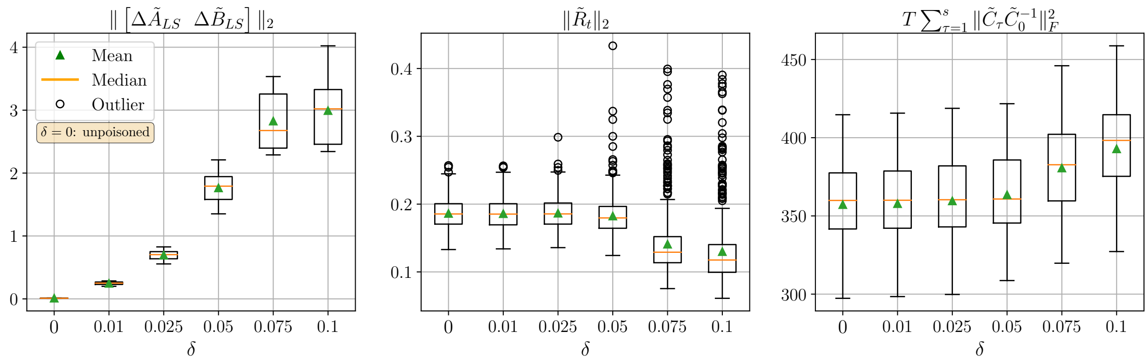

Figure 4: Stealthy attacks on the system used in example 2. From left to right: distribution over 10 different seeds of the impact on the LS error, $\|\tilde{R}_t\|_2$ and $T\sum_{\tau=1}^{s}\|\tilde{C}_\tau\tilde{C}_0^{-1}\|_{\mathrm{F}}^2$.

Numerical results. We applied the attack resulting from (9) on the system used in example 2, with $T = 500$, $s = 25$ and the same value of δ for all constraints. Since $n = 4, m = 1$, there are 2504 variables to optimize. The optimization problem is non-convex, therefore a local solution can be found by means of first-order methods. Results (see fig. 4 and other figures in the appendix) indicate that the resulting poisoning signals can relevantly impact the LS estimate, while the statistical indicators (central and right plot in fig. 4) show no evidence of anomaly. Furthermore, the attack seems to impact have a greater impact on the estimate of A_{LS} than that of B_{LS} (see also the appendix; for $\delta = 0.05$ we obtain $\mathbb{E}\|\Delta\tilde{A}_{\mathrm{LS}}\|_2 \approx 2$ and $\mathbb{E}\|\Delta\tilde{B}_{\mathrm{LS}}\|_2 \approx 0.15$). Preliminary results indicate that this effect may be due to the presence of the constraint g_3 on $(\mathcal{Z}_{\tilde{\mathcal{D}}}, \mathcal{Z}_{\mathcal{D}})$. Lastly, we observe that for small values of δ the residuals do not visually change in a sensible way (refer to the appendix), and an analysis of the outliers, based on the concept of leverage Kannan and Manoj (2015), shows no statistical difference for any values of δ. These findings suggest that it is possible to devise potentially undetectable poisoning attacks without making use of any sparsity assumption.

5. Conclusion

In this work, we have analyzed poisoning attacks on the data collected from a linear dynamical system affected by process noise. We have focused on the problem of poisoning the least-squares estimate of the underlying dynamical system, which is a quantity used by various data-driven controllers and thus can greatly affect their performance. We have established under which conditions the set of models compatible with the data includes the true model parameter, and we analyzed the effect of poisoning on the least-squares error. Based on the analysis of various attack strategies, we have proposed a new stealthy poisoning attack. Results indicate that this attack can relevantly impact the least-squares estimate while being stealthy from a statistical perspective. We conclude that it is possible to craft stealthy attacks that are not necessarily sparse. Possible detection methods include watermarking and/or encryption of the data. Future venues of research include analysis of online poisoning attacks; impact and detection of offline poisoning attacks; recovery of the original system matrices based on the set of compatible models.

Acknowledgments

This work was supported by the Swedish Foundation for Strategic Research through the CLAS project (grant RIT17-0046). In addition, the author is grateful to Prof. Alexandre Proutiere for his unwavering support and for providing the opportunity to work on this project.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

Scott Alfeld, Xiaojin Zhu, and Paul Barford. Data poisoning attacks against autoregressive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 16–25, 2006.

Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1467–1474, 2012.

Sergio Bittanti, Marco Campi, and Fabrizio Lorito. Effective identification algorithms for adaptive control. *International journal of adaptive control and signal processing*, 6(3):221–235, 1992.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Marco C Campi, Andrea Lecchini, and Sergio M Savaresi. Virtual reference feedback tuning: a direct method for the design of feedback controllers. *Automatica*, 38(8):1337–1346, 2002.

Jafar Abbaszadeh Chekan and Cedric Langbort. Regret bounds for lq adaptive control under database attacks (extended version). *arXiv preprint arXiv:2004.00241*, 2020.

Jeremy Coulson, John Lygeros, and Florian Dörfler. Data-enabled predictive control: In the shallows of the deepc. In *2019 18th European Control Conference (ECC)*, pages 307–312. IEEE, 2019.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Claudio De Persis and Pietro Tesi. Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, 65(3):909–924, 2019.

K Drouiche. A new test for whiteness. *IEEE transactions on signal processing*, 48(7):1864–1871, 2000.

Han Feng and Javad Lavaei. Learning of dynamical systems under adversarial attacks. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3010–3017. IEEE, 2021.

Håkan Hjalmarsson. Iterative feedback tuning—an overview. *International journal of adaptive control and signal processing*, 16(5):373–395, 2002.

Jonathan RM Hosking. The multivariate portmanteau statistic. *Journal of the American Statistical Association*, 75(371):602–608, 1980.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 19–35. IEEE, 2018.

K Senthamarai Kannan and K Manoj. Outlier detection in multivariate data. *Applied Mathematical Sciences*, 47(9):2317–2324, 2015.

A Karimi, L Mišković, and D Bonvin. Iterative correlation-based controller tuning. *International journal of adaptive control and signal processing*, 18(8):645–664, 2004.

David G Kleinbaum, Lawrence L Kupper, Azhar Nizam, and Eli S Rosenberg. *Applied regression analysis and other multivariable methods*. Cengage Learning, 2013.

Marius Kloft and Pavel Laskov. Online anomaly detection under adversarial impact. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 405–412. JMLR Workshop and Conference Proceedings, 2010.

Anne Koch, Julian Berberich, and Frank Allgöwer. Verifying dissipativity properties from noise-corrupted input-state data. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 616–621. IEEE, 2020.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017.

Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Lloyd S Nelson. The anderson-darling test for normality. *Journal of Quality Technology*, 30(3): 298, 1998.

Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via ℓ_1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 1–14, 2009.

Alessio Russo and Alexandre Proutiere. Poisoning attacks against data-driven control methods. In *2021 American Control Conference (ACC)*, pages 3234–3241. IEEE, 2021.

Alessio Russo, Marco Molinari, and Alexandre 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)*, pages 53–58. IEEE, 2021.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Mehrdad Showkatbakhsh, Paulo Tabuada, and Suhas Diggavi. Secure system identification. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1137–1141. IEEE, 2016.

Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys (CSUR)*, 2022.

Jan C Willems, Paolo Rapisarda, Ivan Markovsky, and Bart LM De Moor. A note on persistency of excitation. *Systems & Control Letters*, 54(4):325–329, 2005.

Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *international conference on machine learning*, pages 1689–1698. PMLR, 2015.

Yue Yu, Ruihan Zhao, Sandeep Chinchali, and Ufuk Topcu. Poisoning attacks against data-driven predictive control, 2022. URL https://arxiv.org/abs/2209.09108.

Xuezhou Zhang, Xiaojin Zhu, and Laurent Lessard. Online data poisoning attacks. In *Learning for Dynamics and Control*, pages 201–210. PMLR, 2020.