2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

From Passive Monitoring to Active Defence: Resilient Control of Manipulators Under Cyberattacks

Gabriele Gualandi and Alessandro V. Papadopoulos

***Abstract*— Cyber-physical robotic systems are vulnerable to *false data injection attacks* (FDIAs), in which an adversary corrupts sensor signals while evading residual-based *passive* anomaly detectors such as the χ^2 test. Such *stealthy* attacks can induce substantial end-effector deviations without triggering alarms. This paper studies the resilience of redundant manipulators to stealthy FDIAs and advances the architecture from passive monitoring to active defence. We formulate a closed-loop model comprising a feedback-linearized manipulator, a steady-state Kalman filter, and a χ^2-based anomaly detector. Building on this passive monitoring layer, we propose an active control-level defence that attenuates the control input through a monotone function of an anomaly score generated by a novel actuation-projected, measurement-free state predictor. The proposed design provides probabilistic guarantees on nominal actuation loss and preserves closed-loop stability. From the attacker perspective, we derive a convex QCQP for computing one-step optimal stealthy attacks. Simulations on a 6-DOF planar manipulator show that the proposed defence significantly reduces attack-induced end-effector deviation while preserving nominal task performance in the absence of attacks.**

I. Introduction

Robotic manipulators are increasingly deployed in open and networked environments, ranging from industrial assembly to collaborative human-robot interaction. Their tight integration of computation, communication, and actuation, however, makes them vulnerable to cyberattacks that directly compromise safety and reliability. Among cyber-physical threats, attacks on *data integrity* are particularly critical: by corrupting sensor information, an adversary can mislead the controller and alter the robot's behavior without any physical contact [1], [2]. Such threats are especially concerning when they are *stealthy*, i.e., engineered to remain below the threshold of an anomaly detection system (ADS), thereby avoiding alarms while still driving the end-effector away from its intended task [3], [4].

A well-studied class of integrity attacks are *False Data Injection Attacks* (FDIAs), in which adversaries manipulate sensor signals to achieve malicious objectives. In networked control and power systems, stealthy FDIAs have been extensively analyzed: from characterizations of undetectability [5], [6], [7] to optimal attack synthesis. In robotics, however, prior work has primarily focused on passive anomaly detection or on "perfectly undetectable" attacks that completely bypass detection [7]. These perspectives miss a key robotics-specific vulnerability: the widespread use of *feedback linearization* reduces manipulator dynamics to double integrators, which in turn induces an *integrator vulnerability*. As a consequence, persistent sensor corruption can silently accumulate in the closed-loop system, driving large task-space errors even under residual-based χ^2 detection [8].

This paper addresses this gap by introducing an *active defence* strategy that transforms anomaly detection from passive monitoring into a resilience mechanism. The proposed method leverages an *actuation-projected state predictor*, a model-driven state estimate that ignores sensing, to compute an anomaly score immune to direct sensor corruption. Based on this score, we introduce *anomaly-aware command scaling*, which attenuates commanded accelerations as anomalies grow, thereby reducing the adversary's ability to steer the manipulator while preserving nominal performance.

Contributions.: The contributions of this work are the following:

1) We formalize stealthy false data injection attacks (FDIAs) against the sensors of feedback-linearized manipulators, exposing their integrator vulnerability and showing that the attacker's one-step optimal strategy reduces to a convex QCQP.
2) We propose *anomaly-aware command scaling*, which attenuates control inputs based on a measurement-free actuation-projected state predictor.
3) We provide probabilistic guarantees on bounded attenuation in nominal operation and prove closed-loop stability under the proposed defence.
4) Simulations on a 6-DOF redundant manipulator demonstrate that the defence significantly limits attacker-realizable end-effector deviations while preserving task performance in the absence of attacks.

A. Related Work

Anomaly detection in CPS. Residual-based detectors, particularly χ^2 tests on Kalman innovations, are a classical tool for monitoring CPS integrity [9]. These methods provide statistical guarantees on false-alarm rates and are widely used in industrial practice. Extensions include adaptive thresholds [10] and sequential schemes such as CUSUM tests [11]. However, such ADS are inherently *passive*: they detect anomalies but do not alter the control policy, leaving the system structure unchanged.

False Data Injection Attacks. Theoretical studies of FDIAs in networked systems have characterized undetectability conditions [5], [6], affine attack structures [7], and optimal attack strategies [3]. These works typically assume

This work was supported by the Swedish Research Council (VR) with the PSI project No. #2020-05094, and by the Knowledge Foundation (KKS).

G. Gualandi and A.V. Papadopoulos are with the Department of Computer Science and Engineering, Mälardalen University, Västerås, Sweden gabriele.gualandi@mdu.se

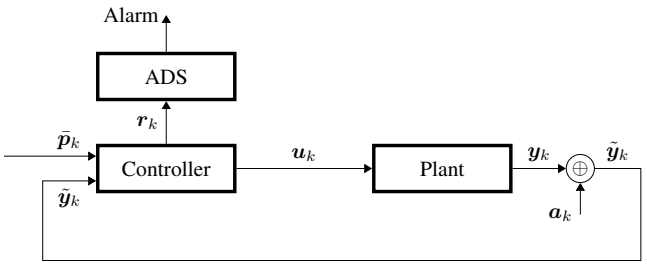

Fig. 1. Closed-loop system under FDIA. Sensor signals are corrupted by injection $\boldsymbol{a}_k$, yielding attacked output $\tilde{\boldsymbol{y}}_k$. The Kalman filter generates the innovation $\boldsymbol{r}_k$, which is monitored by the ADS.

either "perfectly undetectable" attacks (no residual information leaks) or focus on power networks and generic LTI plants. Robotics-specific studies remain limited: most works concentrate on detector design rather than on modifying the controller to actively limit attack effectiveness [4].

Integrator Vulnerability. Control systems with an LTI plant having integral action are subject to integrator vulnerability [8], where the residual generated by any linear observer follows the same distribution during normal operation and under attack in the steady-state regime, making sensor bias injecting attack detectable only during transients.

Our contribution. We bridge these lines of research by combining a residual-based ADS with a novel active defence operating at the control level. First, we show that feedback linearization induces an *integrator vulnerability*: a PD-based FDIA can precisely steer the end-effector despite a residual-based χ^2 detector. We then show that our gain-scaling defence, driven by the anomaly score, reduces the attack's effectiveness while providing probabilistic guarantees in attack-free operation, thereby improving the cyber-physical security of robotic manipulators.

II. System Model

We consider a robotic manipulator operating in closed loop with a state estimator, a task-space controller, and an anomaly detection system (ADS). The architecture, illustrated in Fig. 1, captures the interaction between defender and adversary: (i) the plant dynamics, (ii) a Kalman filter for state estimation, (iii) task-space control, (iv) a residual-based ADS, and (v) an additive adversarial attack on sensor measurements.

A. Closed-Loop Architecture

At each discrete time step $k \in \mathbb{Z}_{\geq 0}$, the plant output $\boldsymbol{y}_k \in \mathbb{R}^p$ is corrupted by an injected signal $\boldsymbol{a}_k$ to yield the attacked measurement

$$\tilde{\boldsymbol{y}}_k = \boldsymbol{y}_k + \boldsymbol{a}_k, \tag{1}$$

B. Plant and State Estimator

The plant is modeled as a discrete-time LTI system:

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k, \tag{2a}$$
$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k, \tag{2b}$$

with state $\boldsymbol{x}_k \in \mathbb{R}^n$, process noise $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$, and measurement noise $\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$. We assume $(\boldsymbol{A}, \boldsymbol{C})$ is detectable, ensuring the existence of a unique stabilizing solution $\boldsymbol{P}$ to the DARE

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^\top + \boldsymbol{Q} - \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^\top\big(\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\top + \boldsymbol{R}\big)^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{A}^\top.$$

State estimates are obtained from a steady-state Kalman filter in single-step innovation form:

$$\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{L}\,\boldsymbol{r}_k, \tag{3a}$$
$$\boldsymbol{r}_k = \tilde{\boldsymbol{y}}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k, \tag{3b}$$

where $\boldsymbol{r}_k \in \mathbb{R}^p$ is the innovation. Its steady-state covariance is $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\top + \boldsymbol{R}$, and the corresponding steady-state gain is

$$\boldsymbol{L} = \boldsymbol{A}\,\boldsymbol{P}\,\boldsymbol{C}^\top\,\boldsymbol{\Sigma}^{-1}. \tag{4}$$

The residual $\boldsymbol{r}_k$ and its covariance $\boldsymbol{\Sigma}$ will later be used for residual-based anomaly detection.

C. Manipulator Model and Task-Space Control

We consider a p-DOF robotic manipulator with sensor readings providing joint values $\boldsymbol{q}$, and having state $\boldsymbol{x} = [\boldsymbol{q}^\top\ \dot{\boldsymbol{q}}^\top]^\top \in \mathbb{R}^n$, where $n = 2 \cdot p$ and $\boldsymbol{q} \in \mathbb{R}^p$.

The continuous-time joint-space dynamics are

$$\boldsymbol{M}(\boldsymbol{q})\,\ddot{\boldsymbol{q}} + \boldsymbol{\nu}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \boldsymbol{\tau}, \tag{5}$$

where $\boldsymbol{M}(\boldsymbol{q}) \succ 0$ is the inertia matrix and $\boldsymbol{\nu}$ collects Coriolis, centrifugal, and gravity terms. With inverse-dynamics compensation, i.e., $\boldsymbol{\tau} = \boldsymbol{M}(\boldsymbol{q})\,\boldsymbol{u}^{\text{nom}} + \boldsymbol{\nu}(\boldsymbol{q}, \dot{\boldsymbol{q}})$, the joint dynamics reduce to decoupled double integrators,

$$\ddot{\boldsymbol{q}} = \boldsymbol{u}^{\text{nom}}, \tag{6}$$

whose discrete-time form is represented by Eq. (2) with $\boldsymbol{x}_k = [\boldsymbol{q}_k^\top\ \dot{\boldsymbol{q}}_k^\top]^\top$ and input $\boldsymbol{u}_k = \boldsymbol{u}_k^{\text{nom}}$.

The controller operates in task space using twist control [12], [13], tracking position references $\{\bar{\boldsymbol{p}}_k, \dot{\bar{\boldsymbol{p}}}_k, \ddot{\bar{\boldsymbol{p}}}_k\}$ and orientation references $\{\bar{\boldsymbol{R}}_k, \bar{\boldsymbol{\omega}}_k, \dot{\bar{\boldsymbol{\omega}}}_k\}$, where $\bar{\boldsymbol{R}}_k \in SO(3)$ and $\bar{\boldsymbol{\omega}}_k \in \mathbb{R}^3$.

Position and orientation errors are defined as

$$\boldsymbol{e}_{\boldsymbol{p},k} = \bar{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}_k, \tag{7}$$
$$\boldsymbol{e}_{\boldsymbol{o},k} = \sin\!\Big(\tfrac{\hat{\theta}_k}{2}\Big)\ \hat{\boldsymbol{r}}_k, \tag{8}$$

where $(\hat{\theta}_k, \hat{\boldsymbol{r}}_k)$ is the angle–axis pair of the rotation error $\bar{\boldsymbol{R}}_k\,\hat{\boldsymbol{R}}_k^\top$ with $\hat{\theta}_k \in [0, \pi]$. The estimates $\hat{\boldsymbol{p}}_k$ and $\hat{\boldsymbol{R}}_k$ are obtained via forward kinematics from $\hat{\boldsymbol{q}}_k$ extracted from the state estimate $\hat{\boldsymbol{x}}_k$.

A PD+feedforward law generates desired task accelerations:

$$\boldsymbol{u}_k^{\text{c}} = \begin{bmatrix} \ddot{\bar{\boldsymbol{p}}}_k + \boldsymbol{K}_{pp}\,\boldsymbol{e}_{\boldsymbol{p},k} + \boldsymbol{K}_{dp}\,\dot{\boldsymbol{e}}_{\boldsymbol{p},k} \\ \dot{\bar{\boldsymbol{\omega}}}_k + \boldsymbol{K}_{po}\,\boldsymbol{e}_{\boldsymbol{o},k} + \boldsymbol{K}_{do}\,\dot{\boldsymbol{e}}_{\boldsymbol{o},k} \end{bmatrix}, \tag{9}$$

where $\dot{\boldsymbol{e}}_{\boldsymbol{p},k} \coloneqq \dot{\bar{\boldsymbol{p}}}_k - \dot{\hat{\boldsymbol{p}}}_k$ and $\dot{\boldsymbol{e}}_{\boldsymbol{o},k} \coloneqq \bar{\boldsymbol{\omega}}_k - \hat{\boldsymbol{\omega}}_k$ are velocity errors. The gains are synthesized via discrete-time LQR on the discretized double-integrator model (6).

Joint accelerations are then computed via the pseudoinverse of the geometric Jacobian $\boldsymbol{J}(\boldsymbol{q}) \in \mathbb{R}^{6\times p}$ [13]:

$$\boldsymbol{u}_k^{\text{nom}} = \boldsymbol{J}^\dagger(\boldsymbol{q})\big(\boldsymbol{u}_k^{\text{c}} - \dot{\boldsymbol{J}}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}}\big). \tag{10}$$

D. Passive Anomaly Detection System

The ADS monitors the innovation sequence $\boldsymbol{r}_k$. With steady-state covariance $\boldsymbol{\Sigma} \succ 0$ (hence invertible), the Mahalanobis distance

$$z_k = \boldsymbol{r}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_k \tag{11}$$

is $\chi^2(p)$-distributed under $\mathcal{H}_0$ (no attack), where $p = \dim(\boldsymbol{y}_k)$. To reduce sensitivity to single-sample fluctuations, we employ a windowed statistic of length $W \in \mathbb{N}$:

$$w_k = \sum_{i=k-W+1}^{k} z_i. \tag{12}$$

During execution, an alarm is triggered if $w_k > \tau$.

Lemma 1 (Innovation whiteness): Under $\mathcal{H}_0$ (no attack), the innovation process satisfies

$$\boldsymbol{r}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \qquad \boldsymbol{r}_k \text{ independent across } k.$$

Proof: This is a standard property of the Kalman filter with correct noise statistics: the innovation sequence is white, zero-mean Gaussian with covariance $\boldsymbol{\Sigma}$, see [9]. ∎

Lemma 2 (Chi-squared distribution): Under $\mathcal{H}_0$, the statistics satisfy

$$z_k \sim \chi^2(p), \qquad w_k \sim \chi^2(pW),$$

where p is the output dimension and W the window length.

Proof: From Lemma 1, $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{r}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $z_k = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{r}_k\|_2^2 \sim \chi^2(p)$ [14]. Independence across time implies w_k is the sum of W i.i.d. $\chi^2(p)$ variables, hence $\chi^2(pW)$, see [10]. ∎

Corollary 1 (Threshold calibration): Given a desired per-step false-alarm probability $\alpha \in (0,1)$, set

$$\tau \;=\; F^{-1}_{\chi^2(pW)}(1-\alpha) \;=\; 2\,\mathrm{P}^{-1}\Big(1-\alpha, \tfrac{pW}{2}\Big), \tag{13}$$

which ensures $\Pr(w_k > \tau \mid \mathcal{H}_0) = \alpha$, where P^{-1} denotes the inverse of the regularized lower incomplete gamma function [10].

E. Attacker Model

We now state the assumptions under which both the defence and attack strategies are developed.

Assumption 1 (Adversary knowledge): The adversary is *omniscient*: it has full knowledge of the plant, controller, state estimator, and defence system(s).

Assumption 2 (Adversary capabilities): The only attack surface is additive injection into sensor measurements. The adversary cannot tamper with actuators, control logic, ADS parameters, timing, or packet ordering.

Assumption 3 (Adversary goal): The adversary's goal is to manipulate the position of the end-effector in task space, as formalized later by an attack objective on the task-space trajectory, subject to the detection mechanism.

Assumption 4 (Converged estimator): Without loss of generality, the attack starts at $k = 0$ and lasts at most T samples. At $k = 0$, the Kalman filter operates at steady state (covariances and gains constant), i.e., the estimator has converged.

III. Proposed Defence Method

We now present a defence strategy based on gain scaling to reduce the impact of sensing FDIA, with probabilistic guarantees on the performance loss under attack-free operation.

A. Attack Estimation

We define the *actuation-projected state* $\tilde{\boldsymbol{x}}_k$ as the measurement-free state prediction driven solely by commanded inputs. To mitigate drift from model mismatch, this predictor is periodically re-synchronized with the Kalman estimate as ground truth. Without loss of generality, assume the most recent re-synchronization occurs at $k = 0$, so

$$\tilde{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0. \tag{14}$$

For subsequent steps, the actuation-projected state follows the noise-free, open-loop predictor

$$\tilde{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\,\tilde{\boldsymbol{x}}_k + \boldsymbol{B}\,\boldsymbol{u}_k, \qquad k \geq 0. \tag{15}$$

We define the *actuation-projected residual* (distinct from the innovation $\boldsymbol{r}_k$) as

$$\tilde{\boldsymbol{r}}_k \;\coloneqq\; \hat{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_k \in \mathbb{R}^n, \qquad n = 2p. \tag{16}$$

Theorem 1 (Actuation-projected residual covariance): The covariance of the residual in Eq. (16) after k steps from initialization, $\boldsymbol{\Sigma}_{\tilde{r},k} \in \mathbb{R}^{n\times n}$, is the $(2,2)$ block of the matrix $\boldsymbol{P}_{z,k} \in \mathbb{R}^{2n\times 2n}$ obtained by iterating

$$\boldsymbol{P}_{z,j+1} = \boldsymbol{F}\boldsymbol{P}_{z,j}\boldsymbol{F}^\top + \boldsymbol{\Pi}, \quad j = 0, \ldots, k-1, \tag{17}$$

from the initial condition $\boldsymbol{P}_{z,0} = \mathrm{diag}(\boldsymbol{P}, \boldsymbol{0})$, where

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{L}\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{L}\boldsymbol{C} & \boldsymbol{A} \end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \tag{18}$$

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{L} \\ \boldsymbol{0} & \boldsymbol{L} \end{bmatrix} \in \mathbb{R}^{2n\times(n+p)}, \tag{19}$$

$$\boldsymbol{\Pi} = \boldsymbol{G}\,\mathrm{diag}(\boldsymbol{Q}, \boldsymbol{R})\boldsymbol{G}^\top \in \mathbb{R}^{2n\times 2n}. \tag{20}$$

Proof: From the system equations, the one-step evolution of the estimation error $\boldsymbol{e}_k := \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k$ and of the actuation-projected residual $\tilde{\boldsymbol{r}}_k := \hat{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_k$ is:

$$\boldsymbol{e}_{k+1} = (\boldsymbol{A} - \boldsymbol{L}\boldsymbol{C})\boldsymbol{e}_k + \boldsymbol{w}_k - \boldsymbol{L}\boldsymbol{v}_k, \tag{21}$$

$$\tilde{\boldsymbol{r}}_{k+1} = \boldsymbol{A}\tilde{\boldsymbol{r}}_k + \boldsymbol{L}\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{L}\boldsymbol{v}_k. \tag{22}$$

Stacking $\boldsymbol{z}_k := [\boldsymbol{e}_k^\top, \tilde{\boldsymbol{r}}_k^\top]^\top$ yields

$$\boldsymbol{z}_{k+1} = \boldsymbol{F}\boldsymbol{z}_k + \boldsymbol{G}\boldsymbol{\eta}_k, \tag{23}$$

with $\boldsymbol{F}, \boldsymbol{G}$ as in Eqs. (18) and (19) and $\boldsymbol{\eta}_k \triangleq [\boldsymbol{w}_k^\top, \boldsymbol{v}_k^\top]^\top$. The covariance propagates as

$$\begin{aligned}\boldsymbol{P}_{z,k+1} = {} & \boldsymbol{F}\,\mathbb{E}[\boldsymbol{z}_k\boldsymbol{z}_k^\top]\boldsymbol{F}^\top + \boldsymbol{F}\,\mathbb{E}[\boldsymbol{z}_k\boldsymbol{\eta}_k^\top]\boldsymbol{G}^\top \\ & + \boldsymbol{G}\,\mathbb{E}[\boldsymbol{\eta}_k\boldsymbol{z}_k^\top]\boldsymbol{F}^\top + \boldsymbol{G}\,\mathbb{E}[\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\top]\boldsymbol{G}^\top.\end{aligned}$$

Since $\boldsymbol{\eta}_k$ is white, zero-mean, and independent of $\boldsymbol{z}_k$, $\mathbb{E}[\boldsymbol{z}_k\boldsymbol{\eta}_k^\top] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\top] = \mathrm{diag}(\boldsymbol{Q}, \boldsymbol{R})$, giving

$$\boldsymbol{P}_{z,k+1} = \boldsymbol{F}\boldsymbol{P}_{z,k}\boldsymbol{F}^\top + \boldsymbol{\Pi},$$

with $\boldsymbol{\Pi}$ as in Eq. (20).

At synchronization $k = 0$, $\mathrm{Cov}(\boldsymbol{e}_0) = \boldsymbol{P}$ and $\tilde{\boldsymbol{r}}_0 = \boldsymbol{0}$, hence

$$\boldsymbol{P}_{z,0} = \begin{bmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \mathrm{diag}(\boldsymbol{P}, \boldsymbol{0}).$$

Iterating (17) k times yields $\boldsymbol{P}_{z,k}$; the desired covariance is the $(2,2)$ block, $\boldsymbol{\Sigma}_{\tilde{r},k}$. ∎

Theorem 2 (Confidence for the Anomaly Measure): Consider the anomaly measure

$$\tilde{z}_k = \tilde{\boldsymbol{r}}_k^\top \boldsymbol{\Sigma}_{\tilde{r},k}^{-1} \tilde{\boldsymbol{r}}_k. \tag{24}$$

Choosing $z_{\mathrm{x}} = F^{-1}_{\chi^2(n)}(\psi)$ ensures $\mathbb{P}(\tilde{z}_k \le z_{\mathrm{x}} \mid \mathcal{H}_0) = \psi$ for any given probability ψ.

Proof: Under $\mathcal{H}_0$, the residual $\tilde{\boldsymbol{r}}_k$ is a zero-mean Gaussian vector, $\tilde{\boldsymbol{r}}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\tilde{r},k})$. The time-varying normalization with the covariance $\boldsymbol{\Sigma}_{\tilde{r},k}$ at each step, computed as in Theorem 1, ensures that the resulting statistic has a stationary distribution, $\tilde{z}_k \sim \chi^2(n)$, for all k. The theorem's result follows from the definition of the inverse CDF, $F^{-1}_{\chi^2(n)}$. ∎

B. Command scaling active defence

As an active defence mechanism, we propose to reduce the magnitude of command $\boldsymbol{u}_k^{\mathrm{nom}}$ of Eq. (10) through a gain scaling function $f(\tilde{z})$, with $\tilde{z}$ as in Eq. (24).

Let $z_{\mathrm{x}} > 0$ be a design abscissa and $\beta \in (0,1)$ the desired gain factor at $\tilde{z} = z_{\mathrm{x}}$. Introduce a shape exponent $\gamma > 0$ and a smooth, strictly decreasing function $f : [0,\infty) \to (0,1]$,

$$f(\tilde{z}) = \exp\left[-\left(\frac{\tilde{z}}{z_{\mathrm{scale}}} \right)^{\gamma} \right], \tag{25}$$

$$z_{\mathrm{scale}} = \frac{z_{\mathrm{x}}}{\big(-\ln(\beta)\big)^{1/\gamma}}, \tag{26}$$

satisfying $f(0) = 1$, $f(z_{\mathrm{x}}) = \beta$, and $f(\tilde{z}) \to 0$ as $\tilde{z} \to \infty$.

Theorem 3: Consider a robotic manipulator with nominal joint-space control input $\boldsymbol{u}_k^{\mathrm{nom}}$ as in (10). Let $\tilde{z}_k \in [0,\infty)$ denote the anomaly score defined in (24), and let $f : [0,\infty) \to (0,1]$ be a strictly decreasing function satisfying

$$f(0) = 1, \quad \lim_{\tilde{z}\to\infty} f(\tilde{z}) = 0.$$

Define the scaled control law

$$\boldsymbol{u}_k = f(\tilde{z}_k)\, \boldsymbol{u}_k^{\mathrm{nom}}. \tag{27}$$

Then the following hold:

1) **Probabilistic actuation guarantee.** Let $\tilde{z}_k \sim \chi^2(n)$ under the null hypothesis, and let $z_{\mathrm{x}} = F^{-1}_{\chi^2(n)}(\psi)$ denote the inverse CDF at confidence level $\psi \in (0,1)$. If design parameters of $f(\cdot)$ are chosen such that
$$f(z_{\mathrm{x}}) \ge \beta,$$
for some $\beta \in (0,1)$, then with probability at least ψ, the scaled input magnitude satisfies
$$\|\boldsymbol{u}_k\| \;\ge\; \beta\, \|\boldsymbol{u}_k^{\mathrm{nom}}\|. \tag{28}$$
2) **Closed-loop stability (frozen gain).** Consider the discretized double-integrator dynamics induced by inverse-dynamics compensation in (6). If the nominal frozen-gain closed loop is Schur stable (i.e., (27) with $f(\cdot) \equiv 1$), then for any constant gain factor $\bar{f} \in (0,1]$ the frozen-gain closed loop under Eq. (27) is Schur stable, and hence exponentially stable.

Proof: **Probabilistic actuation guarantee.** Under $\mathcal{H}_0$, the anomaly score follows $\tilde{z}_k \sim \chi^2(n)$. By definition of the quantile, $\Pr(\tilde{z}_k \le z_{\mathrm{x}}) = \psi$. On this event, since $f(\cdot)$ is non-increasing, we have $f(\tilde{z}_k) \ge f(z_{\mathrm{x}})$. By assumption, $f(z_{\mathrm{x}}) \ge \beta$. Substituting into (27),

$$\|\boldsymbol{u}_k\| = f(\tilde{z}_k)\, \|\boldsymbol{u}_k^{\mathrm{nom}}\| \ge \beta \|\boldsymbol{u}_k^{\mathrm{nom}}\|.$$

Therefore, with probability at least ψ, the commanded input magnitude remains at least a fraction β of the nominal value.

Closed-loop stability (frozen gain). Under inverse-dynamics compensation, and treating the Jacobian pseudoinverse as locally constant, it suffices to consider one joint with $u_{k,j}^{\mathrm{nom}} = -\boldsymbol{K}\boldsymbol{x}_{k,j}$, where $\boldsymbol{x}_{k,j} \in \mathbb{R}^2$ and $\boldsymbol{K} = [K_p \;\; K_d]$. For a frozen $\bar{f} \in (0,1]$, the closed-loop matrix is $\boldsymbol{A}_{\mathrm{cl}}(\bar{f}) = \boldsymbol{A}_j - \bar{f}\,\boldsymbol{B}_j\boldsymbol{K}$, where $\boldsymbol{A}_j, \boldsymbol{B}_j$ are the joint double-integrator matrices. Applying the second-order Jury criterion to the characteristic polynomial of $\boldsymbol{A}_{\mathrm{cl}}(\bar{f})$, Schur stability is equivalent to $\bar{f}T_s^2 K_p > 0$, $2 - \bar{f}T_s K_d > 0$, and $K_d > \frac{T_s}{2} K_p$. These conditions hold at $\bar{f} = 1$ by assumption, hence $K_p > 0$, $2 > T_s K_d$, and $K_d > \frac{T_s}{2} K_p$. Therefore, for every $\bar{f} \in (0,1]$, the first condition holds since $\bar{f} > 0$ and $K_p > 0$, the second becomes less restrictive as $\bar{f}$ decreases, and the third is independent of $\bar{f}$. Thus $\boldsymbol{A}_{\mathrm{cl}}(\bar{f})$ is Schur for all $\bar{f} \in (0,1]$, and the frozen-gain closed loop is exponentially stable. ∎

In practice, $f(\tilde{z}_k)$ is time-varying and state-dependent; however, item 1 of the theorem confines f to the compact set $[\beta, 1]$ with probability ψ, over which frozen stability holds uniformly.

Corollary 2: Fix a desired confidence level $\psi \in (0,1)$ and a desired gain floor $\beta \in (0,1)$. Fix $z_{\mathrm{x}} = F^{-1}_{\chi^2(n)}(\psi)$ in (25) so that $f(z_{\mathrm{x}}) = \beta$. Then, for the control law (27) with $\tilde{z}_k$ as in Eq. (24), item 1 of Theorem 3 guarantees that the magnitude of the final command satisfies (28) with probability ψ, i.e., the final input remains at least $\beta \cdot 100\%$ of the nominal magnitude with probability ψ.

IV. Optimal Stealth Attack

We now formalize the adversary's strategy under Assumptions of Section II-E. The attacker's goal is to induce stealthy malicious end-effector accelerations.

To express the predicted trajectories of different vectors (e.g., acceleration, velocity) from a closed-loop simulation, we introduce the following compact notation.

Let $\mathcal{Z}_{k,j}(\boldsymbol{a})$ denote the *j-step-ahead prediction at time $k+j$, computed at time k*, of the closed-loop system under the attack sequence $[\boldsymbol{a}, \boldsymbol{0}, \ldots, \boldsymbol{0}]$ (of length $j+1$).

For a generic vector $\boldsymbol{v}$, we define

$$\boldsymbol{v}_{k+j}^{\mathrm{SIM}} := \pi_{\boldsymbol{v}}\big(\mathcal{Z}_{k,j}(\boldsymbol{a})\big), \tag{29}$$

where $\pi_{\boldsymbol{v}}(\cdot)$ extracts $\boldsymbol{v}$ from the noise-free simulated trajectory at prediction step j (i.e., the terminal value at time $k+j$). The operator $\mathcal{Z}_{k,j}(\boldsymbol{a})$ models Assumption 1 of Section II-E, as it propagates Eqs. (2), (3a) and (3b) under perfect-model assumptions (exact initial state and noise-free dynamics) and it computes the control input via Eq. (27) including the proposed active defence mechanism.

A. Attack Model and Delay Structure

Due to estimator and control delays, an injected signal at time k, $\boldsymbol{a}_k$, influences the end-effector acceleration two steps later (under the closed-loop dynamics in Eqs. (2), (3a) and (3b)). Specifically,

$$\ddot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+2} = \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_k)\big), \tag{30}$$

where $\mathcal{Z}_{k,2}$ denotes a two-step-ahead simulation of the closed-loop system under attack sequence $[\boldsymbol{a}_k, \boldsymbol{0}, \boldsymbol{0}]$.

The attacker's high-level objective at each time step is to make the predicted end-effector acceleration match a desired target acceleration $\ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2}$. This is formulated as:

$$\min_{\boldsymbol{a}_k} \quad \tfrac{1}{2}\left\|\ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2} - \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_k)\big)\right\|^2. \tag{31}$$

B. Incremental Attack Formulation

Due to feedback linearization, the manipulator's joint-space dynamics reduce to double integrators, which entails an *integrator vulnerability*: a persistent sensor bias can induce drift in the regulated variables. This vulnerability has been analyzed for sensor *bias injection attacks* (BIAs)—i.e., constant sensor injections—in linear systems [8]. Although the joint–task mapping is nonlinear and we consider more general *false data injection attacks* (FDIAs) with arbitrary sensor injections, we argue that effective FDIAs still exploit the integrator vulnerability *locally*. For this reason, we model the attack *incrementally*:

$$\boldsymbol{a}_k \;=\; \boldsymbol{a}_{k-1} + \boldsymbol{\Delta}_k, \qquad k \geq 1, \tag{32}$$

where $\boldsymbol{\Delta}_k$ is an increment and $\boldsymbol{a}_k$ is initialized as $\boldsymbol{a}_0 = \boldsymbol{\Delta}_0$ at the attack onset. This parameterizes an FDIA as deviations from a baseline BIA (via the increments $\boldsymbol{\Delta}_k$), and is convenient for gradient-based synthesis.

A first-order expansion of $\pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a})\big)$ around $\boldsymbol{a}_{k-1}$, with $\ddot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+2}$ as in Eq. (30), gives:

$$\ddot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+2} \;\approx\; \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_{k-1})\big) + \boldsymbol{Z}_k\boldsymbol{\Delta}_k \;+\; \mathcal{O}(\|\boldsymbol{\Delta}_k\|^2), \tag{33}$$

where the Jacobian

$$\boldsymbol{Z}_k \;=\; \frac{\partial}{\partial \boldsymbol{a}}\,\pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a})\big)\bigg|_{\boldsymbol{a}=\boldsymbol{a}_{k-1}} \tag{34}$$

is computed numerically via a central-difference scheme with step-size tuning [15]; $\boldsymbol{Z}_k$ provides the local sensitivity needed to differentiate the objective Eq. (31).

By substituting the linear attack model (33) into (31) and introducing a regularization term, a quadratic objective function is obtained. Specifically, the attack increment $\boldsymbol{\Delta}_k$ minimizes the quadratic cost

$$\tfrac{1}{2}\big\|\boldsymbol{Z}_k\boldsymbol{\Delta}_k - \ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2} + \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_{k-1})\big)\big\|^2 + \tfrac{\zeta}{2}\|\boldsymbol{\Delta}_k\|^2, \tag{35}$$

where $\zeta > 0$ penalizes large increments.

C. Attack Objective

We restrict the attack to the translational DOFs, as these are most relevant to adversarial manipulation. The adversary defines $\ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2}$ of Eq. (31) via a one-step-ahead PD law plus feedforward based on a plan $\{\bar{\boldsymbol{p}}^{\mathrm{A}}_k, \dot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_k, \ddot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_k\}$. Specifically, the PD errors are computed as the difference between the desired and predicted quantities, under the incremental attack approach ($\boldsymbol{\Delta}_k$=0 or equivalently, $\boldsymbol{a}_k = \boldsymbol{a}_{k-1}$):

$$\begin{aligned}\ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2} = {} & \boldsymbol{K}^{\mathrm{A}}_{\mathrm{p}}\big(\bar{\boldsymbol{p}}^{\mathrm{A}}_{k+1} - \boldsymbol{p}^{\mathrm{SIM}}_{k+1}\big) \\ & + \boldsymbol{K}^{\mathrm{A}}_{\mathrm{d}}\big(\dot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_{k+1} - \dot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+1}\big) + \ddot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_{k},\end{aligned} \tag{36}$$

where

$$\begin{aligned}\boldsymbol{p}^{\mathrm{SIM}}_{k+1} &= \pi_{\boldsymbol{p}}\big(\mathcal{Z}_{k,1}(\boldsymbol{a}_{k-1})\big), \\ \dot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+1} &= \pi_{\dot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,1}(\boldsymbol{a}_{k-1})\big),\end{aligned}$$

are the one-step-ahead end-effector position and velocity obtained from the closed-loop simulator under the attack sequence $[\boldsymbol{a}_{k-1}, \boldsymbol{0}]$. The resulting $\ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2}$ is therefore a way to counteract the predicted drift resulting from $\boldsymbol{\Delta}_k$=0.

D. Stealth Constraint

The ADS residual is modeled as

$$\boldsymbol{r}_k = \boldsymbol{\Delta}_k + \boldsymbol{c}_k, \tag{37}$$

with $\boldsymbol{c}_k$ denoting the baseline innovation, defined as

$$\boldsymbol{c}_k = \big(\boldsymbol{y}_k + \boldsymbol{a}_{k-1} - \hat{\boldsymbol{y}}_k\big). \tag{38}$$

The adversary uniformly allocates a per-step anomaly budget

$$\tau' = \tfrac{\tau}{T}, \tag{39}$$

so that the stealth constraint becomes

$$\big(\boldsymbol{\Delta}_k + \boldsymbol{c}_k\big)^\top \boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\Delta}_k + \boldsymbol{c}_k\big) \leq \tau'. \tag{40}$$

E. QCQP Formulation

Considering the objective function Eq. (35) and expanding the constraint of Eq. (40), the adversary's problem reduces to the convex QCQP

$$\begin{aligned}\min_{\boldsymbol{\Delta}_k} \; & \tfrac{1}{2}\boldsymbol{\Delta}_k^\top \boldsymbol{H}\boldsymbol{\Delta}_k + \boldsymbol{g}^\top\boldsymbol{\Delta}_k \\ \text{s.t. } & \boldsymbol{\Delta}_k^\top \boldsymbol{O}\boldsymbol{\Delta}_k + \boldsymbol{b}^\top\boldsymbol{\Delta}_k + c \leq 0,\end{aligned} \tag{41}$$

with parameters

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{Z}_k^\top\boldsymbol{Z}_k + \zeta\boldsymbol{I}, \\ \boldsymbol{g} &= -\boldsymbol{Z}_k^\top\Big(\ddot{\boldsymbol{p}}^{\mathrm{A}}_{k+2} - \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_{k-1})\big)\Big), \\ \boldsymbol{O} &= \boldsymbol{\Sigma}^{-1}, \\ \boldsymbol{b} &= 2\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_k, \\ c &= \boldsymbol{c}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{c}_k - \tau'.\end{aligned}$$

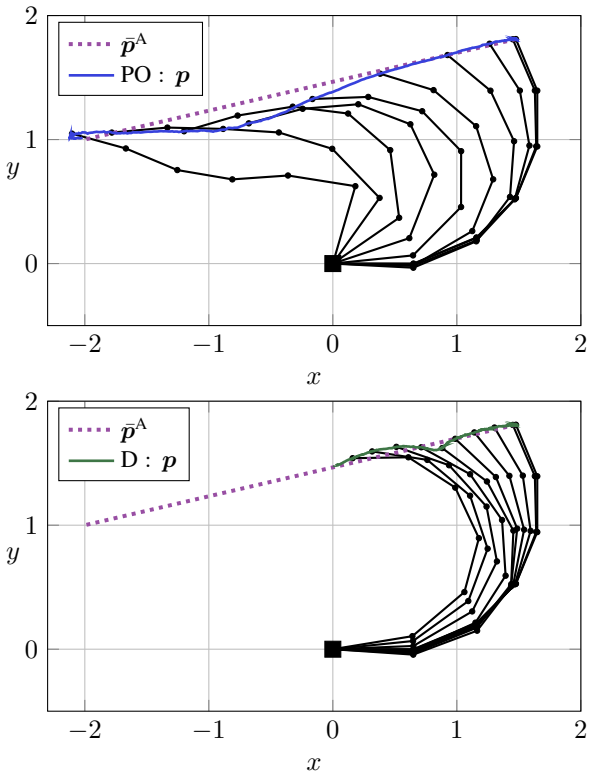

Fig. 2. End-effector trajectories: attacker's reference $\bar{p}^{\mathrm{A}}$, defended only by the passive detector (PO) and by the proposed active defence (D). The latter defence limits drift toward the malicious target.

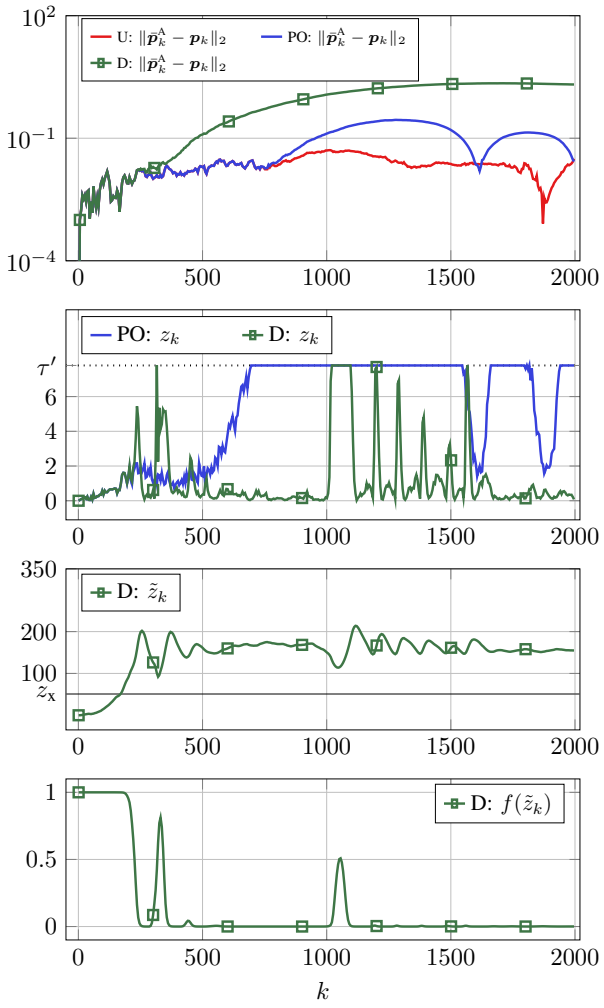

Fig. 3. Attack tracking error $\|\bar{p}_k^{\mathrm{A}} - p_k\|_2$, Mahalanobis distances z_k and $\tilde{z}_k$ of Eqs. (11) and (24), and scaling $f(\tilde{z})$ of Eq. (25). With active defence (D), $\tilde{z}_k$ increases until scaling becomes effective, then settles, limiting attacker-realizable accelerations.

Theorem 4 (Convexity of adversary's QCQP): Problem (41) is convex, and thus admits the global minimizer $\boldsymbol{\Delta}_k^\star$.

Proof: The cost Hessian $\boldsymbol{H} = \boldsymbol{Z}_k^\top \boldsymbol{Z}_k + \zeta \boldsymbol{I} \succeq \zeta \boldsymbol{I} \succ 0$, so the objective is strictly convex. The constraint has Hessian $\boldsymbol{O} = \boldsymbol{\Sigma}^{-1} \succ 0$, yielding a convex (ellipsoidal) feasible set. A strictly convex objective over a non-empty convex feasible set guarantees existence and uniqueness of the global minimizer $\boldsymbol{\Delta}_k^\star$. ∎

Remark 1 (Feasibility): The feasible set is non-empty at every step: the choice $\boldsymbol{\Delta}_k = -\boldsymbol{c}_k$ drives the residual to zero and trivially satisfies the constraint for any $\tau' > 0$.

F. Iterative Attack Construction

The attack evolves incrementally as in Eq. (32)

$$\boldsymbol{a}_k = \boldsymbol{a}_{k-1} + \boldsymbol{\Delta}_k^\star, \tag{42}$$

initialized with $\boldsymbol{a}_0 = \boldsymbol{\Delta}_0^\star$. At each step, the adversary runs an internal simulation to compute $\mathcal{Z}_k$ and $\boldsymbol{Z}_k$, then solves (41) to determine $\boldsymbol{\Delta}_k^\star$.

V. Simulation Results

For reproducibility, the Matlab implementation used in this work has been archived and is publicly available [16].

A. Experimental Setup

The plant is a 6-DOF planar manipulator with link lengths $[0.65, 0.55, 0.45, 0.45, 0.45, 0.45]$ m. The task space comprises planar position (x, y) and one orientation θ_z. The noise covariances are $\boldsymbol{R} = 10^{-4}\boldsymbol{I}$ and $\boldsymbol{Q} = \mathrm{blkdiag}_j\big(q_c\, \boldsymbol{Q}_{\text{base}}\big)$, where $\boldsymbol{Q}_{\text{base}} = [T_s^3/3, T_s^2/2; T_s^2/2, T_s]$ and $q_c = 10^{-2}\,\mathrm{rad}^2/\mathrm{s}^3$, determined with the tuning-knob process.

The nominal task is to maintain the fixed end-effector pose with zero reference velocity and acceleration

$$\bar{p}_0 = \begin{bmatrix} 1.48 \\ 1.81 \end{bmatrix} \text{ m}, \qquad \bar{R}_0 \text{ from } q_0 = [0, \tfrac{\pi}{8}, \cdots, \tfrac{\pi}{8}]^\top.$$

The system is discretized with sampling time $T_s = 3 \cdot 10^{-3}$s. The passive defence (χ^2) uses a sliding window of $W = 100$ samples, and it is calibrated for an Average Run Length (ARL) of 10^6 samples (i.e., one false positive every 0.95 years under $\mathcal{H}_0$), which is realized through Eq. (13) by setting $\tau = 779.28$.

The active defence uses $f(\cdot)$ of Eq. (25) with $\beta = (1 - 10^{-3})$, $\psi = (1 - 10^{-6})$, and $\gamma = 8$, determining $z_p = 50.83$. For Corollary 2, we have formal guarantees

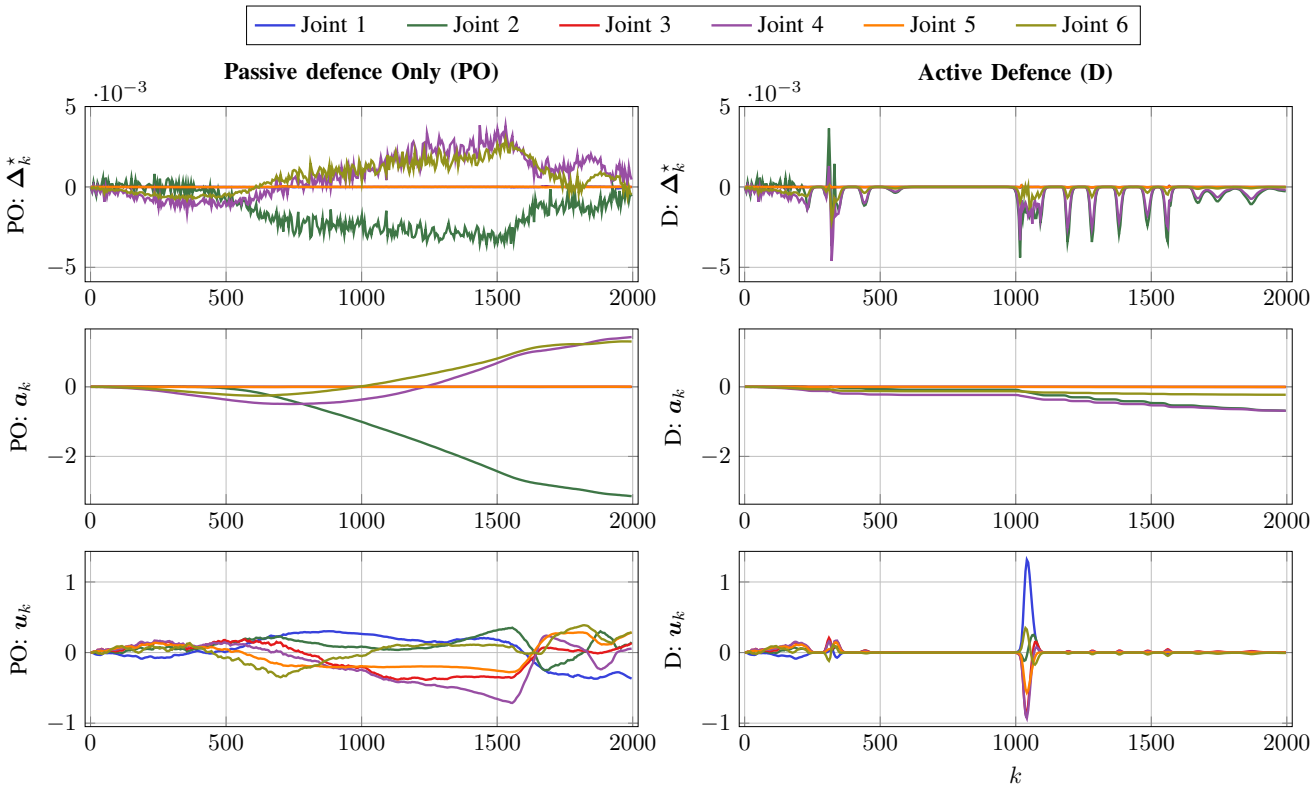

Fig. 4. Adversary and control signals. Optimal increment $\boldsymbol{\Delta}_k^\star$ of Eq. (41), attack sequence $\boldsymbol{a}_k$ of Eq. (32), and resulting actuation $\boldsymbol{u}_k$ of Eqs. (9) and (27). Under the proposed active defence (D), the attacker escalates $\boldsymbol{\Delta}_k^\star$ but scaling suppresses the realized $\boldsymbol{u}_k$, constraining task-space manipulation.

that, on average and under $\mathcal{H}_0$, a single sample of the control command has magnitude below 99.9% of the nominal one every 0.83 hours, which we consider negligible.

The defence adds negligible computational overhead. The QCQP for the attacker (not part of the defence) solves in 40 ms/step with a standard solver (Gurobi).

Attacker.: The attacker's goal is to displace the end-effector from $\bar{\boldsymbol{p}}_0^{\mathrm{A}} = \bar{\boldsymbol{p}}_0$ to $\bar{\boldsymbol{p}}_{(T-1)}^{\mathrm{A}} = [-2, 1]^\top$. This maneuver is interpolated by quintic polynomials with zero boundary velocities and accelerations, generating the reference trajectory $\{\bar{\boldsymbol{p}}_k^{\mathrm{A}}, \dot{\bar{\boldsymbol{p}}}_k^{\mathrm{A}}, \ddot{\bar{\boldsymbol{p}}}_k^{\mathrm{A}}\}$ for $k \in [0, T-1]$. The anomaly budget is distributed uniformly across time, resulting in a per-step constraint $\tau' = \tau/T = 7.79$.

Evaluation Metrics.: Let $\{\boldsymbol{u}_k\}, k = [0, T-1]$ denote a sequence of control commands. The following metric quantifies the control effort:

$$\mathrm{mean}\big(\{\boldsymbol{u}_k\}\big) := \frac{1}{T}\sum_k \big\|\boldsymbol{u}_k\big\|_2.$$

Let $\{\bar{\boldsymbol{p}}_k\}, k = [0, T-1]$ denote a reference hand position trajectory, and $\{\boldsymbol{p}_k\}$ the realized one. The following metrics quantify a task deviation:

$$\mathrm{devmax}\big(\{\bar{\boldsymbol{p}}_k\}, \{\boldsymbol{p}_k\}\big) := \max_k \big\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\big\|_2,$$

$$\mathrm{devRMS}\big(\{\bar{\boldsymbol{p}}_k\}, \{\boldsymbol{p}_k\}\big) := \left(\frac{1}{T}\sum_k \big\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\big\|_2^2\right)^{1/2}.$$

B. Main Results

Undefended (U): When no active or passive defence is in place, the end-effector closely follows the malicious trajectory with a small maximum and RMS deviation from the reference (see first column of Table I).

Passive defence Only (PO): In the presence of the χ^2 ADS using anomaly measure z_k as in Eq. (11), the attacker needs to satisfy the constraint of Eq. (40). The resulting task deviation mildly increases compared to the Undefended case, and the attacker can still reach its final position goal (see the upper diagrams of Fig. 2 and Fig. 3, and the second column of Table I). Throughout, the anomaly measure z_k remains below τ', so the passive ADS never fires alarms (see second diagram of Fig. 3).

With the proposed active defence (D): With the command scaling of Eq. (27) in place, in the initial phase ($k < 250$) the optimal attack increment $\boldsymbol{\Delta}_k^\star$ and the deviation $\|\bar{\boldsymbol{p}}_k^{\mathrm{A}} - \boldsymbol{p}_k\|_2$ mirror PO and U, as $\tilde{z}_k$ is small and $f(\cdot) \approx 1$ (see Figs. 3 and 4). Subsequently, $\tilde{z}_k$ increases determining an increasing scaling of command $\boldsymbol{u}_k$ (see Fig. 4). This incentivizes the attacker to increase injections; z_k quickly climbs to the per-step limit τ' rendering the constraint of Eq. (40) active for a short time.

As the attack progresses, gain reduction $f(\tilde{z}_k)$ disincentivizes the attacker policy of Section IV-C, since increased anomaly would further strengthen scaling. Consequently, $\tilde{z}_k$ tends to settle to an asymptotic value at which the acceleration commands are greatly diminished in magnitude.

TABLE I

TASK DEVIATION AND CONTROL EFFORT METRICS

Metric	U (no defence)	PO (χ^2 only)	D (active defence)
$\mathrm{devmax}\big(\{\bar{\boldsymbol{p}}^{\mathrm{A}}\},\{\boldsymbol{p}\}\big)$	0.05	0.28	2.21
$\mathrm{devmax}\big(\{\bar{\boldsymbol{p}}\},\{\boldsymbol{p}\}\big)$	3.58	3.68	1.5
$\mathrm{devRMS}\big(\{\bar{\boldsymbol{p}}^{\mathrm{A}}\},\{\boldsymbol{p}\}\big)$	0.02	0.13	1.42
$\mathrm{devRMS}\big(\{\bar{\boldsymbol{p}}\},\{\boldsymbol{p}\}\big)$	2.22	2.2	0.82
$\mathrm{mean}\big(\{\boldsymbol{u}_k\}\big)$	0.44	0.46	0.06

Relative to PO, the realized accelerations are significantly attenuated, yielding a much smaller mean control effort (see Table I).

C. Comparative Analysis and Discussion

Without defence, or with passive defence only, stealthy FDIAs can precisely steer the end-effector while remaining within the ADS budget; the passive detector is therefore ineffective in isolation.

With the proposed active defence, the attacker faces the following trade-off: increasing injection raises the proposed anomaly-aware command scaling, while reducing injection makes the stealthy attack less effective. Overall, the attack decreases in effectiveness and the proposed anomaly-aware score $\tilde{z}_k$ settles to an asymptotic value. This plateau represents the sub-optimal injection magnitude the attacker is forced to commit to, akin to an equilibrium in a Stackelberg game. Fig. 3 explains the closed-loop interaction: (i) the attacker increases $\boldsymbol{\Delta}_k^\star$ to counter scaling (Fig. 4, top), (ii) this raises z_k until the per-step constraint becomes active, (iii) meanwhile $\tilde{z}_k$ (based on the actuation-projected predictor) tracks the accumulating state discrepancy and drives $f(\tilde{z}_k)$ down, (iv) the realized $\boldsymbol{u}_k$ is therefore attenuated, limiting achievable task-space acceleration and displacement.

An important consideration regards the safety of the proposed active defence in the absence of attacks. By selecting the design confidence ψ and gain floor β via Corollary 2, unwarranted scaling under $\mathcal{H}_0$ can be made arbitrarily rare, although statistical fluctuations can in principle still trigger noticeable attenuation. However, because the gain scaling acts on the command $\boldsymbol{u}_k$ prior to the inverse-dynamics compensation, it does not alter gravity or Coriolis compensation. Finally, extending the stability analysis to the time-varying case of $f(\tilde{z}_k)$ is left for future work; promising directions include (i) low-pass filtering $f(\tilde{z}_k)$, (ii) enforcing a probabilistic lower bound $f \geq \beta$, and (iii) invoking slowly-varying stability results.

VI. CONCLUSION

This paper addressed the resilience of robotic manipulators against stealthy false data injection attacks (FDIAs). We showed that feedback linearization induces an *integrator vulnerability* that allows persistent sensor corruption to remain undetected by χ^2 anomaly detectors while driving the end-effector off-task. To counter this threat, we introduced *anomaly-aware command scaling*, a lightweight modification of the control law that attenuates inputs as a function of an anomaly score derived from a measurement-free, actuation-projected predictor. Our analysis established two key guarantees: (i) probabilistic bounds on actuation loss in nominal operation, enabling minimally invasive deployment, and (ii) preservation of closed-loop stability under bounded attenuation.

On the adversary side, we derived a convex QCQP formulation of the one-step optimal stealthy attack that incorporates the defender's policy, yielding a Stackelberg-type game-theoretic benchmark against which to evaluate defences. Simulation results on a 6-DOF manipulator confirmed that the proposed defence substantially reduces attacker-induced task-space deviations while maintaining nominal tracking performance in the absence of attacks.

Overall, the proposed approach illustrates how active modification of the control law can transform anomaly detection into a practical resilience mechanism for cybersecure robotic manipulation.

REFERENCES

[1] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security–A Survey," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1802–1831, 2017.

[2] H. Sandberg, V. Gupta, and K. H. Johansson, "Secure networked control systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. 1, pp. 445–464, 2022.

[3] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 4–13, 2017.

[4] A. Intriago, F. Liberati, N. D. Hatziargyriou, and C. Konstantinou, "Residual-based detection of attacks in cyber-physical inverter-based microgrids," *IEEE Trans. Power Syst.*, vol. 39, no. 2, pp. 4020–4038, 2024.

[5] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton Conf. Comm., Control, and Comp. (Allerton)*, 2009, pp. 911–918.

[6] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[7] J. Ueda and J. Blevins, "Affine transformation-based perfectly undetectable false data injection attacks on remote manipulator kinematic control with attack detector," *IEEE Robot. Autom. Lett.*, vol. 9, no. 10, pp. 8690–8697, 2024.

[8] F. E. Tosun, A. M. Teixeira, J. Dong, A. Ahlén, and S. Dey, "Kullback-liebler divergence-based observer design against sensor bias injection attacks in single-output systems," *IEEE Transactions on Information Forensics and Security*, 2025.

[9] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[10] T. R, C. Murguia, and J. Ruths, "Tuning windowed chi-squared detectors for sensor attacks," in *American Control Conference (ACC)*, 2018, pp. 1752–1757.

[11] C. Murguia and J. Ruths, "CUSUM and chi-squared attack detection of compromised sensors," in *IEEE Conf. Control Appl. (CCA)*, 2016, pp. 474–480.

[12] R. M. Murray, Z. Li, and S. Sastry, *A mathematical introduction to robotic manipulation*, 1st ed. Boca Raton: CRC Press, 1994.

[13] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: Modelling, Planning and Control*, ser. Advanced Textbooks in Control and Signal Processing. Springer London, 2009.

[14] G. Gallego, C. Cuevas, R. Mohedano, and N. Garcia, "On the Mahalanobis distance classification criterion for multidimensional normal distributions," *IEEE Trans. Signal Process.*, vol. 61, no. 17, pp. 4387–4396, 2013.

[15] K. M. Ramachandran and C. P. Tsokos, *Sampling distributions*, 3rd ed. Elsevier, 2021, pp. 147–177.

[16] G. Gualandi, "Manipulatorsgainscaling," 2026, v1.0.1. [Online]. Available: https://doi.org/10.5281/zenodo.18877674