Active Defense Against False Data Injection Attacks in Robotic Manipulators

Gabriele Gualandi * **Carl Mikael Larsson** *
Alessandro V. Papadopoulos *

* *Mälardalen University, Västerås, Sweden*
(e-mail: gabriele.gualandi@mdu.se).

Abstract: Robotic systems are vulnerable to *False Data Injection Attacks* (FDIAs), where adversaries corrupt sensor signals to gain malicious control. Feedback linearization exposes robotic systems to *integrator vulnerability*, exposing them to *stealthy* attacks that can cause significant deviations in end-effector behavior without raising alarms. This paper addresses the resilience of manipulators against finite-horizon FDIAs by formalizing two defense methods, namely *anomaly-aware virtual damping* and *manipulability reduction*, with probabilistic guarantees on nominal task execution. Simulations on a 7-DOF redundant manipulator show that the proposed defense substantially reduces the impact of FDIA compared to using solely a threshold-based ADS like the χ^2, while preserving nominal task performance in the absence of attack.

1. 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 (Humayed et al., 2017; Sandberg et al., 2022). 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 detection while still driving the end-effector away from its intended task (Guo et al., 2017; Intriago et al., 2024).

A well-studied class of integrity attacks are *False Data Injection Attacks* (FDIAs) targeting sensing, 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 (Mo and Sinopoli, 2009; Ueda and Blevins, 2024; Fawzi et al., 2014) 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 (Ueda and Blevins, 2024). One perspective has been largely overlooked which relates to a key robotics-specific vulnerability: the widespread use of *feedback linearization* reduces manipulator dynamics to double integrators, which in turn introduces an *integrator vulnerability* (Tosun et al., 2025). As a consequence, persistent sensor corruption can silently accumulate in the closed-loop system.

⋆ This work was supported by the Knowledge Foundation (KKS).

This paper addresses this gap by introducing two *active defense* strategies that limit the impact of finite-horizon FDIA targeting sensors for feedback-linearized robotic arms. The proposed defense methods leverages an *actuation-projected anomaly score* that ignores sensing and hence is immune to sensor corruption. Said anomaly score is used in two ways. Firstly, we introduce *virtual damping* for the joints, to dissipate kinetic energy as anomalies grow, thereby reducing the adversary's ability to steer the manipulator. Secondly, we reconfigure the manipulator's null space to *reduce manipulability* in the estimated attack hand direction, which increases the joint-level anomaly required for a given hand-level displacement. Since the strength of the virtual damping, and the detection of a standard χ^2 ADS are both driven by joint-level anomaly, null-space reconfiguration can further limit the impact of FDIA.

The contributions of this work are the following:

(1) We formalize stealthy FDIAs against feedback-linearized manipulators, exposing their integrator vulnerability, and show that the attacker's one-step optimal strategy reduces to a convex QCQP.
(2) We propose *anomaly-aware virtual damping*, which attenuates joint velocities based on a measurement-free actuation-projected state predictor, with probabilistic guarantees on bounded attenuation in nominal operation. Furthermore, we propose to *decrease manipulability* in the estimated attack direction to further reduce the impact of FDIA, with no impact on task performance in nominal operation.
(3) Simulations on a 7-DOF real-world robot shows that the defenses significantly limits FDIA compared to threshold-based ADSs like the χ^2.

1.1 Related Work

Threshold-based detectors, particularly χ^2 tests on Kalman innovations, are a classical tool for monitoring

CPS integrity (Ding et al., 2018). These methods provide statistical guarantees on false-alarm rates and are widely used in industrial practice. Extensions include adaptive thresholds (Tunga et al., 2018) and sequential schemes such as CUSUM tests (Murguia and Ruths, 2016). However, such ADS are inherently *passive*: they may detect anomalies but do not alter the control policy, leaving the system structure unchanged.

Theoretical studies of FDIAs in networked systems have characterized undetectability conditions (Mo and Sinopoli, 2009), affine attack structures (Ueda and Blevins, 2024), and optimal attack strategies (Guo et al., 2017). These works typically assume 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 (Intriago et al., 2024).

Control systems with a LTI plant having integral action are subject to *integrator vulnerability* (Tosun et al., 2025), where constant attack injections get absorbed by any linear observer, resulting in attacks only being detectable during transients. Classical passive threshold-based detection cannot prevent a stealth FDIA from increasingly deceiving a state estimator.

In (Gualandi and Papadopoulos, 2026), we demonstrated that feedback linearization induces integrator vulnerability in robotic systems, and we proposed gain scaling based on a state-projected (sensing-free) anomaly score to limit the accumulation of kinetic energy during a FDIA. While gain scaling mitigates this energy accumulation, it is fundamentally unable to dissipate it. Therefore, this work introduces a novel *virtual damping* strategy designed to actively dissipate the kinetic energy generated by a FDIA, while maintaining probabilistic performance guarantees during nominal operation. Furthermore, we propose directional *manipulability reduction* to compel the adversary to increase its signal injection for a given task-space displacement, thereby further curtailing its control authority. This establishes kinematic redundancy not only as a mechanism for performance enhancement but also as a tool for cyber-physical security. Although redundant manipulators have long leveraged null-space projections for secondary tasks, such as obstacle clearance (Siciliano and Slotine, 1991), their application to adversarial robustness remains largely unexplored. To the best of our knowledge, exploiting redundancy to *reduce manipulability along an adversarial direction* as an active security defense has not been previously investigated.

2. SYSTEM MODEL

2.1 Closed-Loop Architecture

We consider a robotic manipulator operating in closed loop with a state estimator and a task-space controller. At each discrete time step $k \in \mathbb{Z}_{\geq 0}$, the plant output $\boldsymbol{y}_k$ is corrupted by injected signal $\boldsymbol{a}_k$ to yield measurement

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

The plant has discrete-time linear dynamics with input saturation

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

with state $\boldsymbol{x}_k$, process noise $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$, measurement noise $\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$, and input $\boldsymbol{u}_k$ saturated by the $\mathrm{sat}(\cdot)$ operator within limits $[\boldsymbol{u}_{\min}, \boldsymbol{u}_{\max}]$, $\boldsymbol{u}_{\min} < 0, \boldsymbol{u}_{\max} > 0$. 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 $\boldsymbol{r}_k$ is the innovation (residual). We assume $(\boldsymbol{A}, \boldsymbol{C})$ is detectable, ensuring the existence of a unique stabilizing solution $\boldsymbol{P}$ to the DARE

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^\top + \boldsymbol{Q} - \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^\top\big(\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\top + \boldsymbol{R}\big)^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{A}^\top. \tag{4}$$

Here $\boldsymbol{P}$ is the steady-state covariance of the signal $\boldsymbol{e}_k = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k$. The steady-state covariance of $\boldsymbol{r}_k$ is

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\top + \boldsymbol{R} \succ 0 \tag{5}$$

(hence invertible), and the steady-state gain is

$$\boldsymbol{L} = \boldsymbol{A}\,\boldsymbol{P}\,\boldsymbol{C}^\top\,\boldsymbol{\Sigma}^{-1}. \tag{6}$$

In the remainder of the paper, symbols with the hat $\hat{\cdot}$ are quantities determined using the estimated state $\hat{\boldsymbol{x}}$, which is subject to FDIA.

2.2 Manipulator Model and Task-Space Control

We consider an n-DOF robotic manipulator with joint variable $\boldsymbol{q} \in \mathbb{R}^n$, state $\boldsymbol{x}_k = [\boldsymbol{q}_k^\top\ \dot{\boldsymbol{q}}_k^\top]^\top \in \mathbb{R}^{2n}$ and measurement vector $\boldsymbol{y}_k \in \mathbb{R}^p$ (assuming full state measurement, $p = 2n$). The continuous-time joint-space dynamics with joint torques $\boldsymbol{\tau}$ is:

$$\boldsymbol{B}(\boldsymbol{q})\,\ddot{\boldsymbol{q}} + \boldsymbol{\nu}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \boldsymbol{\tau}, \tag{7}$$

where $\boldsymbol{B}(\boldsymbol{q}) \succ 0$ is the inertia matrix and $\boldsymbol{\nu}$ collects Coriolis, centrifugal, and gravity terms. With inverse-dynamics compensation $\boldsymbol{\tau} = \boldsymbol{B}(\boldsymbol{q})\,\boldsymbol{u} + \boldsymbol{\nu}(\boldsymbol{q}, \dot{\boldsymbol{q}})$, the joint dynamics reduce to decoupled double integrators (*virtual masses system*), $\ddot{\boldsymbol{q}} = \boldsymbol{u}$, whose discrete-time form is (2a) and whose residual is $\boldsymbol{r}_k \in \mathbb{R}^p$ from (3b).

The controller operates in task space using twist control (Siciliano et al., 2009), tracking position $\{\bar{\boldsymbol{p}}_k, \dot{\bar{\boldsymbol{p}}}_k, \ddot{\bar{\boldsymbol{p}}}_k\}$ and orientation $\{\bar{\boldsymbol{R}}_k, \bar{\boldsymbol{\omega}}_k, \dot{\bar{\boldsymbol{\omega}}}_k\}$ references, with $\bar{\boldsymbol{R}}_k \in SO(3)$, $\bar{\boldsymbol{p}}_k \in \mathbb{R}^3$ and $\bar{\boldsymbol{\omega}}_k \in \mathbb{R}^3$. A PD+feedforward law provides task accelerations:

$$\boldsymbol{u}_k^{\mathrm{PD}} = \begin{bmatrix} \ddot{\bar{\boldsymbol{p}}}_k + K_{pp}\,\boldsymbol{e}_{\boldsymbol{p},k} + K_{dp}\,\dot{\boldsymbol{e}}_{\boldsymbol{p},k} \\ \dot{\bar{\boldsymbol{\omega}}}_k + K_{po}\,\boldsymbol{e}_{\boldsymbol{o},k} + K_{do}\,\boldsymbol{e}_{\omega,k} \end{bmatrix}, \tag{8}$$

where: $\boldsymbol{e}_{\boldsymbol{p},k} = \bar{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}_k$ and $\boldsymbol{e}_{\boldsymbol{o},k} = \sin(\hat{\theta}_k/2)\,\hat{\boldsymbol{r}}_k$ are respectively the position and orientation errors; $\boldsymbol{e}_{\boldsymbol{o},k}$ is computed using the angle-axis pair $(\hat{\theta}_k, \hat{\boldsymbol{r}}_k)$ from matrix $\bar{\boldsymbol{R}}_k\,\hat{\boldsymbol{R}}_k^\top$, $\hat{\theta}_k \in [0, \pi]$; $\boldsymbol{e}_{\omega,k} = \bar{\boldsymbol{\omega}}_k - \hat{\boldsymbol{\omega}}_k$; and the gains $K_{pp}, K_{dp}, K_{po}, K_{do}$ are synthesized via discrete-time LQR.

Joint accelerations are computed via pseudoinverse $\boldsymbol{J}^\star(\hat{\boldsymbol{q}})$ of the geometric Jacobian (Siciliano et al., 2009) as:

$$\boldsymbol{u}_k^{\mathrm{nom}} = \boldsymbol{J}_k^\star\left(\boldsymbol{u}_k^{\mathrm{PD}} - \dot{\hat{\boldsymbol{J}}}_k\,\dot{\hat{\boldsymbol{q}}}\right) + \boldsymbol{u}_k^{\mathrm{sec}} \tag{9}$$

where $\boldsymbol{u}_k^{\mathrm{sec}}$ is later used for an active defense lying in the null-space of the Jacobian, therefore not interfering with the primary task under $\mathcal{H}_0$ (no attack). In the absence of active defenses, in (2a) we set $\boldsymbol{u}_k = \boldsymbol{J}_k^\star\left(\boldsymbol{u}_k^{\mathrm{PD}} - \dot{\hat{\boldsymbol{J}}}_k\,\dot{\hat{\boldsymbol{q}}}\right)$.

2.3 χ^2 Anomaly Detection System

The system incorporates a detector to distinguish between the null hypothesis of normal operation, $\mathcal{H}_0$, and the alternative hypothesis of an attack, $\mathcal{H}_1$. The χ^2 detector monitors the statistic (Mahalanobis distance):

$$z_k = \boldsymbol{r}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_k \tag{10}$$

with $\boldsymbol{\Sigma}$ as in (5) and $\boldsymbol{r}_k$ as in (3b). During execution, an alarm is triggered if $z_k > \tau$, where $\tau > 0$ is a parameter (*threshold*). Under $\mathcal{H}_0$, the residual $\boldsymbol{r}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, hence z_k follows a $\chi^2(p)$ distribution, where $p = \dim(\boldsymbol{r}_k)$ (Ding et al., 2018). Given a desired per-step false-alarm probability $\alpha \in (0, 1)$ (equivalently, a desired mean inter-alarm interval $\mathrm{ARL} = 1/\alpha$), setting

$$\tau \;=\; F^{-1}_{\chi^2(p)}(1-\alpha) \;=\; 2\,\mathrm{P}^{-1}\Big(1-\alpha, \frac{p}{2}\Big), \tag{11}$$

ensures that $\mathbb{P}(z_k > \tau \mid \mathcal{H}_0) = \alpha$, where $F^{-1}_{\chi^2(p)}$ is the inverse CDF of the χ^2 distribution with p degrees of freedom, and P^{-1} denotes the inverse of the regularized lower incomplete gamma function (Tunga et al., 2018).

2.4 Assumptions

There are the following assumptions, under which both the defense and attack strategies are developed.

Assumption 1. (Adversary knowledge). The attacker has full knowledge of the plant, controller, state estimator, ADS and active defense system(s).

Assumption 2. (Adversary capabilities). The only attack surface is additive injection into sensor measurements FDIA targeting sensors, as in (1).

Assumption 3. (Adversary goal). The attacker's goal is to control the position of the end-effector in task space while remaining undetected by any threshold-based ADS such as the χ^2 (stealth FDIA).

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 has converged (covariances and gains are constant).

3. PROPOSED DEFENCE METHODS

3.1 Attack Estimation

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

$$\tilde{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0. \tag{12}$$

For subsequent steps, $\tilde{\boldsymbol{x}}_k$ follows the open-loop predictor

$$\tilde{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\,\tilde{\boldsymbol{x}}_k + \boldsymbol{B}\,\boldsymbol{u}_k, \qquad k \geq 0. \tag{13}$$

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

$$\tilde{\boldsymbol{r}}_k = \hat{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_k \in \mathbb{R}^{2n}. \tag{14}$$

Theorem 1. (Actuation-projected residual covariance). Under $\mathcal{H}_0$ (no attack) the covariance of the residual in (14) after k steps from initialization, $\boldsymbol{\Sigma}_{\tilde{r},k} \in \mathbb{R}^{2n\times 2n}$, is the $(2,2)$ block of the matrix $\boldsymbol{P}_{z,k} \in \mathbb{R}^{4n\times 4n}$, where blocks have size $2n \times 2n$, obtained by iterating

$$\boldsymbol{P}_{z,j+1} = \boldsymbol{F}\boldsymbol{P}_{z,j}\boldsymbol{F}^\top + \boldsymbol{\Pi}, \quad j = 0, \ldots, k-1, \tag{15}$$

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

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{L}\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{L}\boldsymbol{C} & \boldsymbol{A} \end{bmatrix} \in \mathbb{R}^{4n\times 4n}, \tag{16}$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{L} \\ \boldsymbol{0} & \boldsymbol{L} \end{bmatrix} \in \mathbb{R}^{4n\times(2n+p)}, \tag{17}$$

$$\boldsymbol{\Pi} = \boldsymbol{Z}\ \mathrm{diag}(\boldsymbol{Q}, \boldsymbol{R})\boldsymbol{Z}^\top \in \mathbb{R}^{4n\times 4n}. \tag{18}$$

and $\boldsymbol{P}$ is defined in (4).

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 $\tilde{\boldsymbol{r}}_k := \hat{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_k$ are:

$$\begin{aligned} \boldsymbol{e}_{k+1} &= (\boldsymbol{A} - \boldsymbol{L}\boldsymbol{C})\boldsymbol{e}_k + \boldsymbol{w}_k - \boldsymbol{L}\boldsymbol{v}_k, \\ \tilde{\boldsymbol{r}}_{k+1} &= \boldsymbol{A}\tilde{\boldsymbol{r}}_k + \boldsymbol{L}\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{L}\boldsymbol{v}_k. \end{aligned}$$

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

$$\boldsymbol{z}_{k+1} = \boldsymbol{F}\boldsymbol{z}_k + \boldsymbol{Z}\boldsymbol{\eta}_k,$$

with $\boldsymbol{F}, \boldsymbol{Z}$ as in (16), (17) and $\boldsymbol{\eta}_k \triangleq [\boldsymbol{w}_k^\top, \boldsymbol{v}_k^\top]^\top$. The covariance propagates as

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

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 (18). At synchronization $k = 0$, for (12) and Assumption 4, we have $\tilde{\boldsymbol{r}}_0 = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}_0) = \boldsymbol{P}$, hence

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

Iterating (15) k times yields $\boldsymbol{P}_{z,k}$; the desired covariance $\boldsymbol{\Sigma}_{\tilde{r},k} \in \mathbb{R}^{2n\times 2n}$ is its lower-right $2n \times 2n$ submatrix. □

Theorem 2. (Confidence for the Anomaly Measure). Consider the statistic (*projected anomaly score*)

$$\tilde{z}_k = \tilde{\boldsymbol{r}}_k^\top \boldsymbol{\Sigma}^{-1}_{\tilde{r},k} \tilde{\boldsymbol{r}}_k. \tag{19}$$

Choosing $z_{\mathrm{x}} = F^{-1}_{\chi^2(2n)}(\psi)$ ensures $\mathbb{P}(\tilde{z}_k \leq z_{\mathrm{x}} \mid \mathcal{H}_0) = \psi$ for any given $z_{\mathrm{x}} > 0$ and 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 Thm. 1, ensures that the resulting statistic has a stationary distribution, $\tilde{z}_k \sim \chi^2(2n)$, for all k. The theorem follows from the definition of the inverse CDF, $F^{-1}_{\chi^2(2n)}$. □

3.2 Virtual Damping Active Defense

We propose virtual damping to enforce a passivity margin on the virtual masses system based on $\tilde{z}_k$ as in (19).

Let $F > 0$ be the desired asymptotic power dissipation, $\gamma > 0$ a design parameter, and let $z_{\mathrm{x}} > 0$ be a design abscissa. Let $\beta \in (0, F)$ be the desired power dissipation ratio at $\tilde{z} = z_{\mathrm{x}}$. We define the *target power dissipation*

ratio, $g(\tilde{z})$, as a smooth, monotonically increasing sigmoid function:

$$g(\tilde{z}_k) = F\left(1 - \exp\left[-\left(\frac{\tilde{z}_k}{z_\mathrm{s}}\right)^\gamma\right]\right), \tag{20}$$

$$z_\mathrm{s} = z_\mathrm{x}\left(-\ln\left(1 - \frac{\beta}{F}\right)\right)^{-1/\gamma}, \tag{21}$$

satisfying $g(0) = 0$, $g(\tilde{z}_k) \geq 0$, $\lim_{z\to\infty} g(z) = F$ and $g(z_\mathrm{x}) = \beta$.

Assumption 5. (Model fidelity). We assume that $\tilde{\boldsymbol{x}}$ always remains accurate over a finite-horizon FDIA on sensors i.e.,

$$\tilde{\boldsymbol{x}}_k \approx \boldsymbol{x}_k, k \in (0, T). \tag{22}$$

Assumption 5 is only needed for the accuracy of the defense control laws based on actuation-projected state $\tilde{\boldsymbol{x}}_k$, as developed in the rest of this work. It is reasonable when process noise remains small over the attack horizon T (which is limited by mechanisms such as periodic software rejuvenation, key rotation, re-establishment of secure channels, or moving-target defenses) and when $\tilde{\boldsymbol{x}}_k$ is regularly resynchronized with $\hat{\boldsymbol{x}}_k$ as in (12).

Theorem 3. (Adaptive damping). Consider a robotic manipulator with nominal joint-space control input $\boldsymbol{u}_k^\mathrm{nom}$ (9) subject to the saturation operator $\mathrm{sat}(\cdot)$ (2a). Let the (projected) nominal power of the j-th joint be defined as

$$\tilde{P}_{k,j}^\mathrm{nom} = \mathrm{sat}(\boldsymbol{u}_k^\mathrm{nom})_j \cdot \dot{\tilde{q}}_{k,j} \tag{23}$$

where $\dot{\tilde{\boldsymbol{q}}}_k$ is extracted from $\tilde{\boldsymbol{x}}_k$. Let $\tilde{z}_k \in [0, \infty)$ denote the anomaly score defined in (19), and let $g : [0, \infty) \to [0, \infty)$ be a non-decreasing function satisfying $g(0) = 0$.

Define the ideal, unconstrained adaptive damping command, $\boldsymbol{u}_k^\mathrm{d,ideal}$, on a per-joint basis as

$$u_{k,j}^\mathrm{d,ideal} = \begin{cases} -g(\tilde{z}_k)\dfrac{|\tilde{P}_{k,j}^\mathrm{nom}|}{\dot{\tilde{q}}_{k,j}}, & \text{if } |\dot{\tilde{q}}_{k,j}| > \epsilon, \\ 0, & \text{otherwise,} \end{cases} \tag{24}$$

with $g(\cdot)$ an arbitrary non-decreasing function satisfying $g(0) = 0$ (such as the one defined in (20)) and $\epsilon \approx 0$.

Let us define the interval $[\mathbf{h}_k^-, \mathbf{h}_k^+]$, where

$$\mathbf{h}_k^- = \boldsymbol{u}_\mathrm{min} - \mathrm{sat}(\boldsymbol{u}_k^\mathrm{nom}) \in [\boldsymbol{u}_\mathrm{min} - \boldsymbol{u}_\mathrm{max}, \mathbf{0}], \tag{25a}$$

$$\mathbf{h}_k^+ = \boldsymbol{u}_\mathrm{max} - \mathrm{sat}(\boldsymbol{u}_k^\mathrm{nom}) \in [\mathbf{0}, \boldsymbol{u}_\mathrm{max} - \boldsymbol{u}_\mathrm{min}], \tag{25b}$$

are the actuator headrooms available for the defense.

Define the *maximal non-saturating damping* command, $\boldsymbol{u}_k^\mathrm{d}$, as:

$$\boldsymbol{u}_k^\mathrm{d} = \max(\mathbf{h}_k^-, \min(\mathbf{h}_k^+, \boldsymbol{u}_k^\mathrm{d,ideal})). \tag{26}$$

If the final control law is

$$\boldsymbol{u}_k = \mathrm{sat}(\boldsymbol{u}_k^\mathrm{nom}) + \boldsymbol{u}_k^\mathrm{d}, \tag{27}$$

then, under Assumption 5:

Claim 1. (Stability). Command $\boldsymbol{u}_k^\mathrm{d}$ acting on the virtual masses system has non-positive joint-wise power, i.e.:

$$u_{k,j}^\mathrm{d} \cdot \dot{q}_{k,j} \leq 0, \quad \forall j = 1, \ldots, n.$$

Hence, for the fundamental results in passivity theory (Van der Schaft, 2000), it does not decrease the stability margin of the closed loop system.

Claim 2. (Probabilistic Magnitude Bound). Let $\tilde{z}_k \sim \chi^2(2n)$ under $\mathcal{H}_0$, and let $z_\mathrm{x} = F^{-1}_{\chi^2(2n)}(\psi)$ denote the inverse CDF at confidence level $\psi \in (0, 1)$. If design parameters of $g(\cdot)$ are chosen such that $g(z_\mathrm{x}) \leq \beta$ for some $\beta > 0$, then with probability at least ψ, the magnitude of the total power dissipated by the defense is bounded by a fraction β of the total absolute nominal power:

$$\sum_{j=1}^n |u_{k,j}^\mathrm{d} \cdot \dot{\tilde{q}}_{k,j}| \leq \beta \sum_{j=1}^n |\tilde{P}_{k,j}^\mathrm{nom}|. \tag{28}$$

Proof. For Assumption 5, any projected power (computed using $\dot{\tilde{\boldsymbol{q}}}$) coincides with its real version (from $\dot{\boldsymbol{q}}$).
To prove Claim 1, let us denote the per-joint power of the ideal command as $\tilde{P}_{k,j}^\mathrm{d,ideal} = u_{k,j}^\mathrm{d,ideal} \cdot \dot{\tilde{q}}_{k,j}$. By substituting (24) into this definition, and observing that $\dot{\tilde{q}}_{k,j}/\dot{\tilde{q}}_{k,j} = 1$, we obtain:

$$\tilde{P}_{k,j}^\mathrm{d,ideal} = -g(\tilde{z}_k)|\tilde{P}_{k,j}^\mathrm{nom}| \leq 0, \tag{29}$$

since $g(\cdot) \geq 0$ and $|\tilde{P}_{k,j}^\mathrm{nom}| \geq 0$. This proves that the ideal command is per-joint passive. Next, consider the final command $u_{k,j}^\mathrm{d}$ from (26). Since $\mathrm{sat}(\boldsymbol{u}_k^\mathrm{nom})$ is within actuator limits $[\boldsymbol{u}_\mathrm{min}, \boldsymbol{u}_\mathrm{max}]$, the headroom intervals $[\mathbf{h}_k^-, \mathbf{h}_k^+]$ from (25a)–(25b) are guaranteed to contain the origin. Therefore, the clipping operation in (26) cannot change the sign, nor increase the absolute value, of $u_{k,j}^\mathrm{d,ideal}$, leading to:

$$|u_{k,j}^\mathrm{d} \cdot \dot{\tilde{q}}_{k,j}| \leq |u_{k,j}^\mathrm{d,ideal} \cdot \dot{\tilde{q}}_{k,j}|, \tag{30}$$

$$\mathrm{sign}(u_{k,j}^\mathrm{d} \cdot \dot{\tilde{q}}_{k,j}) = \mathrm{sign}(u_{k,j}^\mathrm{d,ideal} \cdot \dot{\tilde{q}}_{k,j}). \tag{31}$$

By (30) and (31), the passivity condition (29) still holds after clipping, hence (26) is per-joint passive.

To prove Claim 2, consider the absolute total power of the ideal defense command. Substituting (29) into the definition of total power yields:

$$\sum_{j=1}^n |u_{k,j}^\mathrm{d,ideal} \cdot \dot{\tilde{q}}_{k,j}| = \sum_{j=1}^n |-g(\tilde{z}_k)|\tilde{P}_{k,j}^\mathrm{nom}|| = g(\tilde{z}_k) \sum_{j=1}^n |\tilde{P}_{k,j}^\mathrm{nom}|. \tag{32}$$

Summing (30) over $j = 1, \ldots, n$ and substituting the total ideal power from (32), we obtain the total power bound:

$$\sum_{j=1}^n |u_{k,j}^\mathrm{d} \cdot \dot{\tilde{q}}_{k,j}| \leq \sum_{j=1}^n |u_{k,j}^\mathrm{d,ideal} \cdot \dot{\tilde{q}}_{k,j}| = g(\tilde{z}_k) \sum_{j=1}^n |\tilde{P}_{k,j}^\mathrm{nom}|. \tag{33}$$

Under $\mathcal{H}_0$, the anomaly score follows $\tilde{z}_k \sim \chi^2(2n)$. By the definition of the quantile $z_\mathrm{x} = F^{-1}_{\chi^2(2n)}(\psi)$, we have $\mathbb{P}(\tilde{z}_k \leq z_\mathrm{x}) = \psi$. On this event, since $g(\cdot)$ is non-decreasing, we have $g(\tilde{z}_k) \leq g(z_\mathrm{x})$. By design, $g(z_\mathrm{x}) \leq \beta$. Substituting this into (33) yields

$$\sum_{j=1}^n |u_{k,j}^\mathrm{d} \cdot \dot{\tilde{q}}_{k,j}| \leq \beta \sum_{j=1}^n |\tilde{P}_{k,j}^\mathrm{nom}|.$$

with probability ψ. Therefore, with probability ψ, the total dissipated power is bounded by a fraction β of the nominal absolute power, which completes the proof. □

Claim 2 of Thm. 3 provides a formal method for tuning the defense's nominal impact. By choosing $z_\mathrm{x} = F^{-1}_{\chi^2(2n)}(\psi)$, one can guarantee with an arbitrarily high probability ψ (e.g., $\psi = 0.99$) that the power of the adaptive friction will be less than a desired small fraction β (e.g., $\beta = 0.01$)

of the nominal control power, so that the defense can be configured to be minimally invasive under $\mathcal{H}_0$. A value of $F > 1$ enforces asymptotic strict passivity of the virtual masses system as the projected anomaly $\tilde{z}_k$ grows.

3.3 Manipulability Reduction Active Defense

We now present a defense strategy that leverages the manipulator's kinematic redundancy. Specifically, we reduce manipulability in the estimated attack direction to increase the required anomaly injected for a given hand displacement. We restrict attack direction estimation to the translational DOFs, as these are most relevant to adversarial manipulation. The estimated attack direction is defined as

$$\boldsymbol{d}_k = \frac{\tilde{\boldsymbol{p}}_k - \bar{\boldsymbol{p}}_k + \epsilon}{\|\tilde{\boldsymbol{p}}_k - \bar{\boldsymbol{p}}_k + \epsilon\|_2}, \tag{34}$$

where $\boldsymbol{d}_k \in \mathbb{R}^3, \|\boldsymbol{d}_k\|_2 = 1$, $\tilde{\boldsymbol{p}}_k$ is the hand position from $\tilde{\boldsymbol{q}}_k$, and $\epsilon \approx 0$ avoids degeneracy.

The positional manipulability measure in direction $\boldsymbol{d}_k$ is

$$\eta_k = \boldsymbol{d}_k^\top \boldsymbol{\mathcal{M}}_k \boldsymbol{d}_k, \tag{35}$$

where $\boldsymbol{\mathcal{M}}_k = \tilde{\mathbf{J}}_{\mathrm{p},k} \tilde{\mathbf{J}}_{\mathrm{p},k}^\top$ is the directional manipulability matrix, and $\tilde{\mathbf{J}}_{\mathrm{p},k}$ is the positional geometric Jacobian from the projected joint configuration $\tilde{\boldsymbol{q}}_k$.

We adopt the redundancy resolution scheme based on projected gradient method (De Luca and Oriolo, 1991) with the goal of reducing manipulability along $\boldsymbol{d}_k$, using cost function:

$$\mathcal{C}_k = \tfrac{1}{2}\eta_k^2. \tag{36}$$

Treating $\boldsymbol{d}_k$ as quasi-static allows for the gradient of the cost to be approximated as

$$\nabla_{\boldsymbol{q}} \mathcal{C}_k \approx \eta_k \begin{bmatrix} d_k^\top \frac{\partial \boldsymbol{\mathcal{M}}_k}{\partial q^{(1)}} d_k \\ \vdots \\ d_k^\top \frac{\partial \boldsymbol{\mathcal{M}}_k}{\partial q^{(n)}} d_k \end{bmatrix}. \tag{37}$$

We introduce the following secondary null-space command for control law $\boldsymbol{u}_k^{\mathrm{nom}}$ (9):

$$\boldsymbol{u}_k^{\mathrm{sec}} = \begin{cases} \mathbf{0}, & \text{if } \nabla_{\boldsymbol{q}} \mathcal{C}_k \approx 0 \\ -\nu_k \big(\boldsymbol{I} - \hat{\mathbf{J}}_k^\dagger \hat{\mathbf{J}}_k\big) \nabla_{\boldsymbol{q}} \mathcal{C}_k, & \text{otherwise} \end{cases} \tag{38}$$

where $\hat{\mathbf{J}}_k^\dagger$ is the Moore–Penrose inverse, $\nu_k \in (0, \nu_{\max}]$ is a step size selected by an Armijo line search.

To penalize individual per-joint motion and maintain low manipulability when $\mathcal{C}$ is stationary i.e., $\nabla_{\boldsymbol{q}} \mathcal{C}_k \approx 0$, we use the following weighted pseudoinverse for $\boldsymbol{u}_k^{\mathrm{nom}}$ (9):

$$\boldsymbol{J}^\star = \mathbf{W}_k^{-1} \hat{\mathbf{J}}_k^\top \big(\hat{\mathbf{J}}_k \mathbf{W}_k^{-1} \hat{\mathbf{J}}_k^\top\big)^{-1}, \tag{39}$$

with weighting

$$\mathbf{W}_k = \begin{cases} \boldsymbol{D} + \alpha\big(\nabla_{\boldsymbol{q}}^2 \mathcal{C}_k + \mu_k \boldsymbol{I}\big), & \text{if } \nabla_{\boldsymbol{q}} \mathcal{C}_k \approx 0 \\ \boldsymbol{D}, & \text{otherwise} \end{cases} \tag{40}$$

where $\boldsymbol{D} \succ 0$ is diagonal, $\alpha \geq 0$ is a design scalar and $\mu_k \geq 0$ is chosen opportunely to ensure $\mathbf{W}_k \succ 0$.

Theorem 4. (Existence of $\mathbf{J}^\star$ and C^1-continuity of $\mathbf{W}_k$). Choosing:

$$\mu_k = \log(1 + \exp(\epsilon - \lambda_{\min})), \tag{41}$$

for (40), where $\epsilon > 0, \epsilon \approx 0$, and $\lambda_{\min}$ is the minimum eigenvalue of $\nabla_{\boldsymbol{q}}^2 \mathcal{C}_k$; guarantees existence and uniqueness of $\mathbf{J}^\star$, and C^1-continuity of $\mathbf{W}_k$.

Proof. The Hessian $\nabla_{\boldsymbol{q}}^2 \mathcal{C}$ is symmetric by Schwarz's theorem. Adding $\max(0, \epsilon - \lambda_{\min})\boldsymbol{I}$ to $\nabla_{\boldsymbol{q}}^2$ shifts all eigenvalues to be positive (Liao and Shoemaker, 1991), ensuring $\mathbf{W}_k \succ 0$, which guarantees the existence and uniqueness of $\mathbf{J}^\star$. Subjecting the above max operation to a standard softplus function provides (41). Being the eigenvalues of $\mathbf{W}_k$ continuous in $\boldsymbol{q}$, $\mathbf{W}_k$ is C^1-continuous in $\boldsymbol{q}$ by construction of (40). □

Choosing a large α makes the term $(\nabla_{\boldsymbol{q}}^2 \mathcal{C}_k + \mu_k \boldsymbol{I})$ dominate, while increasing diagonal entries of $\boldsymbol{D}$ penalizes per-joint motion. To enforce the physical limits $[\boldsymbol{u}_{\min}, \boldsymbol{u}_{\max}]$ through the saturation operator $\mathrm{sat}(\cdot)$ (2a) without distorting the commanded motion direction, $\boldsymbol{u}_k$ is computed from $\boldsymbol{u}_k^{\mathrm{nom}}$ (9) via hierarchical magnitude-based scaling. The primary task is prioritized by reserving a fixed quota (e.g., 30%) of the actuation capability for the secondary task $\boldsymbol{u}_k^{\mathrm{sec}}$.

4. OPTIMAL STEALTH ATTACK

Under the assumptions presented in Section 2.4, we formalize the attacker's strategy to control the end-effector while remaining stealth against threshold-based ADS detection. We envisage an attacker that comprehensively simulates the dynamics of all involved system in the absence of noise, like the plant, controller and proposed active defense system(s).

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

Let $\mathcal{Z}_{k,j}(\boldsymbol{a})$ denote the *j-step-ahead prediction* of the closed-loop system, starting from time k, under the attack sequence $[\boldsymbol{a}, \mathbf{0}, \ldots, \mathbf{0}]$ (the only non-zero injection occurs at time k). For a generic vector $\boldsymbol{v}$, we define

$$\boldsymbol{v}_{k+j}^{\mathrm{SIM}} := \pi_{\boldsymbol{v}}\big(\mathcal{Z}_{k,j}(\boldsymbol{a})\big), \tag{42}$$

where $\pi_{\boldsymbol{v}}(\cdot)$ extracts the vector $\boldsymbol{v}$ at time $k + j$ from the simulated evolution.

Under the closed-loop dynamics in (2a), (2b), (3a), and (3b), an injected signal at time k, $\boldsymbol{a}_k$, influences the end-effector acceleration *two* time steps later. Specifically,

$$\ddot{\boldsymbol{p}}_{k+2}^{\mathrm{SIM}} = \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_k)\big). \tag{43}$$

At each time step, the attacker's high-level objective is to make this predicted acceleration match a desired target acceleration $\ddot{\bar{\boldsymbol{p}}}_{k+2}^{\mathrm{A}}$. This is realized by minimizing

$$\tfrac{1}{2}\left\|\ddot{\bar{\boldsymbol{p}}}_{k+2}^{\mathrm{A}} - \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_k)\big)\right\|^2. \tag{44}$$

4.1 Incremental Attack Formulation

Due to feedback linearization, the manipulator's joint-space dynamics reduce to double integrators, which entails an *integrator vulnerability*. This vulnerability has been analyzed for sensor *Bias Injection Attacks* (BIAs), i.e., *constant* sensor injections, in linear systems (Tosun et al., 2025). Although the joint-task mapping is nonlinear, and we consider the more general *False Data Injection Attacks* (FDIAs) we argue that effective FDIAs still exploit the integrator vulnerability *locally*. For this reason, we model the attack *incrementally* as:

$$\boldsymbol{a}_k = \begin{bmatrix}(\boldsymbol{a}_k)_q \\ (\boldsymbol{a}_k)_{\dot{q}}\end{bmatrix} = \begin{bmatrix}(\boldsymbol{a}_{k-1})_q + \boldsymbol{\Delta}_k \\ \dfrac{\boldsymbol{\Delta}_k}{T_s}\end{bmatrix} = \begin{bmatrix}(\boldsymbol{a}_{k-1})_q \\ \boldsymbol{0}\end{bmatrix} + \boldsymbol{M}\boldsymbol{\Delta}_k, \tag{45}$$

where $(\cdot)_q$, $(\cdot)_{\dot{q}}$ extract respectively the position and velocity components of the attack vector, $\boldsymbol{\Delta}_k \in \mathbb{R}^n$ is an increment, and $\boldsymbol{M} = [\boldsymbol{I}_n \ \frac{1}{T_s}\boldsymbol{I}_n]^\top \in \mathbb{R}^{2n\times n}$ maps the joint increment to the full sensor space. This parameterizes an FDIA as deviations from a baseline BIA and it provides a convenient model for gradient-based synthesis.

A first-order expansion of $\pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\cdot)\big)$ around $\boldsymbol{a}_{k-1}$, gives:

$$\ddot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+2} \approx \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_{k-1})\big) + \boldsymbol{G}_k\boldsymbol{\Delta}_k, \tag{46}$$

with $\ddot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+2}$ as in (43) and the Jacobian matrix

$$\boldsymbol{G}_k = \boldsymbol{Z}_k\boldsymbol{M} \in \mathbb{R}^{3\times n}, \ \text{with} \tag{47}$$

$$\boldsymbol{Z}_k = \frac{\partial}{\partial \boldsymbol{a}}\,\pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a})\big)\Big|_{\boldsymbol{a}=\boldsymbol{a}_{k-1}} \in \mathbb{R}^{3\times 2n} \tag{48}$$

is computed numerically via a central-difference scheme. This $\boldsymbol{G}_k$ provides the local sensitivity needed to differentiate the objective function in (44).

Substituting the linear approximation from (46) into the objective (44), through the relation in (43), and introducing a regularization term, the quadratic cost function is obtained. Specifically, the attack increment $\boldsymbol{\Delta}_k$ minimizes

$$\tfrac{1}{2}\big\|\boldsymbol{G}_k\boldsymbol{\Delta}_k - \ddot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_{k+2} + \pi_{\ddot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,2}(\boldsymbol{a}_{k-1})\big)\big\|^2 + \frac{\zeta}{2}\|\boldsymbol{\Delta}_k\|^2, \tag{49}$$

where $\zeta > 0$ penalizes large increments.

The adversary defines $\ddot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_{k+2}$ in (44) via a one-step-ahead PD law to counteract the drift predicted for $k{+}1$ under $\boldsymbol{\Delta}_k{=}0$ (i.e., using $\boldsymbol{a}_{k-1}$) as:

$$\ddot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_{k+2} = \boldsymbol{K}^{\mathrm{A}}_{\mathrm{p}}\big(\bar{\boldsymbol{p}}^{\mathrm{A}}_{k+1} - \boldsymbol{p}^{\mathrm{SIM}}_{k+1}\big) + \boldsymbol{K}^{\mathrm{A}}_{\mathrm{d}}\big(\dot{\bar{\boldsymbol{p}}}^{\mathrm{A}}_{k+1} - \dot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+1}\big), \tag{50}$$

where

$$\boldsymbol{p}^{\mathrm{SIM}}_{k+1} = \pi_{\boldsymbol{p}}\big(\mathcal{Z}_{k,1}(\boldsymbol{a}_{k-1})\big), \tag{51}$$

$$\dot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+1} = \pi_{\dot{\boldsymbol{p}}}\big(\mathcal{Z}_{k,1}(\boldsymbol{a}_{k-1})\big), \tag{52}$$

are the one-step-ahead predicted position and velocity, and $\boldsymbol{K}^{\mathrm{A}}_{\mathrm{p}}$, $\boldsymbol{K}^{\mathrm{A}}_{\mathrm{d}}$ are constant PD gains.

4.2 χ^2 Stealth Constraint

Due to the incremental modeling of the attack, the ADS residual of (10) is modeled as

$$\boldsymbol{r}_k = \boldsymbol{M}\boldsymbol{\Delta}_k + \boldsymbol{c}_k, \tag{53}$$

where $\boldsymbol{M}$ is defined in Section 4.1, and $\boldsymbol{c}_k \in \mathbb{R}^{2n}$ denotes the baseline innovation, defined as

$$\boldsymbol{c}_k = \big(\boldsymbol{y}_k + \boldsymbol{a}_{k-1} - \hat{\boldsymbol{y}}_k\big). \tag{54}$$

The stealth constraint becomes

$$\big(\boldsymbol{M}\boldsymbol{\Delta}_k + \boldsymbol{c}_k\big)^\top\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{M}\boldsymbol{\Delta}_k + \boldsymbol{c}_k\big) \le \tau. \tag{55}$$

4.3 QCQP Formulation

Considering the objective function in (49) and expanding the constraint of (55), the problem reduces to the QCQP

$$\begin{aligned}&\min_{\boldsymbol{\Delta}_k}\ \frac{1}{2}\boldsymbol{\Delta}_k^\top\boldsymbol{H}\boldsymbol{\Delta}_k + \boldsymbol{g}^\top\boldsymbol{\Delta}_k \\ &\text{s.t. } (\boldsymbol{M}\boldsymbol{\Delta}_k)^\top\boldsymbol{O}(\boldsymbol{M}\boldsymbol{\Delta}_k) + \boldsymbol{b}^\top(\boldsymbol{M}\boldsymbol{\Delta}_k) + c \le 0,\end{aligned} \tag{56}$$

Table 1. Scenarios definitions.

	Task		Defence		
Scenario	Nominal	Attacker	χ^2	Damp.	Man. red.
A1	Circle	-	□	□	□
A2	Circle	-	□	■	□
A3	Circle	-	□	■	■
B1	Still	Circle	□	□	□
B2	Still	Circle	■	□	□
B3	Still	Circle	■	■	□
B4	Still	Circle	■	■	■

with parameters

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

Theorem 5. (Convexity of adversary's QCQP). Problem (56) is convex and admits a unique global minimizer $\boldsymbol{\Delta}_k^\star$.

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

The attack evolves incrementally as in (45)

$$\boldsymbol{a}_k = \boldsymbol{a}_{k-1} + \boldsymbol{\Delta}_k^\star, \tag{57}$$

initialized with $\boldsymbol{a}_0 = \boldsymbol{\Delta}_0^\star$. At each step, the adversary run internal simulations to compute $\boldsymbol{G}_k$, $\boldsymbol{p}^{\mathrm{SIM}}_{k+1}$ and $\dot{\boldsymbol{p}}^{\mathrm{SIM}}_{k+1}$, then solves (56) to determine $\boldsymbol{\Delta}_k^\star$, $k \in [0, T-1]$.

5. SIMULATION RESULTS

5.1 Experimental Setup

We define the following tasks:
Still Task: maintain the fixed hand position $\bar{\boldsymbol{p}}^{\mathrm{task}} = [0.65, 0.17, 0.4]^\top$ m and RPY orientation $\theta^{\mathrm{task}} = [3.05, -0.68, 1.26]$ rad.
Circle Task: Starting from the pose of the Still Task, perform a semi-circled trajectory with center in $\boldsymbol{0}$ (robot base) and final point $[-0.23, 0.26, 0.45]^\top$ m (quintic polynomials with zero boundary velocity/acceleration). Orientation is not controlled.

Table 1 defines Scenarios, indicating the tasks (nominal and for the attacker) and what defense system(s) are in place. Scenarios A1–A3 and B1–B4 represents respectively the absence ($\mathcal{H}_0$) and presence of FDIA as in (57).

We introduce the following metrics. Let $\bar{\boldsymbol{p}}_k$ and $\boldsymbol{p}_k \in \mathbb{R}^3$ denote the nominal and actual hand trajectories. Then:

$$\mathrm{RMS}^{T-1}_{k=0}\big(\bar{\boldsymbol{p}}_k, \boldsymbol{p}_k\big) \coloneqq \left(\frac{1}{T}\sum_{k=0}^{T-1}\big\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\big\|_2^2\right)^{1/2},$$

$$\mathrm{MND}^{T-1}_{k=0}\big(\bar{\boldsymbol{p}}_k, \boldsymbol{p}_k\big) \coloneqq \max_{0\le k\le T-1}\big\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\big\|_2.$$

The total absolute power of the virtual masses system is:

$$|P_k^{\text{tot}}| = \sum_{j=1}^{n} |u_{k,j} \cdot \dot{q}_{k,j}|.$$

The virtual kinetic energy respectively injected and dissipated by the defence damping of (26) is:

$$\mathcal{E}_{\text{inj}} = \sum_{k=0}^{T-1} \max\big(0, \dot{\boldsymbol{q}}_k^\top \boldsymbol{u}_k^{\text{d}}\big) T_s, \quad \mathcal{E}_{\text{diss}} = \sum_{k=0}^{T-1} \big|\min\big(0, \dot{\boldsymbol{q}}_k^\top \boldsymbol{u}_k^{\text{d}}\big)\big| T_s.$$

The plant is a simulated 7-DOF Kinova Gen3 manipulator without joints limit, discretized with sampling time $T_s = 3 \cdot 10^{-3}$ s, and acceleration limits $\boldsymbol{u}_{\max} = -\boldsymbol{u}_{\min} = [1, 1, 1, 1, 10, 10, 10]$. Both joint positions and velocities are sensed, resulting in a measurement vector $\boldsymbol{y}_k \in \mathbb{R}^p$ with $p = 14$, sensing covariance $\boldsymbol{R} = 2.5 \cdot 10^{-9} \boldsymbol{I}_p$, and process noise covariance $\boldsymbol{Q} = \boldsymbol{Q}_{\text{base}} \otimes (q_c \boldsymbol{I}_n)$, where $\boldsymbol{Q}_{\text{base}} = \begin{bmatrix} T_s^3/3 & T_s^2/2 \\ T_s^2/2 & T_s \end{bmatrix}$ is the exact discretization of the Continuous White Noise Acceleration (CWNA) model $q_c = 10^{-5}\,\text{rad}^2/\text{s}^3$ is the process noise intensity, and $\otimes$ is the Kronecker product.

The χ^2 detector of Section 2.3 is calibrated for an ARL of 10^9 samples (833 hours), realized by threshold $\tau = 71.57$ for (11). The proposed adaptive damping defense of Section 3.2 is tuned with: $F = 3$, $\gamma = 6$, $\beta = 0.01$, $\psi = (1 - 0.01)$, determining $z_x = 29.14$. For Theorem 3, we have formal guarantees that, on average and under $\mathcal{H}_0$, a single sample of the control command deviates from 1% of the nominal power every 100 samples. For the proposed manipulability reduction defense of Section 3.3 we set: $\alpha = 5 \cdot 10^3$, $\boldsymbol{D} = \text{diag}(10, 10, 10, 10, 1, 1, 1)$, and $\nu_{\max} = 50$. All tasks has a duration of $T = 5000$ samples (15 s).

5.2 Comparative Analysis and Discussion

The proposed defense adds negligible computational overhead. The QCQP for the attacker (external to the defense) is solved in approximately 40 ms/step using a standard solver (Gurobi). The key performance metrics for all evaluated scenarios are summarized in Table 2. For scenarios A1–A3, the reported values are averaged over 100 Monte Carlo simulations. Figure 1 illustrates the time-series signals for a representative single run of each scenario.

Scenario A1 provides the reference nominal performance for the execution of the circle task under $\mathcal{H}_0$. The proposed active defenses (A2, A3) introduce a negligible disturbance to the primary task, with hand displacement inaccuracy increasing by only $\approx 10^{-5}$ m. Compared to A1, the null-space motion required in A3 increases the mean and maximum power consumption ($|P_k^{\text{tot}}|$) by $+\,950.3\%$ and $+\,2089.3\%$, respectively.

Scenario B1 provides the reference performance for the circle task compromised by FDIA to the sensors. The attack can steer the end-effector with relatively high accuracy. The presence of the χ^2 detector alone (B2) proves ineffective, as z_k remains well below the threshold τ. With the introduction of anomaly-aware virtual damping (B3, B4), the attacker faces a fundamental trade-off: increasing the injection raises the dissipation rate, while reducing the injection makes the stealthy attack less effective. Simultaneously, increasing the injection is bounded by the activation of the stealthiness constraint in (55) enforcing $z_k < \tau$. Overall, the defense significantly reduces the kinetic energy of the double mass integrator, and the attack causes $g(\tilde{z}_k)$ to progressively approach 1, representing the limit above which the system becomes *strictly passive*. With the introduction of manipulability reduction (B4), a larger anomaly at the joint level is required to produce the same desired hand motion; this anticipates the activation of the stealthiness constraint and further decreases the impact of the attack. Notably, all attack impact metrics are reduced, e.g., $\text{RMS}_{k=0}^{T-1}(\bar{\boldsymbol{p}}_k, \boldsymbol{p}_k)$ by $-\,28.1\%$, $\text{RMS}_{k=0}^{T-1}(\bar{\boldsymbol{p}}_k^{\text{A}}, \boldsymbol{p}_k)$ by $+14.2\%$, and the hand's path length by -14.8%. However, the mean and max of $|P_k^{\text{tot}}|$ increase by $+\,5801.6\%$ and $+1967.7\%$, respectively. As $\tilde{z}_k$ increases, the virtual damping active defense (B3, B4) drives the actuators for joints q_2, q_3, q_4 toward their saturation limits, since the induced virtual friction progressively fills the actuator headroom reserved for defense in (25a) and (25b).

Across all experiments, the state deviation $\|\tilde{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_2$ has an average of 0.11 and a maximum of 0.27. It is worth

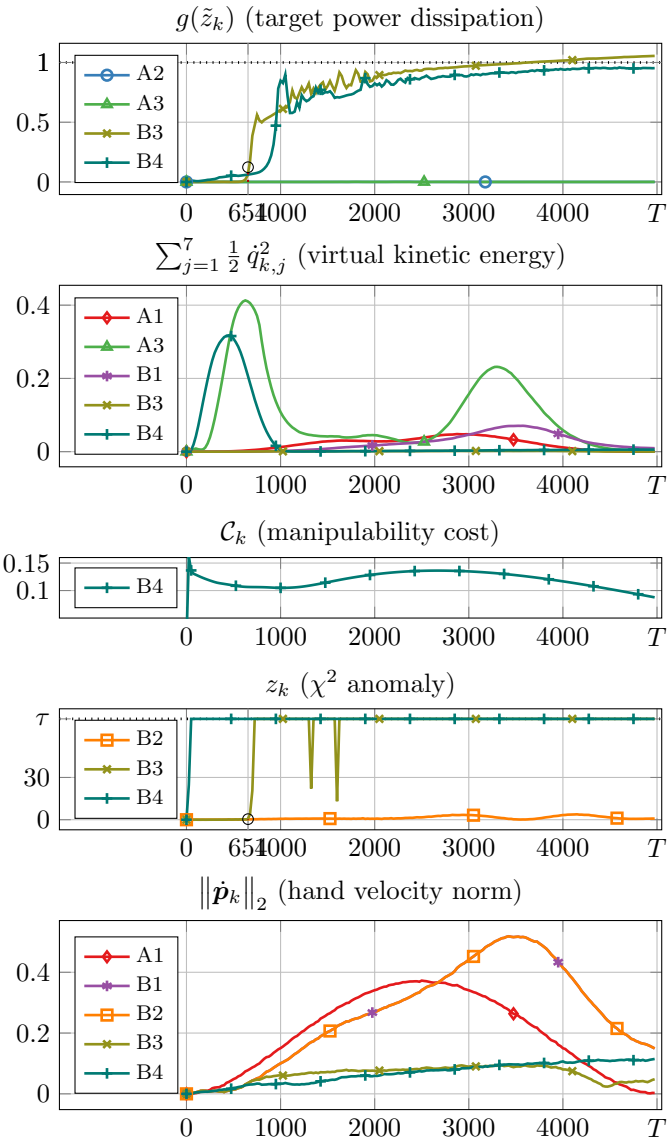

Fig. 1. Signals in the Scenarios as the sampling time k varies in $[0, T-1]$.

Table 2. Comparative metrics results.

ID	Metric	A1	A2	A3	B1	B2	B3	B4
M1.1	$\mathrm{RMS}_{k=0}^{T-1}\left(\bar{\boldsymbol{p}}_k, \boldsymbol{p}_k\right)$, [m]	$1.35 \cdot 10^{-4}$	$1.34 \cdot 10^{-4}$	$1.9 \cdot 10^{-4}$	0.58	0.58	0.27	0.19
M1.2	$\mathrm{MND}_{k=0}^{T-1}\left(\bar{\boldsymbol{p}}_k, \boldsymbol{p}_k\right)$, [m]	$3.19 \cdot 10^{-4}$	$3.19 \cdot 10^{-4}$	$6.06 \cdot 10^{-4}$	0.88	0.88	0.45	0.39
M2.1	$\mathrm{RMS}_{k=0}^{T-1}\left(\bar{\boldsymbol{p}}_k^{\mathrm{A}}, \boldsymbol{p}_k\right)$[m]	∅	∅	∅	0.04	0.04	0.44	0.5
M2.2	$\mathrm{MND}_{k=0}^{T-1}\left(\bar{\boldsymbol{p}}_k^{\mathrm{A}}, \boldsymbol{p}_k\right)$, [m]	∅	∅	∅	0.06	0.06	0.69	0.72
M3.1	mean $\|\dot{\boldsymbol{q}}_k\|_2$	0.17	0.17	0.32	0.19	0.19	0.04	0.16
M3.2	max $\|\dot{\boldsymbol{q}}_k\|_2$	0.31	0.31	0.82	0.38	0.38	0.07	0.8
M3.3	mean $\lvert P_k^{\mathrm{tot}}\rvert$	$9.73 \cdot 10^{-3}$	$9.73 \cdot 10^{-3}$	0.08	0.01	0.01	$8.05 \cdot 10^{-4}$	0.05
M3.4	max $\lvert P_k^{\mathrm{tot}}\rvert$	0.03	0.03	0.58	0.03	0.03	0.02	0.44
M3.5	path length of $\boldsymbol{p}_k$, [m]	1.22	1.22	1.22	1.24	1.24	0.48	0.41
M3.6	defence injected energy $\mathcal{E}_{\mathrm{inj}}$, [J]	∅	$2.9 \cdot 10^{-7}$	$4.84 \cdot 10^{-7}$	∅	∅	0.05	$3.88 \cdot 10^{-4}$
M3.7	defence dissipated energy $\mathcal{E}_{\mathrm{diss}}$, [J]	∅	$7.29 \cdot 10^{-5}$	$5.77 \cdot 10^{-4}$	∅	∅	0.14	0.11

discussing the impact of this deviation under $\mathcal{H}_0$ (A1, A2, A3), i.e., when $\hat{\boldsymbol{x}}_k \approx \boldsymbol{x}_k$. For the manipulability reduction defense, such deviation does not interfere with the primary task, because in (38) we project onto the null-space of $\hat{\mathbf{J}}_k \approx \mathbf{J}_k$. Regarding the damping defense, an excessive deviation could invalidate Assumption 5 and break the stability established in Theorem 3, i.e., if $\mathcal{E}_{\mathrm{inj}} > \mathcal{E}_{\mathrm{diss}}$. In our experiments, $\mathcal{E}_{\mathrm{inj}} \ll \mathcal{E}_{\mathrm{diss}}$; hence, stability is far from being compromised. We acknowledge that this phenomenon establishes an upper bound for the re-synchronization duration interval in (12).

6. CONCLUSIONS

This paper addressed the resilience of robotic manipulators against finite-horizon False Data Injection Attacks (FDIAs) targeting sensors. Feedback linearization induces an *integrator vulnerability* that allows 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 virtual damping* and *manipulability reduction* in the attack direction, that disincentivate anomaly injection as a function of an anomaly score derived from a measurement-free, actuation-projected predictor. We established two key guarantees: (i) probabilistic bounds on power 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, providing a principled benchmark against which to evaluate defenses. Simulation results on a 7-DOF manipulator confirmed that the proposed defenses substantially improves the resilience against FDIA to robotic arms compared to using a χ^2 detection alone.

REFERENCES

De Luca, A. and Oriolo, G. (1991). Issues in acceleration resolution of robot redundancy. In *IFAC Symp. Rob. Contr. (SYROCO)*, 93–98.

Ding, D. et al. (2018). A survey on security control and attack detection for industrial cyber-physical systems. *Neurocomputing*, 275, 1674–1683.

Fawzi, H., Tabuada, P., and Diggavi, S. (2014). Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Trans. Autom. Contr.*, 59(6).

Gualandi, G. and Papadopoulos, A.V. (2026). From passive monitoring to active defense: Resilient control of manipulators under cyberattacks. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Guo, Z., Shi, D., Johansson, K.H., and Shi, L. (2017). Optimal linear cyber-attack on remote state estimation. *IEEE Trans. Control Netw. Syst.*, 4(1), 4–13.

Humayed, A., Lin, J., Li, F., and Luo, B. (2017). Cyber-physical systems security–A Survey. *IEEE Internet Things J.*, 4(6), 1802–1831.

Intriago, A. et al. (2024). Residual-based detection of attacks in cyber-physical inverter-based microgrids. *IEEE Trans. Power Syst.*, 39(2), 4020–4038.

Liao, L.Z. and Shoemaker, C. (1991). Convergence in unconstrained discrete-time differential dynamic programming. *IEEE Trans. Autom. Contr.*, 36(6), 692–706.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. In *Allerton Conf. Comm., Contr. & Comp.*, 911–918.

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

Sandberg, H., Gupta, V., and Johansson, K.H. (2022). Secure networked control systems. *Annual Review of Contr., Rob., & Auton. Syst.*, 5(1), 445–464.

Siciliano, B. and Slotine, J.J. (1991). A general algorithm for managing multiple tasks in highly redundant robotic systems. In *Int. Conf. Adv. Rob. (ICAR)*, 1211–1216.

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

Tosun, F.E., Teixeira, A.M.H., Dong, J., Ahlén, A., and Dey, S. (2025). Kullback-Leibler divergence-based observer design against sensor bias injection attacks in single-output systems. *IEEE Trans. Information Forensics and Security*, 20, 2763–2777.

Tunga, R., Murguia, C., and Ruths, J. (2018). Tuning windowed chi-squared detectors for sensor attacks. In *Am. Contr. Conf. (ACC)*, 1752–1757.

Ueda, J. and Blevins, J. (2024). Affine transformation-based perfectly undetectable false data injection attacks on remote manipulator kinematic control with attack detector. *IEEE Robot. Autom. Lett.*, 9(10), 8690–8697.

Van der Schaft, A. (2000). *L2-gain and passivity techniques in nonlinear control*. Springer.