Optimal Myopic Attacks on Nonlinear Estimation

R. Spencer Hallyburton, Amir Khazraei, and Miroslav Pajic

***Abstract*— Prior works have analyzed the security of estimation and control (E&C) for linear, time-invariant systems; however, there are few analyses of nonlinear systems despite their broad safety-critical use. We define two attack objectives on nonlinear E&C and illustrate that realizing the optimal attacks against the widely-adopted extended Kalman filter with industry-standard χ^2 anomaly detection is equivalent to solving convex quadratically-constrained quadratic programs. Although these require access to the true state of the system, we provide practical relaxations on the optimal attacks to allow for execution at runtime given a specified amount of attacker knowledge. We show that the difference between the optimal and relaxed attacks is bounded by the attacker knowledge.**

I. Introduction

Security analysis of estimation and control (E&C) in cyber-physical systems (CPS) has attracted considerable research interest due to safety-critical CPS applications. Most of the influential work in E&C CPS security has centered on linear, time-invariant (LTI) systems. For instance, [1]–[4] exploited vulnerabilities of LTI E&C with information models ranging from full system access to single-sensor level knowledge to demonstrate concerning vulnerabilities in some of the most widely used E&C algorithms. After the discovery of LTI E&C vulnerabilities, subsequent works proposed algorithms for detecting attacks and architectures for attack-resilient state estimation [5]–[8]. As a response, recent focus has been directed towards undetectable or "stealthy" attacks on LTI CPS (e.g., [9]).

However, insights from analysis of LTI CPS lack practical relevance because controlled physical processes of safety-critical importance are often nonlinear. For example, automotive applications with inertial measurement units (IMUs) are nonlinear in control. Add in relative-range sensor or tightly-coupled global positioning systems (GPS), and the problem is also nonlinear in the measurements. Airborne applications such as drones are similarly often highly nonlinear.

A handful of works have attempted to analyze nonlinear, time-invariant control from a security perspective. For example, [10] investigated nonlinear AC control in power grids and designed false-data injection attacks. However, these often consider highly specialized attack goals, e.g., [10] derived attacks in closed-form with precise dynamical equations; [11] analyzed the extended Kalman filter (EKF) but considered stochastic attacks rather than an optimal attack. These works, while interesting case studies, provide little in advancing a broad understanding of CPS security.

Thus, there is a gap in existing literature. LTI system analyses leverage the simplicity of the dynamics to derive provably optimal attacks and accurate resilient estimators (e.g., [6], [7]). Unfortunately, few of these ideals can be transferred to nonlinear systems. The complexity and suboptimality of nonlinear estimators has correspondingly allowed for few established guarantees in nonlinear theory and applications; hence, recent works mainly focused on the use of deep-learning for effective attack design (yet, without any guarantees) on system with nonlinear dynamics (e.g., [12]).

Consequently, to address this shortcoming, in this work, we establish optimal and stealthy false-data-injection attacks against the widely-used EKF. We select a permissive information model and describe two myopic (one-step) attack objectives. The first is a *myopic maximum deviation* (MMD) attack that maximally deviates the state estimation error in an attacker-defined subspace of the state space. The second is a *myopic adversarial state approach* (MASA) attack that optimally pushes the victim's state towards an adversarial state in an attacker-defined subspace. We show that the designs of both attacks can be captured as convex optimization problems that are solvable in polynomial time.

Several of the derived optimal attacks are practically infeasible because they require more knowledge than the attacker may be able to acquire. In such cases, we pursue practical relaxations of the original objective based on an information model and derive guarantees on the boundedness of the suboptimality for the relaxed case. Finally, we demonstrate the effectiveness of attacks in a case study and find that attacking nonlinear estimation is effective and has robust performance guarantees. With strong guarantees and efficient runtime performance, our proposed attacks establish a new framework for security analysis of nonlinear dynamical systems.

The paper is organized as follows: Section II presents the state estimation models of linear and nonlinear systems. Section III introduces the security model including the attacker's knowledge and goals. Section IV then derives the optimal myopic attacks on nonlinear Kalman filtering and provides guarantees on practical relaxations. Finally, Section V covers case studies and Monte Carlo simulations to evaluate the optimal attacks and derived bounds.

Notation: $\mathbb{N}$ and $\mathbb{R}$ denote the sets of natural and real numbers, respectively. $\mathbb{R}^n_+$ is the non-negative subspace of $\mathbb{R}^n$. Pr denotes the probability for a random variable. $\mathcal{N}(\mu,\Sigma)$ denotes a Gaussian distribution with mean vector μ and covariance matrix Σ. We represent positive-(semi)-definiteness of a matrix M, as $M \succ (\succeq) 0$.

This work is sponsored in part by the ONR under agreement N00014-20-1-2745, AFOSR under award number FA9550-19-1-0169, and by the NSF under CNS-1652544 award and the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant CNS-2112562.

R. S. Hallyburton, A. Khazraei, and M. Pajic are with Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA; miroslav.pajic@duke.edu

II. System Model and Preliminaries

In this section, we formally introduce the model of nonlinear estimation in CPS.

A. State Estimation

We consider a discrete-time nonlinear time-invariant physical process modeled in the standard state-space form as

$$\begin{aligned} x_k &= f(x_{k-1}, u_k) + w_k, \\ z_k &= h(x_k) + v_k; \end{aligned} \tag{1}$$

here, $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, $z_k \in \mathbb{R}^p$ are the state, input and output vectors of the plant at time $k \in \mathbb{N}$; f and h are nonlinear functions capturing state transition and measurement models, respectively. Finally, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^p$ are the process and measurement noises that are assumed to be Gaussian with zero mean and Q and R covariance matrices, respectively.

1) Extended Kalman Filter (EKF): If f and h are nonlinear and at least differentiable to first-order, the EKF is a practical way to estimate states. The EKF uses a propagation step to mix control signal and dynamical equations and an update step to fuse measurements.

Propagation: Linearizing f as $F_k \coloneqq \frac{\partial f}{\partial x}\Big|_{x=\hat{x}_{k-1|k-1}}$, the state is propagated using the control signal,

$$\begin{aligned} \hat{x}_{k|k-1} &= f(\hat{x}_{k-1|k-1}, u_k) \\ P_{k|k-1} &= F_k P_{k-1|k-1} F_k^T + Q_k, \end{aligned} \tag{2}$$

where $P \succ 0$ is the state covariance matrix.

Update: Linearizing h as $H_k \coloneqq \frac{\partial h}{\partial x}\Big|_{\hat{x}_{k|k-1}}$, the state is updated with the innovation, $\tilde{y}_k$ (i.e., the residual),

$$\begin{aligned} \tilde{y}_k &= z_k - h(\hat{x}_{k|k-1}); \quad S_k = H_k P_{k|k-1} H_k^T + R_k, \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \tilde{y}_k; \quad K_k = P_{k|k-1} H_k^T S_k^{-1}, \end{aligned} \tag{3}$$

with S_k the innovation covariance and K_k the Kalman gain.

2) Anomaly Detection: If the system is truly *linear* with Gaussian noise, the innovations are white (i.e., $\tilde{y}_k \sim \mathcal{N}(0, S_k)$) and the scalar $g_k^{\chi^2} \coloneqq \tilde{y}_k^T S_k^{-1} \tilde{y}_k$ follows a χ^2 distribution with p degrees of freedom. This leads to a statistical anomaly detection function for incoming measurements:

$$\begin{aligned} &\text{reject measurement if: } g_k^{\chi^2} > \tau \\ &\text{where } \beta = \Pr(V \leq \tau), \quad g_k^{\chi^2} \coloneqq \tilde{y}_k S_k^{-1} \tilde{y}_k; \end{aligned} \tag{4}$$

i.e., a measurement is rejected if $g_k^{\chi^2}$ exceeds a threshold τ. That threshold is set such that, for a perfect χ^2 random variable V, the smallest β (e.g., β= 99% [13]) are accepted.

The χ^2 anomaly detector is still used in non-linear systems in practice using the linearizations and assuming the dynamical models capture the behavior of the plant.

III. State Estimation Security Model

We make two assumptions on the attacker. First, the attacker has access to a "full-reactive" suite of knowledge, defined in Section III-A. Second, the attack goal is *myopic*. A fully general attack could trade short-term loss for long-term gains. However, as described in Section III-B, this can be challenging to formalize and compute in real-time.

A. Threat Model

1) Knowledge: We consider four elements of knowledge important for CPS controllers. Namely, these are:

- **System Goal State** – Knowledge of the intended future state of the system;
- **Control Signals** – Access to the control signals, u_k, and the state propagation in (2);
- **Measurement Models** – Access to the state update of (3) including the measurement model and measurement noise;
- **Sensor Data** – Access to sensor data from one or more sensors in real-time.

In this work, we analyze cases where the attacker has near-complete knowledge. Specifically, we consider a "full reactive" set of knowledge where the attacker has all knowledge except the system's goal state.

2) Capability: We assume the attacker can only modify existing sensor data and cannot send additional sensor data nor modify the measurement timestamp, consistent with e.g., [3], [6], [14]. We also assume the attacker cannot reliably compromise control signals. In general, such attacks can be modeled as an adversarial bias – i.e., $z_k^a \coloneqq z_k + a_k$.

Thus, the state update of (3) with anomaly detection of (4) under such an attack can be captured as

$$\hat{x}_{k|k}(a_k) = \begin{cases} \hat{x}_{k|k-1} & \text{if } \tilde{y}_k^T(a_k) S_k^{-1} \tilde{y}_k(a_k) > \tau \\ \hat{x}_{k|k-1} + K_k \tilde{y}_k(a_k) & \text{otherwise} \end{cases} \tag{5}$$

with $\tilde{y}_k(a_k) \coloneqq z_k + a_k - h(\hat{x}_{k|k-1})$; i.e., *any measurement triggering the detector is not included in the estimation update*.

B. Attack Goal

A fully general attack could trade short-term loss for long-term gain. However, it is challenging to formalize an attacker planning for short and long term horizons when the attack goal may be unbounded in state space (e.g., maximum deviation). Thus, we formalize attacks as myopic (one-step) optimization problems. We define two classes of attacker goal for E&C: the myopic maximum deviation (MMD) and the myopic adversarial state approach (MASA) attacks.

The MMD attack maximizes the error between the victim's state and the true state of the system.

Definition 1 (Myopic Maximum Deviation (MMD)): An attack $a_k^* \in \mathbb{R}^p$ is a myopic maximum deviation attack if

$$a_k^* = \arg\max_a \ \frac{1}{2} \left\| C(x_k - \hat{x}_{k|k}(a_k)) \right\|^2, \tag{6}$$

where $C \in \mathbb{R}_+^{w \times n}$, $w \leq n$, is an attacker-specified projection (e.g., weight) matrix.

The MASA attack optimally moves towards an adversary-defined state at each step, implemented with two subvariants.

Definition 2 (Myopic Adversarial-State Approach (MASA)): Let placeholder $\nu_k \in \mathbb{R}^n$ be a function of attack $a_k \in \mathbb{R}^p$. Let $C \in \mathbb{R}_+^{w \times n}$ be an attacker-specified projection matrix, $w \leq n$. Let $\mathscr{X}_k^a \in \mathbb{R}^w$ be an attacker-specified state. Then, ν_k approaches $\mathscr{X}_k^a$ under the attack if

$$\|C\nu_k(a_k) - \mathscr{X}_k^a\| < \|C\nu_k(0) - \mathscr{X}_k^a\|.$$

Such an attack is myopic optimal if

$$a_k = a_k^* = \arg\min_a \frac{1}{2} \ \|C\nu_k(a) - \mathscr{X}_k^a\|^2. \tag{7}$$

Definition 2.1: An *estimated-state MASA attack* is a MASA attack with $\nu_k := \hat{x}_{k|k}$. In addition, a *true-state MASA attack* is a MASA attack with $\nu_k := x_k$.

In the remainder of this work, we derive optimal, polynomial-time realizations of MMD, estimated-state MASA, and true-state MASA attacks. We also provide practical relaxations to for a "full reactive" knowledge model.

IV. Optimal Attacks

We derive polynomial-time optimal attacks for MMD and MASA objectives. Under the full-reactive knowledge, the MMD optimization is infeasible due to the required knowledge of the true state. Thus, we propose a feasible plant-state relaxation to the MMD attack. We find the estimated-state MASA is feasible while the true-state MASA is infeasible and requires relaxation. However, we do not show guarantees on the relaxed true-state MASA attack.

a) Additional notation: To simplify our notation, we use $\hat{x} := \hat{x}_{k|k}$, $\hat{x}^- := \hat{x}_{k|k-1}$, and $\hat{h}^- := h(\hat{x}_{k|k-1})$. Since the attacks are myopic, we safely drop time (k) subscripts for any E&C element. Below, we define the substitutions used to transform nonlinear attack objectives into quadratically-constrained quadratic program (QCQPs), as in Propositions 1, 4, and 5, and introduce subscripts only to differentiate between the objective (A_0, b_0) and constraints (A_1, b_1, d_1). We also define the following terms (the "Substitutions"):

$$\begin{aligned}
A_0 &:= (CK_k)^T CK_k \geq 0 \\
b_0 &:= -(CK_k)^T C\left(x_k - (\hat{x}_{k|k-1} + K_k(z_k - h(\hat{x}_{k|k-1}))\right) \\
\breve{b}^0 &:= -(CK_k)^T C\left(\breve{x}_k - (\hat{x}_{k|k-1} + K_k(z_k - h(\hat{x}_{k|k-1}))\right) \\
\bar{b}^0 &:= (CK_k)^T C(\hat{x}_{k|k-1} + K_k(z_k - h(\hat{x}_{k|k-1}))) - (CK_k)^T \mathscr{X}^a \\
\breve{\bar{b}}^0 &:= (CK_k)^T C(\hat{x}_{k|k-1} + K_k(z_k - h(\hat{x}_{k|k-1}))) - (CK_k)^T \mathscr{X}^b \\
\mathscr{X}^b &:= C\hat{x}_{k|k-1} + C\breve{x}_{k|k} - \mathscr{X}^a \\
A_1 &:= 2S_k^{-1} > 0 \\
b_1 &:= 2S_k^{-1}(z_k - h(\hat{x}_{k|k-1})) \\
d_1 &:= (z_k - h(\hat{x}_{k|k-1}))^T S_k^{-1}(z_k - h(\hat{x}_{k|k-1})) - \tau.
\end{aligned}$$

$$\begin{aligned}
\text{Objectives: } J(a) &:= {}^1\!/_2 \ a^T A_0 a + b_0^T a \\
\breve{J}(a) &:= {}^1\!/_2 \ a^T A_0 a + \breve{b}_0^T a \\
\bar{J}(a) &:= {}^1\!/_2 \ a^T A_0 a + \bar{b}_0^T a \\
\breve{\bar{J}}(a) &:= {}^1\!/_2 \ a^T A_0 a + \breve{\bar{b}}_0^T a \\
\text{Constraint: } G(a) &:= {}^1\!/_2 \ a^T A_1 a + b_1^T a + d_1 \leq 0.
\end{aligned}$$

A. Design of MMD Attacks

We now consider how to implement optimal and practical MMD attacks introduced in Definition 1.

Proposition 1: The MMD attack (from Definition 1) can be obtained as the solution of the optimization problem

$$\begin{aligned}
a_{\text{mmd}}^* = \arg\max_a J(a) = \arg\max_a \ & \frac{1}{2} a^T A_0 a + b_0^T a, \\
\text{subject to } G(a) = & \frac{1}{2} a^T A_1 a + b_1^T a + d_1 \leq 0.
\end{aligned} \tag{8}$$

Intuitively, Proposition 1 states that the most effective attack is *stealthy* for the employed attack detector (i.e., does not trigger the anomaly detector (4)) because, due to (5), sensor measurements that trigger the detector are rejected.

Proof: We begin with (6) and perform transformations that do not change the optimization. We consider $\hat{x}(a)$ according to (5) which is piecewise with cases as follows.

Case (1): when $\tilde{y}^T(a) S^{-1} \tilde{y}(a) \leq \tau$. Then, from (6), using $l := x - \hat{x}^- - Kz + K\hat{h}^-$, it holds that

$$\begin{aligned}
a^* &= \arg\max \left\| C(x - (\hat{x}^- + K\tilde{y}(a))) \right\|^2 \\
&= \arg\max \ (l - Ka)^T C^T C (l - Ka) \\
&= \arg\max \ a^T K^T C^T CKa - 2lC^T CKa + l^T C^T Cl \\
&= \arg\max \ \frac{1}{2} a^T K^T C^T CKa - lC^T CKa \\
&= \arg\max \ J(a),
\end{aligned}$$

Case (2): when $\tilde{y}^T(a) S^{-1} \tilde{y}(a) > \tau > 0$. Then, from (6),

$$\begin{aligned}
a^* &= \arg\max \ \left\| C(x - \hat{x}^-) \right\|^2 \\
&= \arg\max \ \left\| C(x - \hat{x}(a - \hat{h}^- - z)) \right\|^2.
\end{aligned}$$

Thus, any attack causing the χ^2-detector to exceed the threshold τ has the same effect on $\hat{x}$ as the stealthy attack $a^0 = \hat{h}^- - z$. Therefore, it is sufficient to consider only stealthy attacks $\{a \mid g_k^{\chi^2}(a) \leq \tau\}$, which is equivalent to imposing the constraint $G(a) \leq 0$. ■

The MMD attack is thus a QCQP with a single constraint, which is solvable in polynomial time regardless of the convexity of the objective and constraint functions [15]. Nevertheless, the MMD QCQP is convex ($A_0 \succeq 0$, $A_1 \succ 0$).

Proposition 2: $G(a^*) = 0$ for a^* the optimal MMD attack. Equivalently, $y(a^*)^T S^{-1} y(a^*) = \tau$.

Proof: The MMD objective is a convex maximization problem. The global maximum of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is attained at an extreme feasible point over the domain of f. With continuous convex constraints, this point satisfies the constraint with equality (see e.g., [16] Theorem I.1). ■

1) Practical Relaxations: It is not possible to know b_0 due to the dependence on the true plant state x. We therefore propose a plant-state relaxation of the MMD attack using an attacker's uncompromised estimate of the plant, $\breve{x} := \breve{x}_{k|k}$.

Definition 3 (Plant-State MMD Attack): An attack $a^\dagger \in \mathbb{R}^p$ is a plant-state MMD attack if

$$a^\dagger = \arg\max_a \ \frac{1}{2} \left\| C(\breve{x} - \hat{x}(a)) \right\|^2, \tag{9}$$

where $\breve{x}$ is the attacker's *uncompromised* estimate of x.

Proposition 3: The plant-state MMD attack is the solution

$$\begin{aligned}
a^\dagger = \arg\max_a \breve{f}(a) = \arg\max_a \ & \frac{1}{2} a^T A_0 a + \breve{b}_0^T a, \\
\text{subject to } G(a) = & \frac{1}{2} a^T A_1 a + b_1^T a + d_1 \leq 0.
\end{aligned} \tag{10}$$

Proof: Follows Proposition 1, replacing x with $\breve{x}$. ■

The plant-state MMD attack is *feasible* at runtime under full-reactive knowledge with the attacker's estimate of the

true state. However, the optimal plant-state MMD attack from (10) will be suboptimal on the MMD objective from (6) compared to the optimal MMD attack from (8). We therefore seek to bound the performance loss in the following result.

Theorem 1 (Optimal Attack Error Absolutely Bounded): The error between the plant-state and true-state MMD attacks is bounded by

$$\left\|a^* - a^\dagger\right\| \leq 2\sqrt{\tau\lambda_{\max}(S)}, \tag{11}$$

where $\lambda_{\max}(S)$ is the largest eigenvalue of the innovation covariance matrix S and τ is the χ^2 threshold.

Proof: Let a^*, $a^\dagger$ be solutions to the true-state and plant-state MMD problems (i.e., (8), (10)). From Prop. 1, 3, all choices a satisfies $y(a)^T S^{-1} y(a) \leq \tau$. Since $\|y(a)\|^2 \lambda_{\min}(S^{-1}) \leq y(a)^T S^{-1} y(a)$, $\|y(a)\|^2 \leq \frac{\tau}{\lambda_{\min}(S^{-1})} = \tau\lambda_{\max}(S)$. Finally, $\left\|y(a^*) - y(a^\dagger)\right\| = \left\|a^* - a^\dagger\right\| \leq \|y(a^*)\| + \left\|y(a^\dagger)\right\| \leq 2\sqrt{\tau\lambda_{\max(S)}}$, completing the proof. ∎

With a full-reactive knowledge model, this result provides a bound on the error between the optimal attack of the feasible plant-state MMD problem compared to the optimal attack of the infeasible MMD problem.

Estimators tend not to be provably optimal for non-linear systems except in special cases. However, methods such as the EKF have shown consistent performance in practice. Often, with Monte Carlo simulation or trials on real data, a bound on the estimation error can be experimentally determined. We use the idea that the estimation error may be unknown but bounded to pursue guarantees on the attack performance in terms of the objective function, J.

Definition 4 (Subspace-Bounded): An estimate of some state w_{est} is subspace bounded from the true value w_{true} by δ if $\|C(w_{\text{est}} - w_{\text{true}})\| \leq \delta$, for a predefined projecting (e.g., weight) matrix C.

Specifically, we continue with the idea that the error of the attacker's estimate of the plant state is unknown but subspace bounded by satisfying $\|C(x - \check{x})\| \leq \delta$ at each timestep.

Lemma 1: If the error of the attacker's estimate of the plant state is subspace bounded by δ, then the error between the true and observable QCQP linear coefficients, b_0 and $\check{b}_0$, in the objective function at each timestep is bounded by

$$\left\|b_0 - \check{b}_0\right\| \leq \delta\sigma_{\max}(CK), \tag{12}$$

where $\sigma_{\max}(CK)$ is the largest singular value of CK, C is an attacker-defined weight matrix, and K the Kalman gain.

Proof: Let us define $w \coloneqq \hat{x}^- + K(z - \hat{h}^-)$. Then,

$$\begin{aligned}\left\|b_0 - \check{b}\right\| &= \left\|-(CK)^T C(x-w) + (CK)^T C(\check{x}-w)\right\| \\ &= \left\|K^T C^T C(\check{x}-x)\right\| \\ &\leq \|CK\|\,\|C(\check{x}-x)\| \leq \delta\,\|CK\| = \delta\sigma_{\max}(CK),\end{aligned}$$

completing the proof. ∎

Lemma 2: If the error of the attacker's state estimate is subspace bounded by δ, then for any a_1, $a_2 \in \mathbb{R}^p$ such that $J(a_1) \geq J(a_2)$, the difference in the objectives is bounded by

$$\begin{aligned}&0 \leq J(a_1) - J(a_2) \\ &\leq \frac{1}{2}\varepsilon^2\lambda_{\max}(A_0) + \varepsilon\left(\|A_0 a_2\| + \delta\sigma_{\max}(CK) + \left\|\check{b}_0\right\|\right),\end{aligned} \tag{13}$$

where $\varepsilon \coloneqq \|a_1 - a_2\|$, and $\lambda_{\max}(A_0)$ is the largest eigenvalue of A_0, while $\sigma_{\max}(CK)$ is the largest singular value of CK, A_0 is defined by the Substitutions, C is the attacker-defined weight matrix, and K is the Kalman gain.

Proof: Using $e \coloneqq a_1 - a_2$, $A_0 = A_0^T$, it follows that

$$\begin{aligned}0 \leq J(a_1) - J(a_2) &= {}^1\!/_2\; a_1^T A_0 a_1 + b_0^T a_1 - {}^1\!/_2\; a_2^T A_0 a_2 - b_0^T a_2 \\ &= {}^1\!/_2\; e^T A_0(e + 2a_2) + b_0^T e \\ &\leq {}^1\!/_2\; \|e\|^2 \|A_0\| + \|e\|\,\|A_0 a_2 + b_0\| \\ &= {}^1\!/_2\; \varepsilon^2\lambda_{\max}(A_0) + \varepsilon\,\|A_0 a_2 + b_0\|.\end{aligned}$$

In addition,

$$\begin{aligned}\|A_0 a_2 + b_0\| &\leq \|A_0 a_2\| + \left\|(b_0 - \check{b}_0) + \check{b}_0\right\| \\ &\leq \|A_0 a_2\| + \delta\sigma_{\max}(CK) + \left\|\check{b}_0\right\|\end{aligned}$$

from Lemma 1, thus completing the proof. ∎

Theorem 2 (Suboptimality in Plant-State MMD): If the error of the attacker's state estimate is subspace bounded by δ, then the difference between the MMD objective evaluated on the solutions of (8) and (10) (a^* and $a^\dagger$), is bounded by

$$\begin{aligned}0 \leq J(a^*) - J(a^\dagger) &\leq 2\tau\lambda_{\max}(S)\lambda_{\max}(A_0) + \\ &+ 2\sqrt{\tau\lambda_{\max}(S)}\left(\|A_0 a_2\| + \delta\sigma_{\max}(CK) + \left\|\check{b}_0\right\|\right).\end{aligned} \tag{14}$$

Proof: The proof follows by from Lemma 2 and using $\varepsilon \coloneqq \left\|a^* - a^\dagger\right\| \leq 2\sqrt{\tau\lambda_{\max}(S)}$ from Theorem 1. ∎

All quantities on the right-hand-side of Theorem 2 are available at runtime under full-reactive knowledge without access to the true state of the plant. This bound can thus be computed online and dictates how far the attacker can be from the optimal attack impact.

Finally, we bound the difference between the attacker's perceived impact and the true impact of an attack.

Theorem 3 (Perceived vs. True Impact): If the error of the attacker's estimate of the plant state is subspace bounded by δ, then the difference between the true impact and the perceived impact of an attack a is bounded by

$$|J(a) - \check{J}(a)| \leq \delta\sigma_{\max}(CK)\,\|a\|; \tag{15}$$

here, $\sigma_{\max}(CK)$ is the largest singular value of CK, C is a weight matrix, and K the Kalman gain.

Proof: The result directly holds since from Lemma 1, $|J(a) - \check{J}(a)| = |(b_0 - \check{b}_0)^T a| \leq \left\|b_0 - \check{b}_0\right\|\,\|a\| \leq$
$\leq \delta\sigma_{\max}(CK)\,\|a\|$. ∎

B. Design of MASA Attacks

We now employ the same procedure to design MASA attacks and bound online attack performance. We start with the following result for fully optimal MASA attack design.

Proposition 4: The estimated-state MASA attack (Def. 2.1) is the solution of the optimization problem

$$\begin{aligned}a^*_{\text{masa, es}} = \arg\min_a\ &\frac{1}{2}a^T A_0 a + \bar{b}_0^T a \\ \text{subject to } &\frac{1}{2}a^T A_1 a + b_1^T a + d_1 \leq 0.\end{aligned} \tag{16}$$

Proof: Follows directly from Proposition 1 by replacing Cx with $\mathscr{X}^a$ where $\mathscr{X}^a$ is the attacker specified state. ∎

Note that the estimated-state MASA attack does not require knowledge of the true state of the plant and is thus feasible (i.e., can be executed online).

1) Relaxations: The true-state MASA attack from Definition 2.1 requires the true plant state which is unavailable to the attacker. Furthermore, the relaxation using $\check{x}$ instead of the true state is not sufficient in the MASA attack due to the delayed dependence of $\check{x}$ on a – i.e., the attack impacts $\check{x}$ after control corrects for errors in $\hat{x}$ (which *does* depend on a through the compromised update in (5)). Such delayed dependence can be highly non-linear and depends on the victim's goal and controller which are not fully available under a full-reactive knowledge model.

Therefore, we propose an alternative relaxation using a reflection of the attacker's estimated state of the plant.

Definition 5 (Reflected True-State MASA): Let $\mathscr{X}^a \in \mathbb{R}^w$ be an attacker-specified state for true-state MASA (Def. 2.1). Let C be an attacker-specified projection matrix. An attack $a^\dagger$ is a reflected true-state MASA attack if obtained as

$$\begin{aligned} a^\dagger_{\text{masa, ts}} = \arg\min_a \ & \|C\hat{x}(a) - \mathscr{X}^b\|_2^2 \\ \text{s.t.} \quad & \tilde{y}(a) S^{-1} \tilde{y}^T(a) \leq \tau \\ \mathscr{X}^b := & C\hat{x}^- + C(\check{x} - \mathscr{X}^a). \end{aligned} \tag{17}$$

Now, we can capture the following result.

Proposition 5: The reflected true-state MASA attack can be obtained as a solution to the following problem

$$\begin{aligned} a^\dagger_{\text{masa, ts}} = \arg\min_a \ & \frac{1}{2} a^T A_0 a + \breve{b}_0^T a \\ \text{subject to } & \frac{1}{2} a^T A_1 a + b_1^T a + d_1 \leq 0. \end{aligned} \tag{18}$$

Proof: Follows from Proposition 1, by replacing Cx with $\mathscr{X}^b$ "where $\mathscr{X}^b$ is the reflection of the attacker's goal state across the current state. ■

The reflected true-state MASA attack will myopically push $\hat{x}$ in the *opposite* direction of the *attacker's* goal state, $\mathscr{X}^a$, which intuitively will cause the control to compensate *towards* the adversary's goal state. However, without access to the control module, i.e., without knowing how the control will react to the state estimate error, this has few guarantees and the worst-case error may be difficult if not impossible to bound. That said, we find that it works well in practice.

V. Evaluation

We demonstrate impact of the myopic attacks on nonlinear state estimation in a kinematic case study. Subsequently, the bounds from Theorems 1, 2, and 3, and Lemma 1 are validated using Monte Carlo (MC) simulations.

A. Case Studies

We use a nonlinear kinematic state estimation application. Still, the presented principles and experiments generalize to all applications of linear and extended Kalman filtering.

1) Model: We simulate a dynamic target tracking scenario by modeling spherical coordinate returns from a radar sensor with component-wise Gaussian noise according to [17]. We simulate range, azimuth, and elevation measurements (ρ, θ, ϕ) relative to a fixed sensor platform and use an EKF

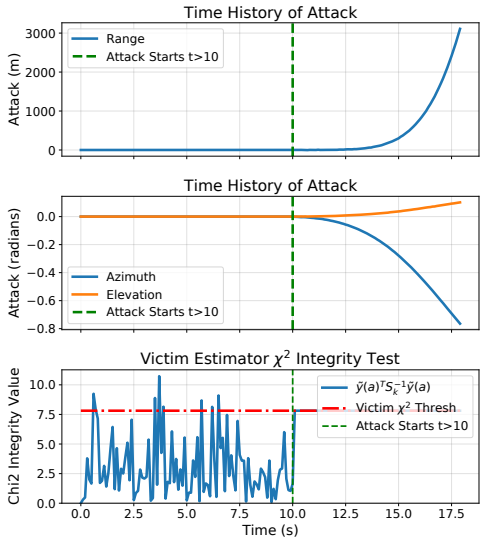

Fig. 1: Nonlinear estimation is easily compromised by myopic MMD attacks. Attack quickly and stealthily compromises spherical coordinate measurements. Estimation errors grow beyond the uncertainty bounds suggested by the EKF.

to process measurements. We estimate position, velocity, and acceleration states using a nearly-constant-acceleration model from [17]. We allow the filter to converge over $t = 10.0\ s$ before starting the attacks and running until $t = 18.0\ s$.

2) Methods: Each attack objective is convex with closed-form gradients and Hessians. We pre-condition following [15] using the Cholesky factorization of the inverse constraint Hessian to achieve faster convergence. This step is essential to obtaining real-time convergence, particularly in the case of order-of-magnitude scaling discrepancies between measurements (i.e., $\rho >> \theta,\ \phi$).

We choose a constrained trust-region optimization algorithm and find that the optimization runs faster than the simulation rate, easily keeping up with real-time.

3) Case Study I – MMD: Fig. 1 shows results of the MMD attacks with the projection matrix set as $C_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$, $w = 3$, $n = 9$. The attack quickly compromises the victim's (i.e., plant) state estimate, even with nonlinearities in E&C. Fig. 1 illustrates that the attack never exceeds the threshold set by the χ^2 anomaly detector meaning the attack remains stealthy, entirely in accordance with Proposition 2.

4) Case Study II– MASA: Fig. 2 shows the same model with an estimated-state MASA attack following (7). The attacker drives $\hat{x}$ towards a specified goal state, $\mathscr{X}^a$. In this kinematic application, we find that solely specifying attack goal as a position state (i.e., $C \in \mathbb{R}_+^{3\times 9}$) does succeed in rapidly pushing the state estimate towards the attacker goal, but that overshoot occurs. This is expected since the attack was formulated as a myopic optimization. Thus, we choose an attacker goal state that has both position and velocity. Specifically, $C_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$, $w = 6$, $n = 9$. We observe the objective remains constant at 0 in Fig. 2 without overshoot.

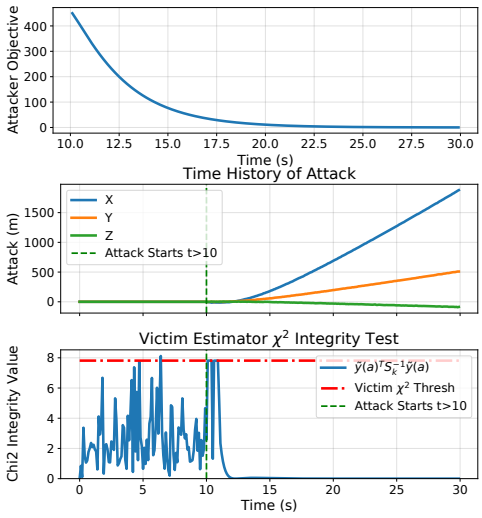

Fig. 2: MASA attack quickly drives the estimated state towards attacker-specified state. The optimization objective drops to 0 once the attacker perfectly reaches the state.

B. MC Bound Simulation

Next, we use Monte Carlo simulations without a dedicated dynamics model to investigate the bounds derived in Sec. IV.

1) Methods: Given a fixed true state $x_k \in \mathbb{R}^9$, $\hat{x}_{k|k-1}$ and $\check{x}_{k|k}$ are sampled from a Gaussian distribution given a fixed victim-state covariance matrix, $P \in \mathbb{R}_+^{9\times 9}$, $P \succ 0$. A measurement model creates a measurement from the true state for the EKF. We choose $C \in \mathbb{R}_+^{3\times 9}$ to maximize the deviation in the first three states.

2) Results: $N = 10000$ Monte Carlo trials are used to observe behavior of the myopic attacks. Fig. 3 shows histograms of quantities derived in Theorems 1, 2, 3 and Lemma 1. All bounds are order-of-magnitude tight.

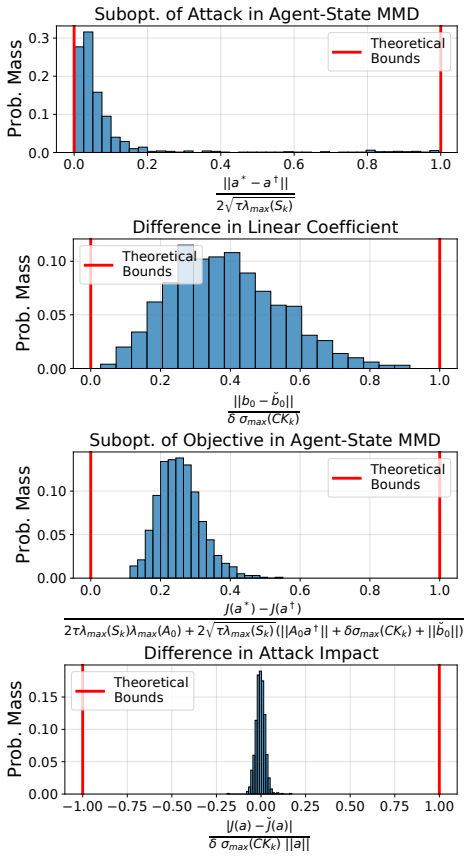

Fig. 3: Derived bounds for MMD attack from $N = 1000$ Monte Carlo simulations normalized by upper bound to fit between [{-1, 0}, 1]: Comparison with bounds from (a) Theorem 1, (b) Lemma 1, (c) Theorem 2, and (d) Theorem 3.

VI. Conclusion

We defined myopic maximum deviation and myopic adversarial state approach attacks. When attacking EKFs with a χ^2 anomaly detector, each attacker goal can be formulated as a convex QCQP. We provided practical relaxations to ensure run-time feasibility given an appropriate attacker knowledge model. Finally, we showed that the difference between the optimal and relaxed problems is bounded. Future work will use this as a basis to derive attacks with relaxed information models and develop robust estimators for nonlinear systems.

References

[1] S. Amin, X. Litrico, S. S. Sastry, and A. M. Bayen, "Stealthy deception attacks on water scada systems," 2010, pp. 161–170.

[2] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Info. Syst. Sec*, vol. 14, p. 33, 2011.

[3] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. S. Sastry, "Cyber security analysis of state estimators in electric power systems." IEEE, 2010, pp. 5991–5998.

[4] N. Hashemi, C. Murguia, and J. Ruths, "A comparison of stealthy sensor attacks on control systems," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 973–979.

[5] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, 2017.

[6] M. Pajic, I. Lee, and G. J. Pappas, "Attack-resilient state estimation for noisy dynamical systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 82–92, 2017.

[7] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. J. Pappas, "Robustness of attack-resilient state estimators," in *2014 ACM/IEEE ICCPS*, 2014, pp. 163–174.

[8] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[9] A. Khazraei and M. Pajic, "Attack-resilient state estimation with intermittent data authentication," *Automatica*, vol. 138, 2022.

[10] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks against nonlinear state estimation in smart power grids," in *2013 IEEE Power & Energy Society General Meeting*, 2013, pp. 1–5.

[11] S. Liu, G. Wei, Y. Song, and Y. Liu, "Extended kalman filtering for stochastic nonlinear systems with randomly occurring cyber attacks," *Neurocomputing*, vol. 207, pp. 708–716, 2016.

[12] A. Khazraei, S. Hallyburton, Q. Gao, Y. Wang, and M. Pajic, "Learning-based vulnerability analysis of cyber-physical systems," in *2022 ACM/IEEE ICCPS*. IEEE, 2022, pp. 259–269.

[13] I. Jovanov and M. Pajic, "Relaxing integrity requirements for attack-resilient cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4843–4858, 2019.

[14] 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.

[15] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[16] R. Horst and H. Tuy, *Global optimization: Deterministic approaches*. Springer Science & Business Media, 2013.

[17] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I. Dynamic models," *IEEE Transactions on aerospace and electronic systems*, vol. 39, no. 4, pp. 1333–1364, 2003.