Learning-based Attacks in Cyber-Physical Systems

Mohammad Javad (MJ) Khojasteh
Center for Autonomous Systems and Technologies (CAST)
California Institute of Technology

Joint work with:

- Anatoly Khina, Tel Aviv University
- Massimo Franceschetti, University of California, San Diego
- Tara Javidi, University of California, San Diego

Cloud robots and automation systems

Security

We need to address physical security in addition to cyber security

News reports

Port of San Diego suffers cyber-attack, second port in a week after Barcelona

Hacker jailed for revenge sewage attacks

Job rejection caused a bit of a stink

HACKERS REMOTELY KILL A JEEP ON THE HIGHWAY—WITH ME IN IT

News reports

The Stuxnet outbreak

A worm in the centrifuge

The Economist

An unusually sophisticated cyber-weapon is mysterious but important

Computer virus Stuxnet a 'game changer,' DHS official tells Senate

CNN

Symantec

"It has changed the way we view the security threat"

The man in the middle

A fictitious plant for
the controller

Plant

Controller

A malicious controller
for the plant

Mathematical formulation

- Linear dynamical system

$$X_{k+1} = aX_k + U_k + W_k$$

$$\{W_k\} \text{ are i.i.d. } \mathcal{N}(0, Var[W])$$

- The controller, at time k, observes Y_k and generates a control signal U_k as a function of all past observations Y_1^k.

$Y_k = X_k$ Under normal operation

$Y_k = V_k$ Under attack

- The attacker feeds a malicious input $\tilde{U}_k$ to the plant.
- How can the controller detect that the system is under attack?

Anomaly detection

- The controller is armed with a detector that tests for anomalies in the observed history Y_1^k.
- Under legitimate system operation we expect

$$Y_{k+1} - aY_k - U_k(Y_1^k) \sim \text{ i.i.d. } \mathcal{N}(0, Var[W])$$

- The detector performs the variance test

$$\frac{1}{T}\sum_{k=1}^{T}\left[Y_{k+1} - aY_k - U_k(Y_1^k)\right]^2 \in (Var[W] - \delta, Var[W] + \delta).$$

- What kind of attacks can we detect?

The man in the middle attack types

Stuxnet

Replay attack

Y. Mo, B. Sinopoli (2009)

Statistical-duplicate attack

$$X_{k+1} = aX_k + U_k + W_k$$

B. Satchidanandan,
P. R. Kumar (2017)
R. S. Smith (2011)

Learning-based attack

$$X_{k+1} = aX_k + U_k + W_k$$

MJ Khojasteh et al.
(2020)

Learning-based attack

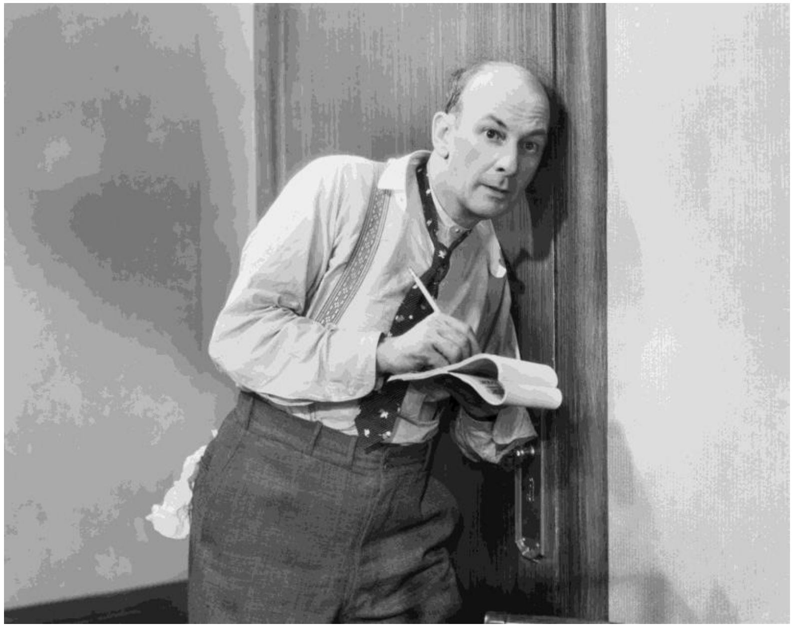
$$X_{k+1} = aX_k + U_k + W_k.$$

- The attacker has access to both X_k and U_k and knows the distribution of W_k and of the initial condition X_0, but it should learn the open loop gain a of the plant.
- For analysis purposes, we can assume the open loop gain of the plant is a random variable A with a distribution known to the attacker and for any event C we let

$$\mathbb{P}_a(C) = \mathbb{P}(C|A = a).$$

Two phases of the learning-based attack

Learning (exploration) phase

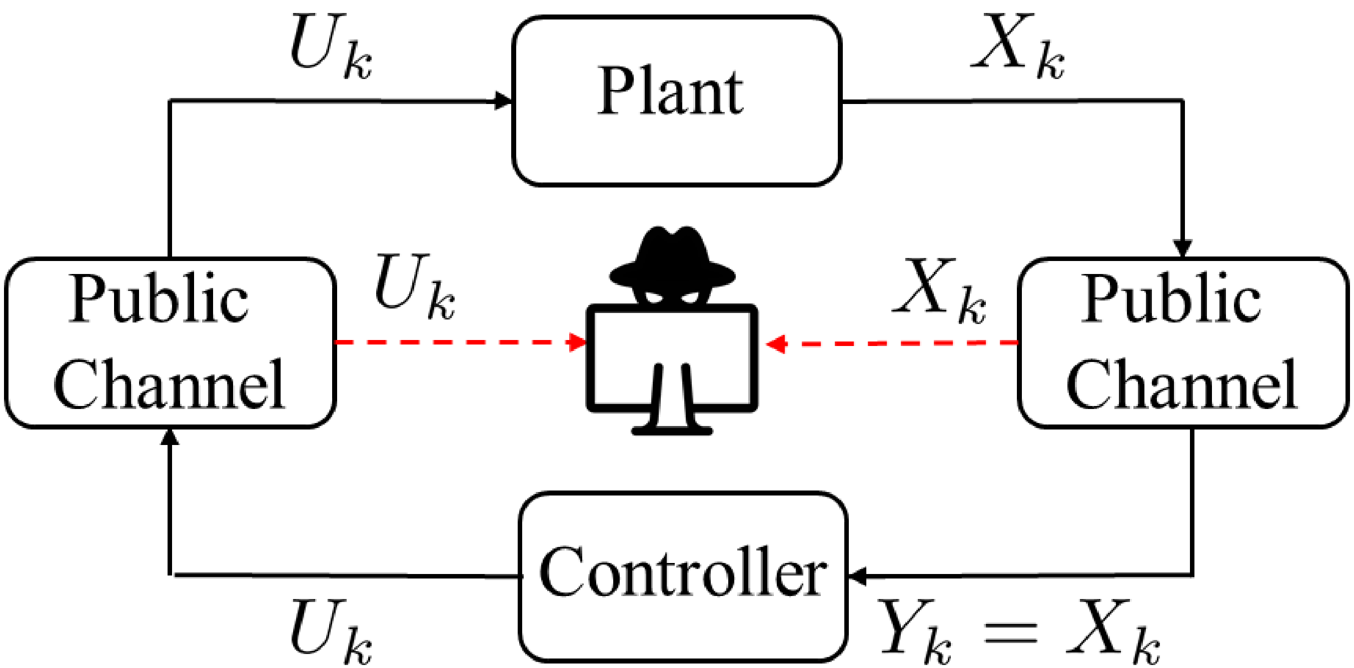


Eavesdropping and learning

Hijacking (exploitation) phase

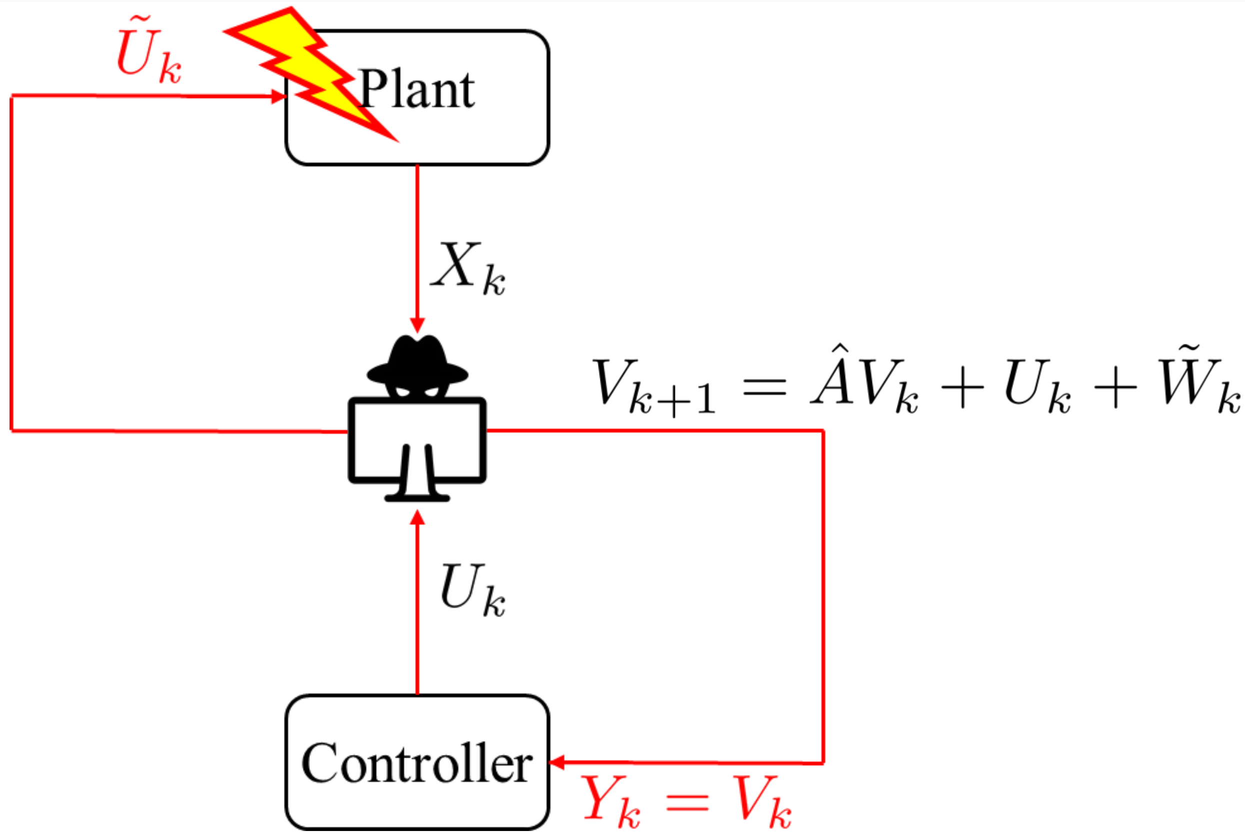
Hijacking the system

Learning (exploration) phase



- For $k \in [0, L]$, the attacker observes the plant state and control input, and tries to learn the open-loop gain a.

Hijacking (exploitation) phase



- For $k = L+1, \ldots, T$, the attacker feeds the fake signal V_k to the controller, reads the next input U_k, and drives the system to an undesired state by feeding $\tilde{U}_k$ to the plant.

Detecting the attack

- Let Θ_T be the indicator of the attack at any time before T
- The controller uses Y_1^T to construct an estimate $\hat{\Theta}_T$ of Θ_T according to the variance test
- Define the deception probabilities

$$P_{dec}^{a,T} \triangleq \mathbb{P}_a \left(\hat{\Theta}_T = 0 \middle| \Theta_T = 1 \right)$$

$$P_{dec}^{T} \triangleq \mathbb{P} \left(\hat{\Theta}_T = 0 \middle| \Theta_T = 1 \right) = \int_{-\infty}^{\infty} P_{dec}^{a,T} f_A(a) da$$

- Assume the power of the fictitious sensor reading converges a.s.

$$\lim_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{k=L+1}^{T} V_k^2 = \tfrac{1}{\beta} < \infty$$

Results

- We provide lower and upper bounds on the deception probability
- The lower bound is based on a given learning algorithm and holds for any measurable control policy
- The upper bound holds for any learning algorithm, and any measurable control policy

Lower bound

- Assuming the attacker uses a least-square learning algorithm to learn the plant, such that

$$\hat{A} = \arg\min_{A} \|X_{k+1} - AX_k - U_k\| = \frac{\sum_{k=1}^{L-1}(X_{k+1} - U_k)X_k}{\sum_{k=1}^{L-1} X_k^2}$$

- This algorithm is consistent, namely

$$\hat{A} \xrightarrow{P} a \quad \text{as} \quad L \to \infty$$

K. J. Åström, P. Eykhoff (1971), L Ljung (1982)

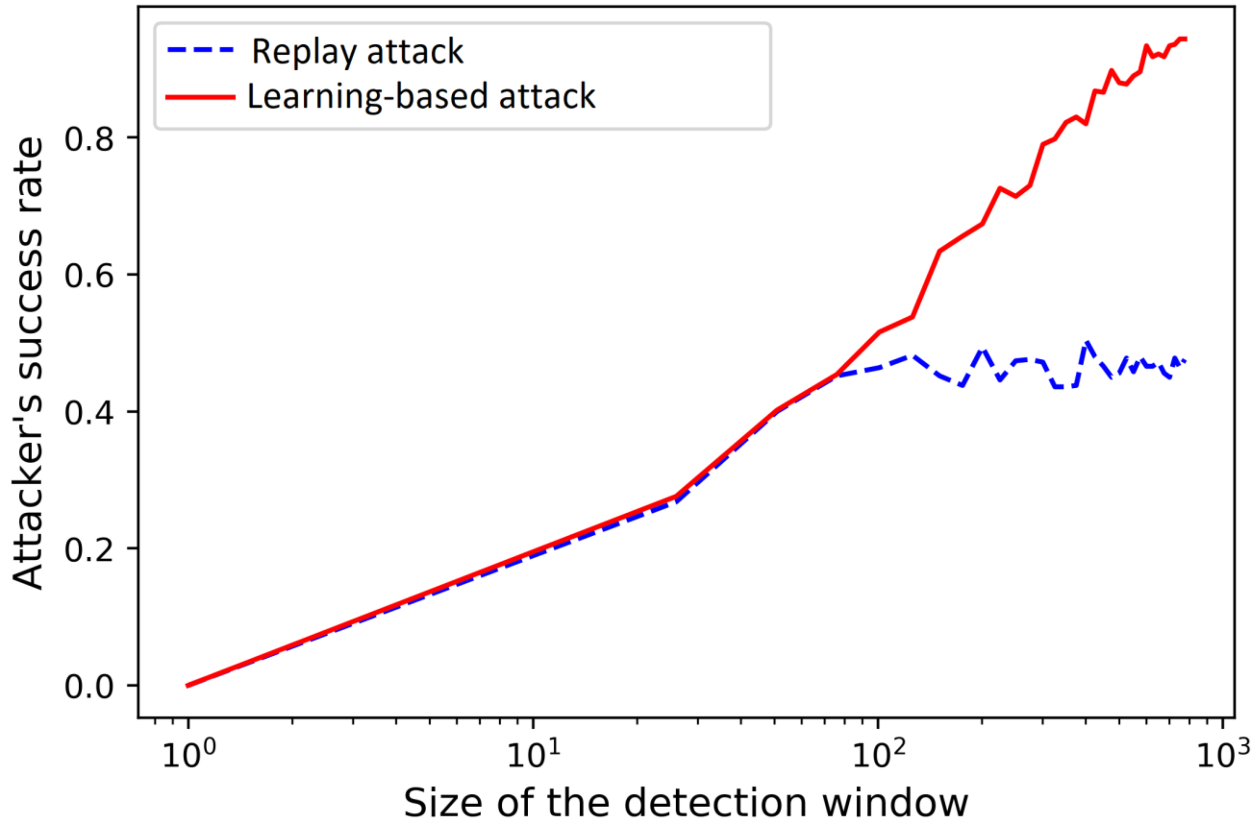
Lower bound

- On the other hand, for any fixed L the deception probability depends on the ability to learn the plant, and we can show

$$\lim_{T\to\infty} P^{a}_{\text{dec}} = \mathbb{P}_a\left(|\hat{A} - a| < \sqrt{\delta\beta}\right)$$
$$\geq 1 - \frac{2}{(1+\delta\beta)^{L/2}}$$

Using concentration bound of A. Rantzer 2018

Comparison with a replay attack



MJ Khojasteh et al. (2020)

Upper bound on the deception probability

- If A is distributed uniformly in $[-R, R]$, then letting $Z_1^k = (X_1^k, U_1^k)$, we have

$$\lim_{T\to\infty} P_{dec} \leq \frac{I(A; Z_1^L) + 1}{\log(R/\sqrt{\delta\beta})}.$$

- The numerator represents the information revealed about A from the observation of the random variable Z.
- The denominator represents the intrinsic uncertainty of A when it is observed at resolution $\epsilon = \sqrt{\delta\beta}$ corresponding to the entropy of the quantized random variable $H(A_\epsilon)$.

Upper bound on the deception probability

- In addition, if $A \to (X_k, Z_1^{k-1}) \to U_k$ is a Markov chain for all $k \in \{1, \ldots, L\}$, then

$$\lim_{T\to\infty} P_{dec} \le \frac{I(A; Z_1^L) + 1}{\log(R/\sqrt{\delta\beta})}$$

$$\le \frac{\sum_{k=1}^{L} D\left(\mathbb{P}_{X_k|Z_1^{k-1},A} \middle\| \mathbb{Q}_{X_k|Z_1^{k-1}} \middle| \mathbb{P}_{Z_1^{k-1},A}\right) + 1}{\log(R/\sqrt{\delta\beta})}$$

any sequence of probability measures $\left\{\mathbb{Q}_{X_k|Z_1^{k-1}}\right\}$, provided

$\mathbb{P}_{X_k|Z_1^{k-1}} \ll \mathbb{Q}_{X_k|Z_1^{k-1}}$ for all $k \in \{1, \ldots, L\}$.

The Gaussian case

- The freedom in choosing the auxiliary probability measure $\left\{\mathbb{Q}_{X_k|Z_1^{k-1}}\right\}$ make the second bound a useful bound.
- Gaussian plant disturbance $W_k \sim \mathcal{N}(0, Var[W])$
- By choosing $\mathbb{Q}_{X_k|Z_1^{k-1}} \sim \mathcal{N}(0, Var[W])$ we have

$$\lim_{T\to\infty} P_{dec} \leq G(Z_1^L),$$

where $$G(Z_1^L) \triangleq \frac{\frac{\log e}{2\sigma^2}\sum_{k=1}^{L}\mathbb{E}(AX_{k-1}+U_{k-1})^2+1}{\log\left(R/\sqrt{\delta\beta}\right)}.$$

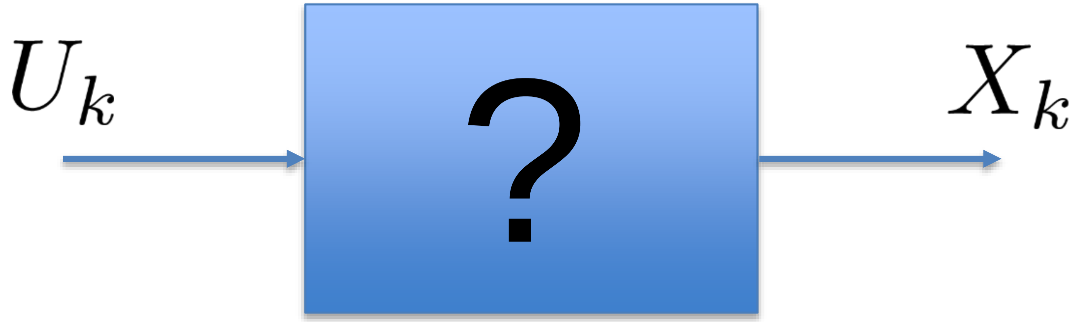
Privacy-enhancing signal

Impede the learning process of the attacker

$$\underbrace{U_k = \underbrace{\bar{U}_k}_{\text{Nominal control policy}} + \overbrace{\Gamma_k}^{\text{Privacy-enhancing signal}}}$$

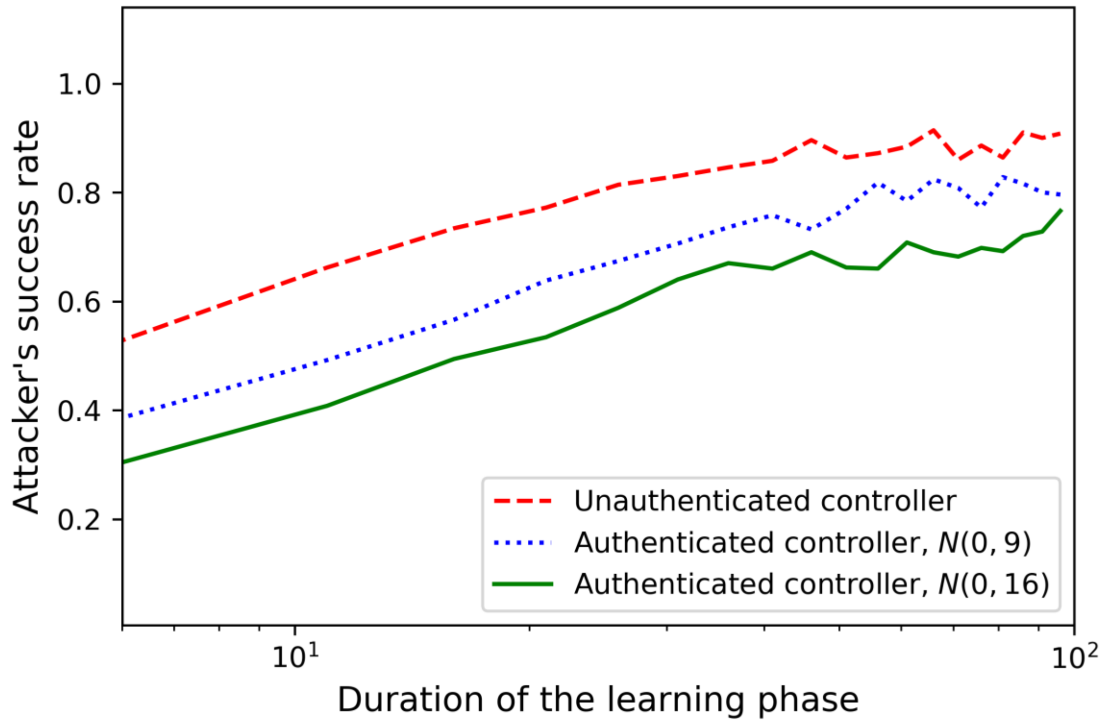
Privacy-enhancing signal

- Injecting a strong noise may in fact speed up the learning process



- Carefully crafted watermarking signals provide better guarantees on the deception probability

Defense against learning-based attack



MJ Khojasteh et al. (2020)

Vector systems

$$A \longrightarrow \mathbf{A} = \left[\ ?\ \right]$$

$$\mathrm{Var} \longrightarrow \mathrm{Cov}$$

$$\hat{A} = \frac{\sum_{k=1}^{L-1} (X_{k+1} - U_k) X_k}{\sum_{k=1}^{L-1} X_k^2}$$

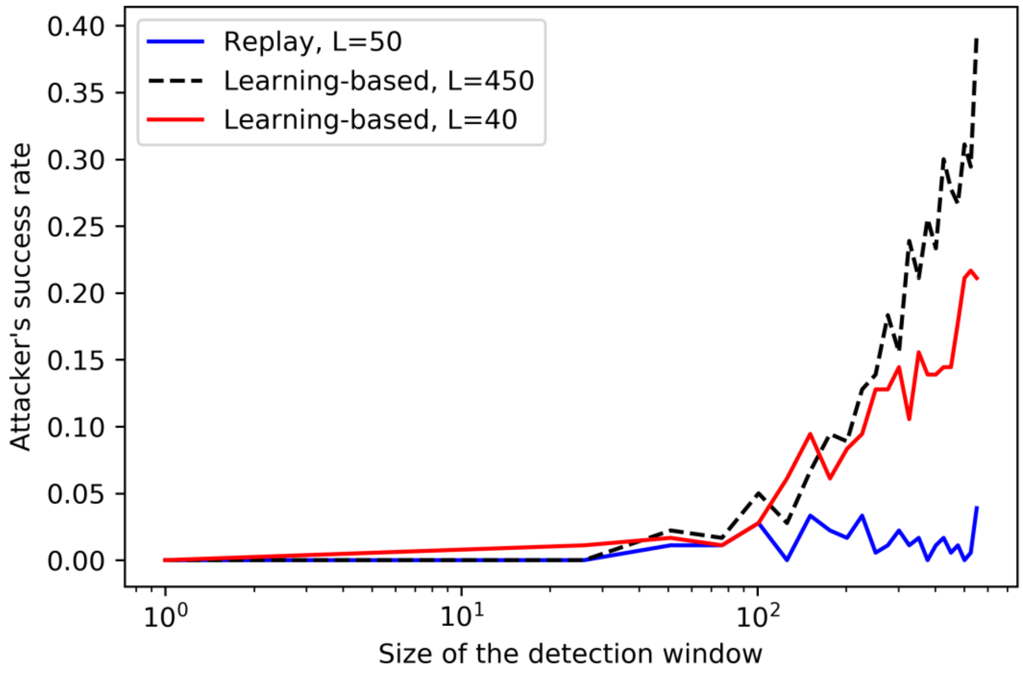
$$\downarrow$$

$$\mathbf{A} = \begin{cases} \mathbf{0}_{n \times n}, & \det(\mathbf{G}_{L-1}) = 0; \\ \sum_{k=1}^{L-1} \left((\mathbf{X}_{k+1} - \mathbf{U}_k) \mathbf{X}_k^{\dagger} \right) \mathbf{G}_{L-1}^{-1}, & \text{otherwise.} \end{cases}$$

Learning-based attack: vector systems

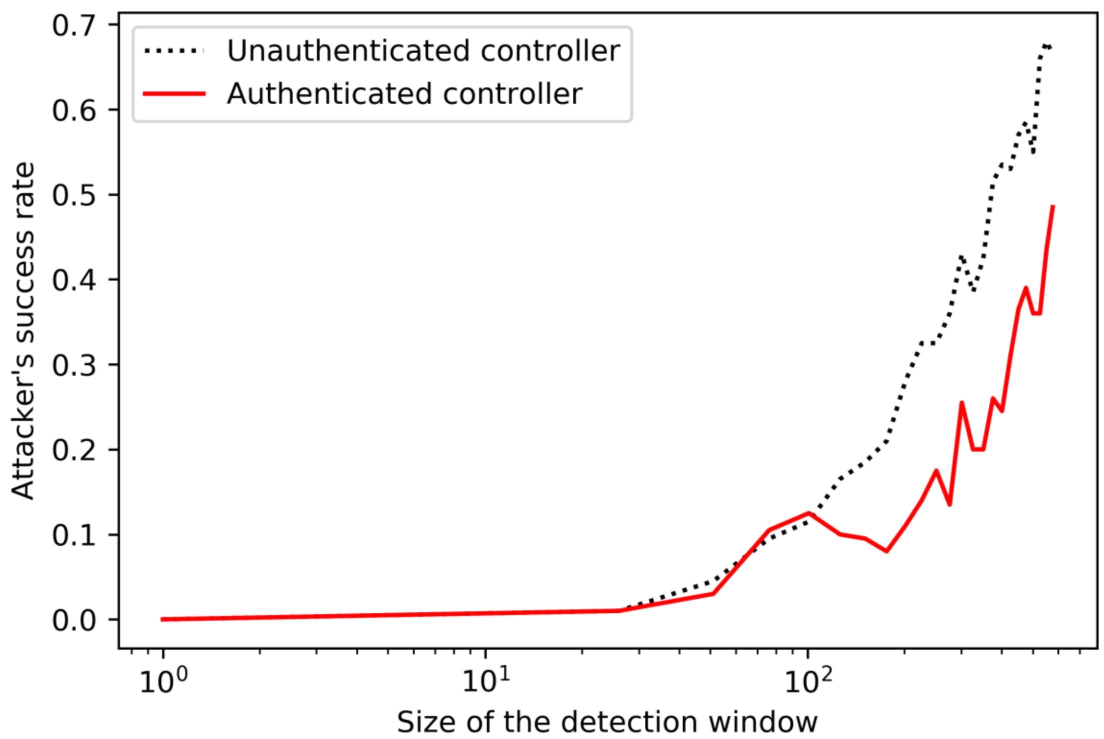
$A \longrightarrow \mathbf{A} = \left[\ ?\ \right]$

$\mathrm{Var} \longrightarrow \mathrm{Cov}$



MJ Khojasteh et al. (2020)

Defense against vector learning-based attack

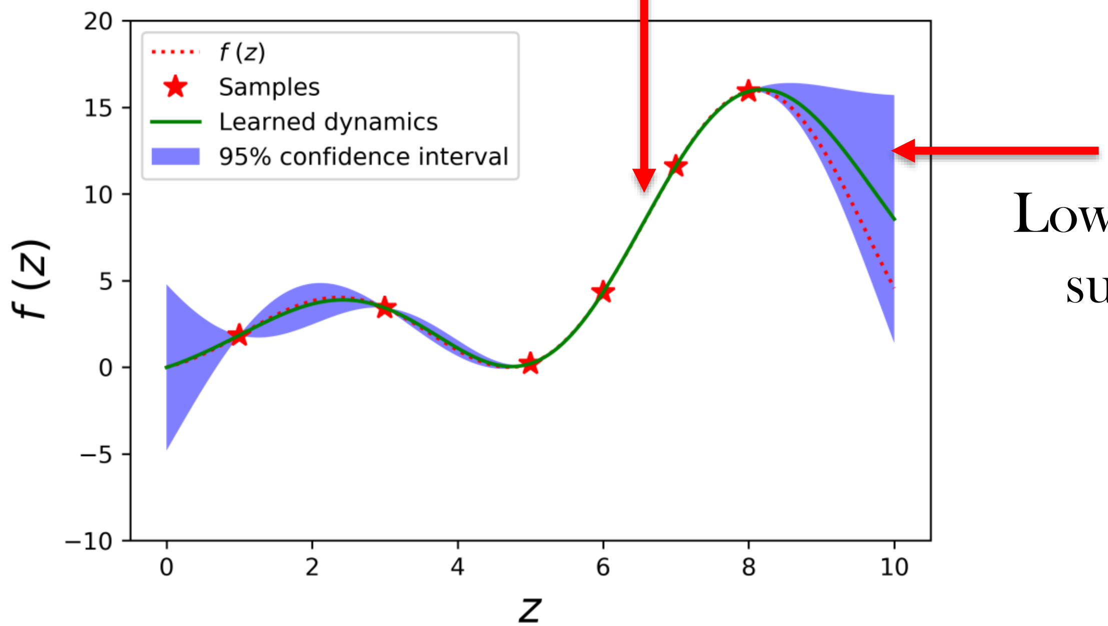


Nonlinear learning-based attack

A → $f(X, U) \in$ Reproducing Kernel Hilbert Space (RKHS)

Linear regression → Bayesian learning: Gaussian processes (GP)

Vulnerable region



Lower attacker's success rate

References

- Khojasteh MJ, Khina A, Franceschetti M, Javidi T.
  Authentication of cyber-physical systems under learning-based attacks.
  IFAC-PapersOnLine. 2019 Jan 1; 52(20): 369-74.

- Khojasteh, M.J., Khina, A., Franceschetti, M. and Javidi, T.
  Learning-based attacks in cyber-physical systems.
  arXiv preprint arXiv:1809.06023, 2020.