ADVERSARIAL REGRESSION FOR DETECTING ATTACKS IN CYBER-PHYSICAL SYSTEMS

AMIN GHAFOURI, YEVGENIY VOROBEYCHIK, XENOFON KOUTSOUKOS

INSTITUTE FOR SOFTWARE INTEGRATED SYSTEMS

VANDERBILT UNIVERSITY

Control Systems and the Quest for Autonomy: A Symposium in Honor of Panos Antsaklis
University of Notre Dame, Oct. 27-28, 2018

VANDERBILT UNIVERSITY

MOTIVATION: RESILIENT AUTONOMOUS CPS

Cyber-physical systems (CPS), such as self-driving cars and process control systems, deeply intertwine physical and software components. Their failure has *physical consequences*.

- **In 2008, a major Turkish oil pipeline suffered a cyber-attack**
  - Attackers **disabled** the pressure and flow **sensors**, which allowed them to super-pressurize the oil in the pipeline, causing an **explosion**
  - Control room did not learn about the blast until 40 minutes after it happened

- **Hackers can mess with traffic lights to jam roads and reroute cars (2014)**
  - Wireless vehicle detection systems based on magnetic sensors embedded in roadways
  - Unsecure communication protocol lacks integrity protection
  - Attacker needs to be physically near the sensors

Sensys Networks VDS240

Attributes of Resilient Autonomous CPS

- Functional correctness (by design)
- Robustness to *reliability* failures
- **Survivability against *cyber* attacks**

SENSOR ATTACKS

Sensors may be under attack by adversaries that exploit zero-day vulnerabilities and/or physical access

Attackers can falsify sensor data (i.e., integrity attack)

Undetected attacks on *critical sensors* may cause significant damage, such as reactor explosion

- Controllers often attempt to maintain the physical system state in a “safe” range
- If an observed sensor value (pressure) is too low, the controller will increase pressure

Safety monitoring typically relies on anomaly detection but stealthy attacks are possible

Cyber-attack on German steel plant (2014)

REGRESSION-BASED ANOMALY DETECTION

1. **Predictor**
   - Predicts **sensor measurements as a function of measurements of other sensors**
   - Learn $\hat{y_s} = f_s(y_{-s})$, predicted measurement of each sensor *s* as a function of *measured* values of other sensors
2. **Detector**
   - Given **residuals** (i.e., difference between observed and predicted), **determines** whether to raise an alarm
   - $|y_s - \hat{y_s}| \leq \tau_s$ where τ_s is a predefined threshold to trigger an anomaly alarm
3. **But anomaly detectors can be vulnerable to sensor attacks themselves**

ATTACK MODEL

Capability

- Compromise a subset of **sensors** and **perturb** their values
  - Can compromise at most *B* sensors (attack budget)

Knowledge

- Attacker has **complete knowledge** of the system and implementations

Objective

- Maximize/minimize the observed value for some **critical sensor** to cause damage
- **Constraint**: Remain **undetected by the anomaly detector** (stealthy attack)

ATTACKER'S PROBLEM

Given:

- A collection of regression-based anomaly detectors $\{|y_s - \hat{y}_s| \leq \tau_s\}$
- A critical sensor s_c
- A budget constraint *B* (the number of sensors that can be attacked)

Compute the optimal *stealthy* (undetected) attack (which sensors to compromise, and what their observed measurements should be) to maximize (minimize) measured value of the critical sensor

- For example, minimizing *observed* sensor value of pressure can lead the controller to increase actual pressure

$$\min y_{s_c}$$

Stealth $s.t.\!: |y_s - f(y_{-s})| \leq \tau_s$

Budget $||y - y_{true}||_0 \leq B$

ATTACKER'S PROBLEM

✓**Proposition: Attacker's Problem is NP-Hard *even when linear regression is used for anomaly detection*.**

✓**We devise:**

- ✓Exact solution for linear regression models (integer linear program)
- ✓Iterative algorithm for the nonlinear (e.g., neural network regression) case (heuristic)

Amin Ghafouri, Yengeniy Vorobeychik, and Xenofon Koutsoukos. "Adversarial Regression for Detecting Attacks in Cyber-Physical Systems", *27th International Joint Conference on Artificial Intelligence and 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018)*, Stockholm, Sweden, July 13-19, 2018.

SPECIAL CASE: LINEAR REGRESSION

$|y_s - f(y_{-s})| \leq \tau_s$ **: can be represented using linear constraints (since *f()* is linear)**

$||y - y_{true}||_0 \leq B$ **: can be represented using linear constraints if we add binary variables indicating which sensors are attacked**

Thus, the problem can be solved using a Mixed-Integer Linear Program (MILP)

GENERALIZING

$|y_s - f(y_{-s})| \leq \tau_s$ **: cannot be represented using linear constraints for arbitrary non-linear *f()***

ALGORITHM FOR ATTACKING GENERAL NON-LINEAR MODELS

1. **Obtain a linearized model by a first-order Taylor expansion around the solution estimate**
2. **Transform the problem to a MILP**
3. **Constrain solutions to be close to previous iterate (trust region)**
4. **If the solution of MILP is infeasible w.r.t. stealth constraint,**

   reduce trust region
5. **Repeat.**

CASE STUDY: TENNESSEE-EASTMAN PROCESS CONTROL SYSTEM (TE-PCS)

Involving two simultaneous gas-liquid exothermic reactions for producing two liquid products

$$A(g) + C(g) + D(g) \rightarrow G(liq), \quad \text{Product } 1,$$
$$A(g) + C(g) + E(g) \rightarrow H(liq), \quad \text{Product } 2.$$

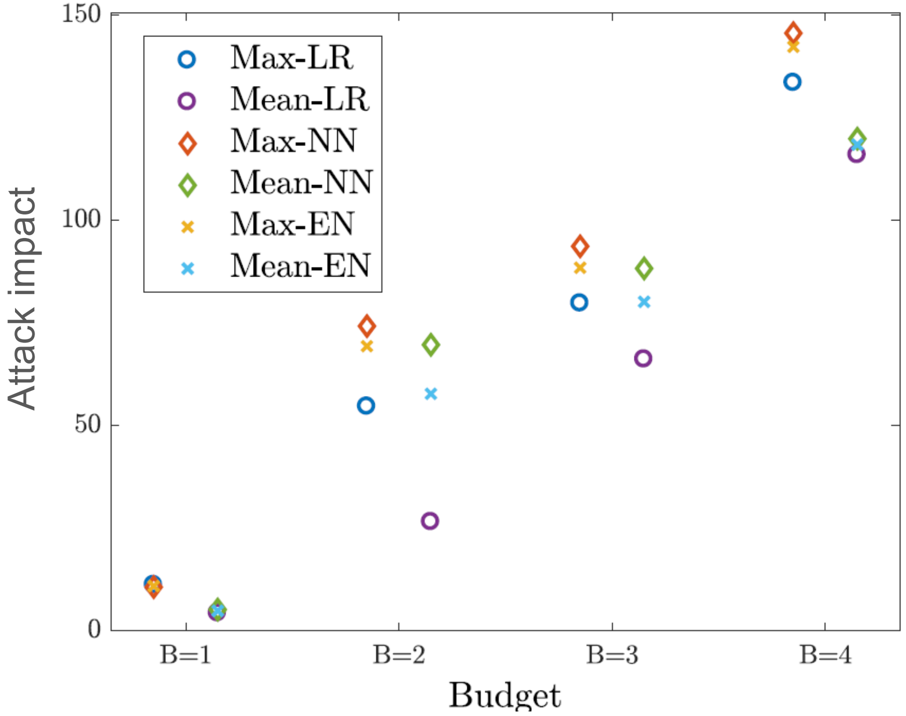
Five major units: reactor, condenser, vapor-liquid separator, recycle compressor, and product stripper.

Safety monitoring using 41 measurement outputs and 12 control inputs.

Consider linear regression and neural network regression for anomaly detection

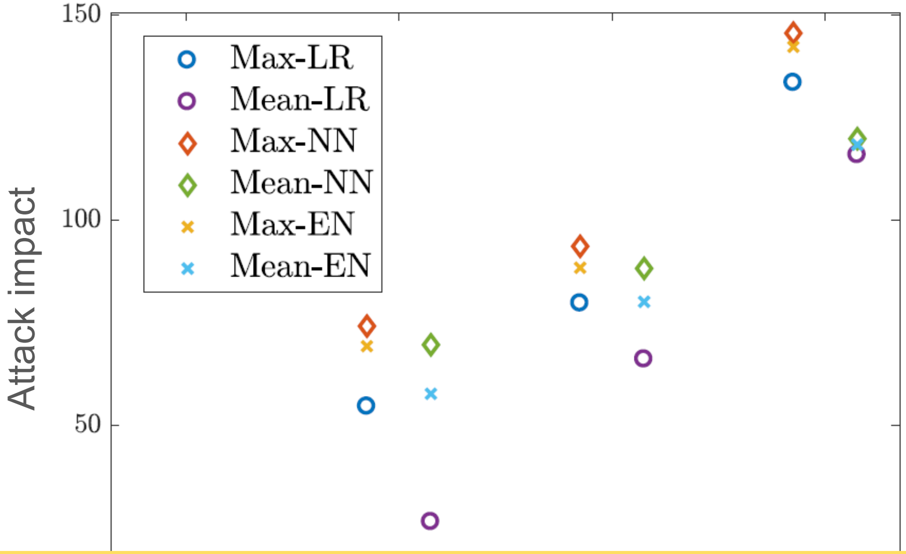
ATTACKING PRESSURE OF REACTOR

Maximum and mean of the solution of adversarial regression:



ATTACKING PRESSURE OF REACTOR

Maximum and mean of the solution of adversarial regression:



Neural network (diamonds) is more vulnerable than linear regression (circles)!

DEFENDING AGAINST ATTACKS

In the anomaly detection system, the defender can leverage the stealth constraint of the attacker's problem by appropriately choosing the detector thresholds

Trade off:

- Impact of attack (maximum distortion of critical sensor values induced by the attacker)
- False alarm rate

Problem:

- Minimize impact of attack (optimal solution to attacker's problem)
- Subject to: False alarm rate is at most z

HEURISTIC ALGORITHM FOR OPTIMIZING THRESHOLDS

Start with a baseline detector with false alarm rate *z*

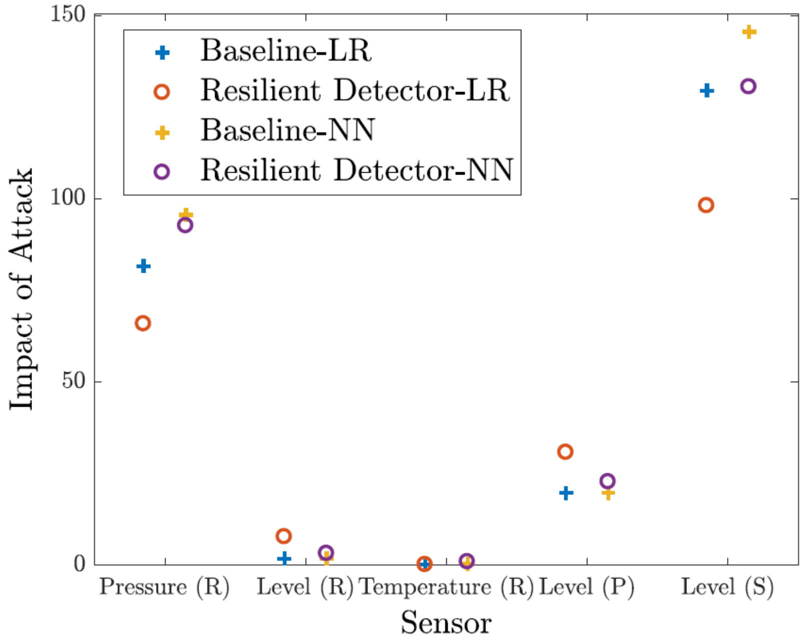
Iteratively:

- Find optimal attack
  - A : Sensors with largest attack impact
  - B : Sensors with smallest impact
- Reduce threshold on sensors in A
- Increase threshold on sensors in B to keep false alarm rate at *z*
- Stop when no longer reducing overall attack impact

EXPERIMENTS: RESILIENT DETECTOR

Same setting as before

Maintain the same # of false alarms as for an initial non-resilient detector



Significant reduction in attack impact relative to baseline for most vulnerable sensors

SUMMARY

Described a general regression learning framework for Anomaly Detection in CPS

Studied stealthy attacks in CPS considering

- Linear regression
- General regression (illustrated using Neural Networks)

Proposed an approach to design a more Resilient Detector while maintain the same overall false alarm rate as for a baseline detector

Resilient anomaly detection can improve survivability against cyber attacks and increase trust in autonomous CPS