



Cyber-Physical Security in Energy Systems

Part II: Attack Modeling and Detection Methods

DTU PES Summer School 2026

Henrik Sandberg (hsan@kth.se)

KTH EECS, Decision and Control Systems

Outline

- DoS attack modeling and dynamic instability
- FDI attack modeling and detection
 - Unknown state and input estimation
 - CUSUM
 - PCA
- Main references:
 - [S. Sundaram, "Fault-Tolerant and Secure Control Systems," Lecture Notes, 2010: https://engineering.purdue.edu/~sundara2/misc/ft_control_lecture_notes.pdf]
 - [H. Sandberg, Lecture notes, EL2850, KTH, 2025]

Disclaimer (😊)



Bart De Moor
(1960-...)

“At the end of the day,
the only thing we really understand,
is linear algebra”.

📌 Pinned



Peyman Milanfar ✓ @docmilanfar · 1h

99% of everything that works in practice is calculus and linear algebra. The other 1% doesn't really work.

Attack Model: Denial-of-Service (DoS) on Actuator Channel

Confidentiality attack: disclosure (c_u, c_y)
 Integrity attack: false data injection (i_u, i_y)
 Availability attack: DoS attack (a_u, a_y)

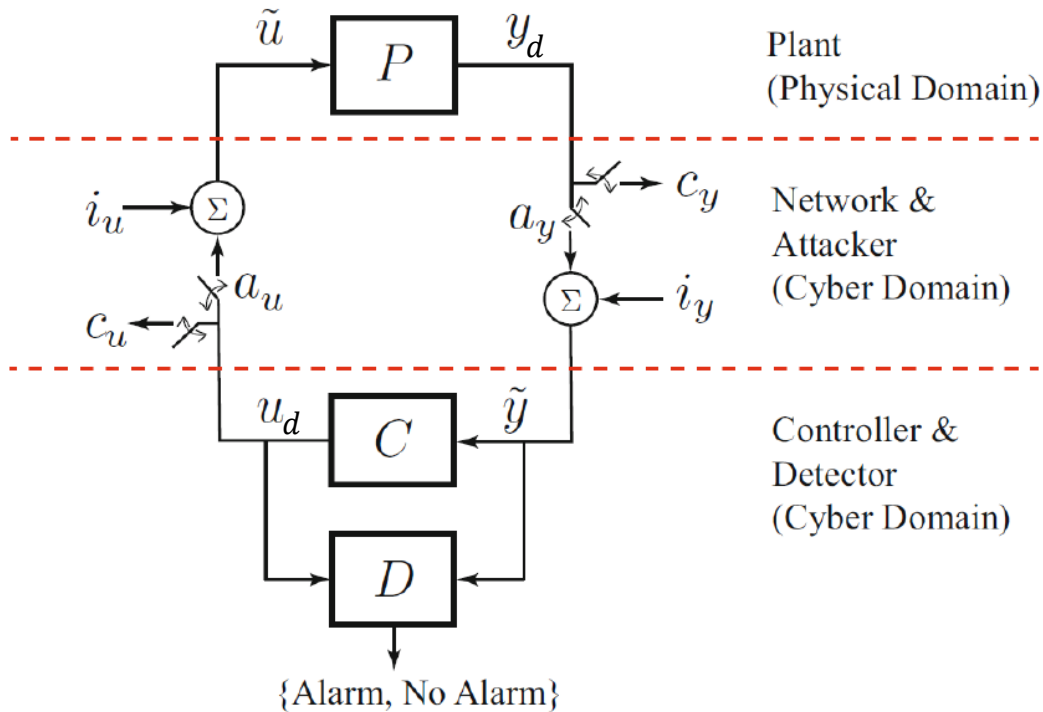
Assume plant P can be modeled by linear discrete-time state-space model (time index $k \geq 0$):

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k]$$

$$\mathbf{u}[k] = \begin{cases} -\mathbf{K}\mathbf{x}[k], & \text{no drop with probability } p, \\ 0, & \text{drop with probability } q := 1 - p \end{cases}$$

Theorem: Suppose $\mathbf{B}$ is square and has full rank. The expected value of the state $\mathbf{x}[k]$ is bounded, iff $\rho(\mathbf{A})^2 q < 1$
 ($\rho(\mathbf{A})$ is largest magnitude of eigenvalues of $\mathbf{A}$)

Inherently unstable plants P are much more sensitive to a few dropped control commands. Stable plants lose performance, but *not* stability



Example 1: DoS Attack in F-8 Longitudinal Dynamics

2.3.2 Longitudinal Dynamics of an F-8 Aircraft

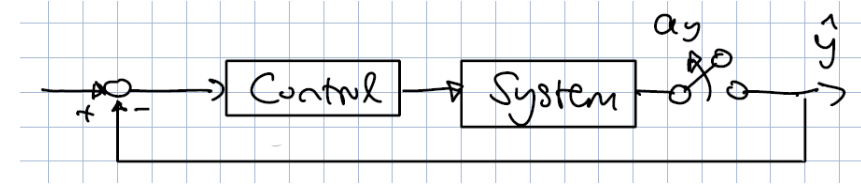
Consider a model for the (sampled) linearized longitudinal dynamics of an F-8 aircraft [87]:

$$\underbrace{\begin{bmatrix} V[k+1] \\ \gamma[k+1] \\ \alpha[k+1] \\ q[k+1] \end{bmatrix}}_{\mathbf{x}[k+1]} = \begin{bmatrix} 0.9987 & -3.2178 & -4.4793 & -0.2220 \\ 0 & 1 & 0.1126 & 0.0057 \\ 0 & 0 & 0.8454 & 0.0897 \\ 0.0001 & -0.0001 & -0.8080 & 0.8942 \end{bmatrix} \underbrace{\begin{bmatrix} V[k] \\ \gamma[k] \\ \alpha[k] \\ q[k] \end{bmatrix}}_{\mathbf{x}[k]} + \begin{bmatrix} -0.0329 \\ 0.0131 \\ -0.0137 \\ -0.0092 \end{bmatrix} \mathbf{u}[k], \quad (2.4)$$

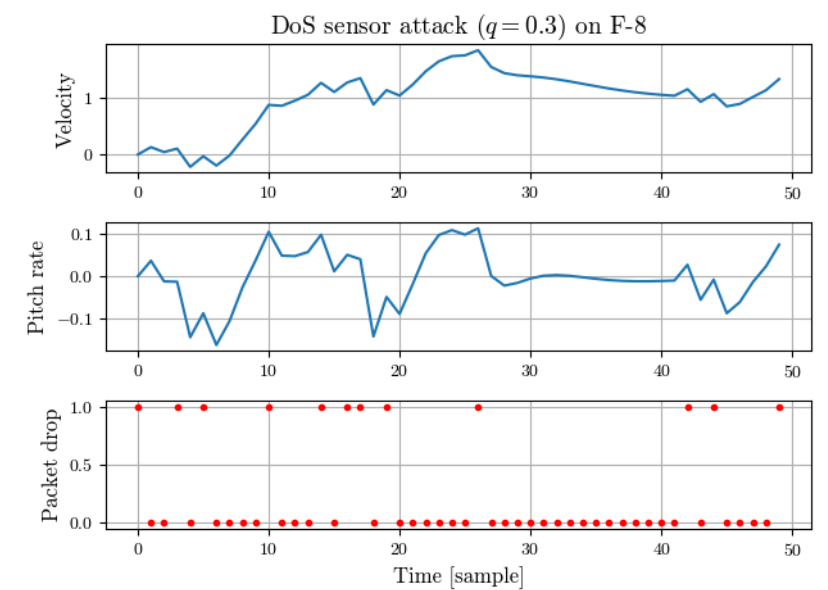
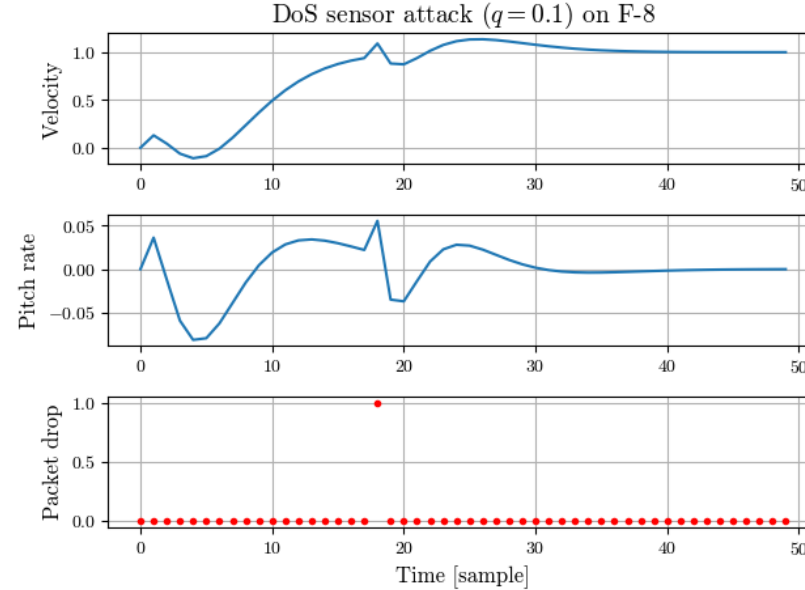
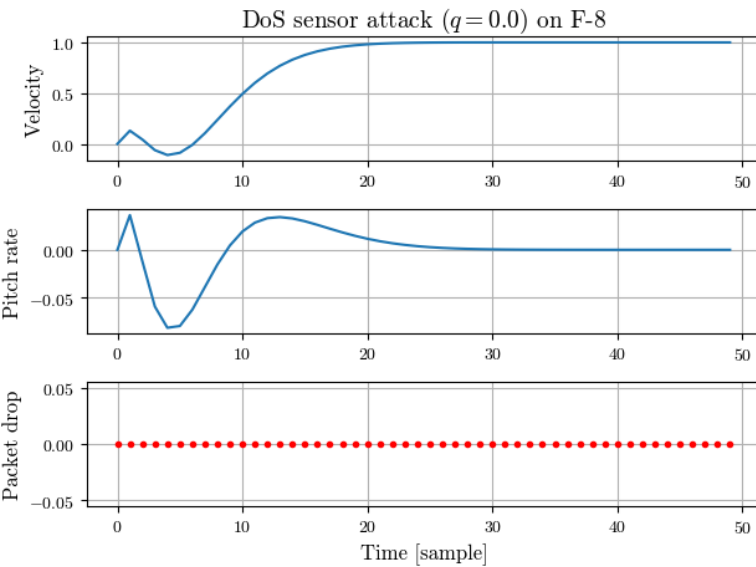
where $V[k]$ is the velocity of the aircraft, $\gamma[k]$ is the flight-path angle, $\alpha[k]$ is the angle-of-attack, and $q[k]$ is the pitch rate. The input to the aircraft is taken to be the deflection of the elevator flaps.

The output of the system will depend on the sensors that are installed on the aircraft. For example, if there is a sensor to measure the velocity and the pitch rate, the output would be given by

$$\mathbf{y}[k] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}[k].$$

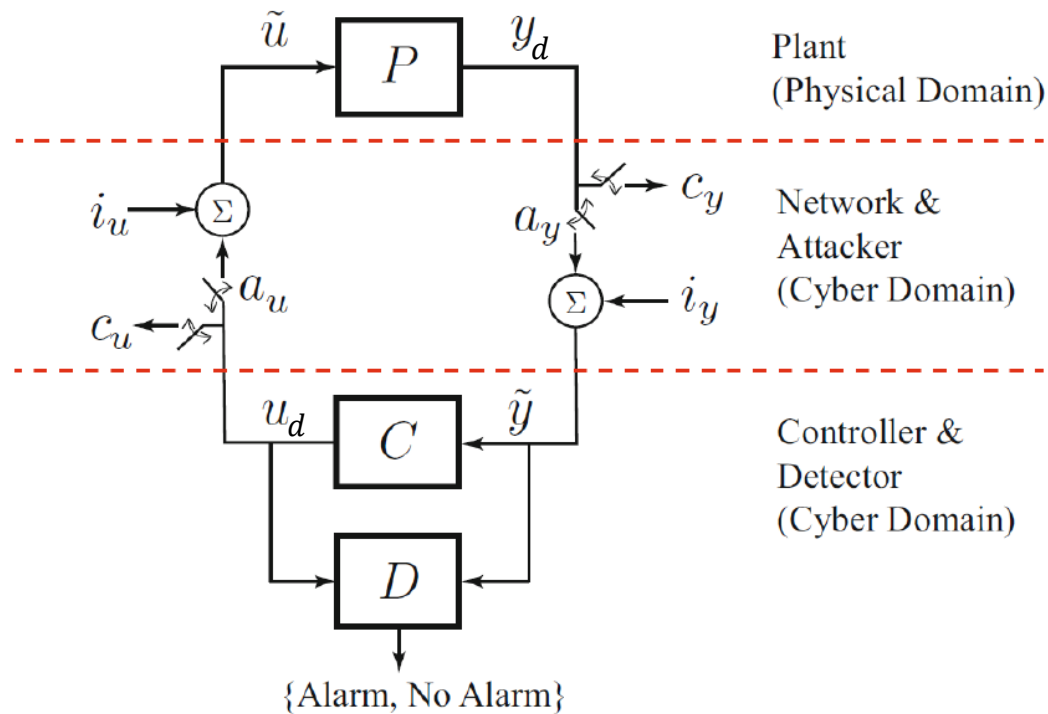


Example 1: DoS Attack in F-8 Longitudinal Dynamics



[DoS_F8.py]

Attack Model: False Data Injection (FDI)



Confidentiality attack: disclosure (c_u, c_y)
 Integrity attack: false data injection (i_u, i_y)
 Availability attack: DoS attack (a_u, a_y)

Assume plant P can be modeled by linear discrete-time state-space model (time index $k \geq 0$):

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}'\tilde{\mathbf{u}}[k], & \mathbf{x}[k] \in \mathbb{R}^n, \tilde{\mathbf{u}}[k] \in \mathbb{R}^{m'} \\ \tilde{\mathbf{y}}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{D}'\tilde{\mathbf{u}}[k] + \mathbf{i}_y[k], & \tilde{\mathbf{y}}[k], \mathbf{i}_y[k] \in \mathbb{R}^p \end{aligned}$$

Sensors and actuators can be corrupted by (i_u, i_y)

Problem: How, and when, can we design a Detector that alarms when $(i_u, i_y) \neq 0$?

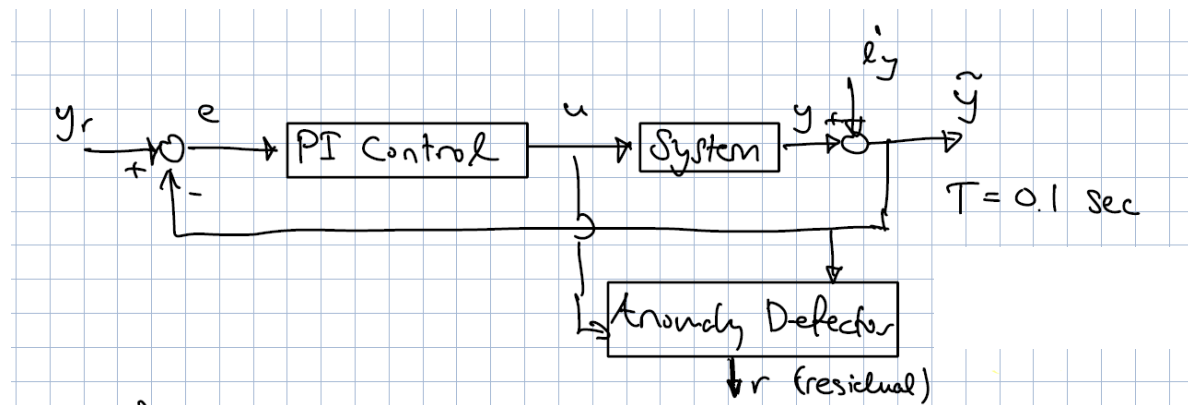
Example 2: FDI Attack in Car Speed Control

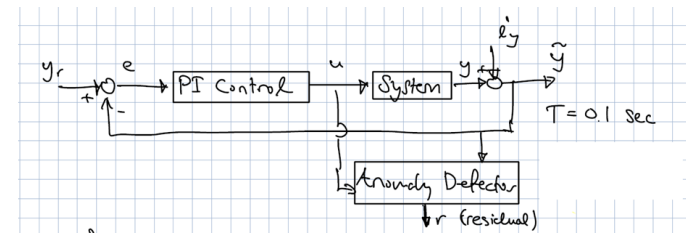
2.3.1 A Simple Model for a Car

Consider again the model of the car from Chapter 1, with speed $v(t)$ at any time t , and acceleration input $a(t)$. Along the lines of Example 1.1, suppose that we *sample* the velocity of the car every T seconds, and that the acceleration is held constant between sampling times. We can then write

$$v[k+1] = v[k] + Ta[k], \quad (2.1)$$

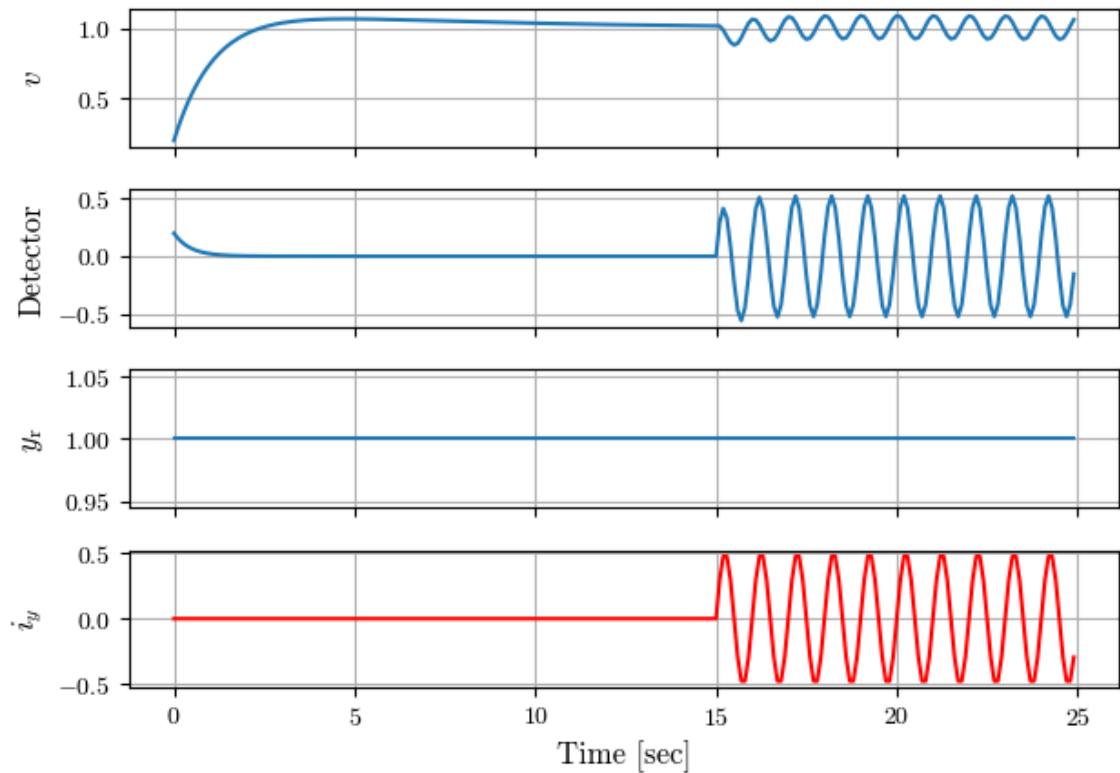
where $v[k]$ is shorthand for $v(kT)$, $k \in \mathbb{N}$, and $a[k]$ is the acceleration that is applied at time $t = kT$. Now, suppose that we wish to also consider the amount of fuel in the car at



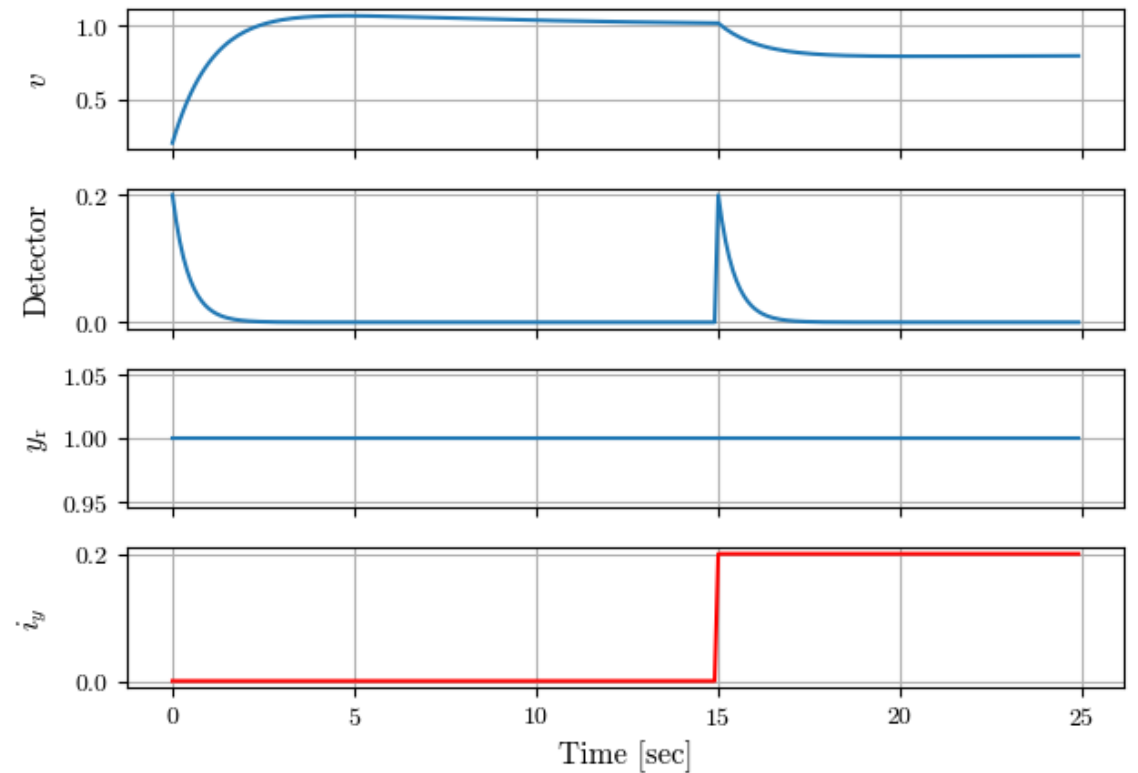


Example 2: FDI Attack in Car Speed Control

Sinusoidal attack on sensor



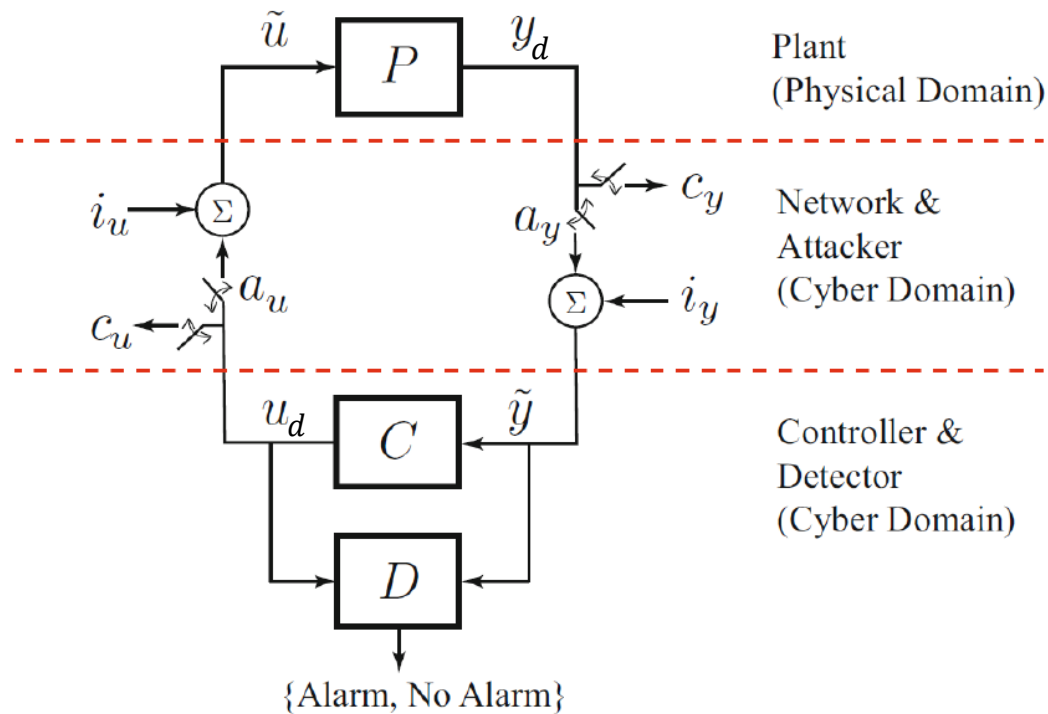
Bias attack on sensor



[FDI_car.py]

When can Detector estimate i_y ?

Attack Model: False Data Injection (FDI)



Confidentiality attack: disclosure (c_u, c_y)
 Integrity attack: false data injection (i_u, i_y)
 Availability attack: DoS attack (a_u, a_y)

Assume plant P can be modeled by linear discrete-time state-space model (time index $k \geq 0$):

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}'\tilde{\mathbf{u}}[k], & \mathbf{x}[k] \in \mathbb{R}^n, \tilde{\mathbf{u}}[k] \in \mathbb{R}^{m'} \\ \tilde{\mathbf{y}}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{D}'\tilde{\mathbf{u}}[k] + \mathbf{i}_y[k], & \tilde{\mathbf{y}}[k], \mathbf{i}_y[k] \in \mathbb{R}^p \end{aligned}$$

Sensors and actuators can be corrupted by (i_u, i_y)

Problem: How, and when, can we design a Detector that alarms when $(i_u, i_y) \neq 0$?

Assumptions on Detector

- Assume desired control u_d known
- Use linearity and subtract the desired control from the model, and collect *all unknown perturbations in a new attack signal* $\mathbf{u}[k] = (\mathbf{i}_u[k], \mathbf{i}_y[k]) \in \mathbb{R}^m$
- New model

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k], & \mathbf{x}[k] &\in \mathbb{R}^n, \mathbf{u}[k] \in \mathbb{R}^m \\ \mathbf{y}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k], & \mathbf{y}[k] &\in \mathbb{R}^p \end{aligned}$$

- **Known (typically):** Model $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ and output $\mathbf{y}[k], k \geq 0$
- **Unknown (typically):** Input $\mathbf{u}[k]$ and state $\mathbf{x}[k], k \geq 0$

Reformulated problem: Alarm as soon as possible when $\mathbf{u}[k] \neq 0$

Prototypical Estimation Problems

- Depending on assumptions on what variables are known and unknown, several estimation problems can be posed, and solved using linear algebra:
 - Problem 1: Estimation of $\mathbf{x}$ from $\mathbf{y}$ and $\mathbf{u}$ (*observability*)
 - Problem 2: Estimation of $\mathbf{u}$ from $\mathbf{y}$ and $\mathbf{x}[0]$
 - Problem 3: Estimation of $\mathbf{u}$ and $\mathbf{x}$ from $\mathbf{y}$ (*strong observability*)
 - Problem 4: Asymptotic Estimation of $\mathbf{u}$ and $\mathbf{x}$ from $\mathbf{y}$ (*strong detectability*)

Useful Result from Linear Algebra

Consider the model ($\mathbf{y}$ received measurement, $\mathbf{x}$ unknown state, $\mathbf{u}$ unknown attack):

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \quad \mathbf{y} \in \mathbb{R}^p, \mathbf{x} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m$$

(U1): There is at least one exact solution $\mathbf{x}, \mathbf{u}$ iff $\text{rank} [\mathbf{C} \quad \mathbf{D} \quad \mathbf{y}] = \text{rank} [\mathbf{C} \quad \mathbf{D}]$

(Compute one solution using $[\mathbf{x} \quad \mathbf{u}]^T = [\mathbf{C} \quad \mathbf{D}]^+ \mathbf{y}$, where $+$ denotes a pseudoinverse)

(U2): The solution $\mathbf{u}$ is unique iff $\text{rank} [\mathbf{C} \quad \mathbf{D}] = \text{rank} [\mathbf{C}] + m$

- Sample applications:
 - **(U1 - anomaly detection):** Is $\mathbf{y}$ consistent with the state and attack model?
 - **(U2 - attack reconstruction):** Can I uniquely reconstruct the attack $\mathbf{u}$ from $\mathbf{y}$?

From Dynamics to a Linear Equation

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] \\ \mathbf{y}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k] \end{aligned}$$

$$\mathbf{x}[k+1] = \mathbf{A}^{k+1}\mathbf{x}[0] + \sum_{i=0}^k \mathbf{A}^{k-i}\mathbf{B}\mathbf{u}[i]$$

$$= \mathbf{A}^{k+1}\mathbf{x}[0] + \underbrace{[\mathbf{A}^k\mathbf{B} \quad \mathbf{A}^{k-1}\mathbf{B} \quad \dots \quad \mathbf{A}\mathbf{B} \quad \mathbf{B}]}_{\mathbf{C}_k} \underbrace{\begin{bmatrix} \mathbf{u}[0] \\ \mathbf{u}[1] \\ \vdots \\ \mathbf{u}[k-1] \\ \mathbf{u}[k] \end{bmatrix}}_{\mathbf{u}[0:k]}$$

$$\equiv \mathbf{A}^{k+1}\mathbf{x}[0] + \mathbf{C}_k\mathbf{u}[0:k].$$

Problem 1: Estimation of $\mathbf{x}$ from $\mathbf{y}$ and $\mathbf{u}$

- State estimation: estimate **unknown state** from **known input and output**
- Stack data over time horizon $[0, L]$:

$$\underbrace{\begin{bmatrix} \mathbf{y}[0] \\ \mathbf{y}[1] \\ \mathbf{y}[2] \\ \vdots \\ \mathbf{y}[L] \end{bmatrix}}_{\mathbf{y}[0:L]} = \underbrace{\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^L \end{bmatrix}}_{\mathcal{O}_L} \mathbf{x}[0] + \underbrace{\begin{bmatrix} \mathbf{D} & 0 & 0 & \dots & 0 \\ \mathbf{CB} & \mathbf{D} & 0 & \dots & 0 \\ \mathbf{CAB} & \mathbf{CB} & \mathbf{D} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{CA}^{L-1}\mathbf{B} & \mathbf{CA}^{L-2}\mathbf{B} & \mathbf{CA}^{L-3}\mathbf{B} & \dots & \mathbf{D} \end{bmatrix}}_{\mathcal{J}_L} \underbrace{\begin{bmatrix} \mathbf{u}[0] \\ \mathbf{u}[1] \\ \mathbf{u}[2] \\ \vdots \\ \mathbf{u}[L] \end{bmatrix}}_{\mathbf{u}[0:L]}$$

with *observability matrix* $\mathcal{O}_L$ and *invertibility matrix* $\mathcal{J}_L$

- Rearranging we obtain $\mathbf{y}[0:L] - \mathcal{J}_L \mathbf{u}[0:L] = \mathcal{O}_L \mathbf{x}[0]$ and we can uniquely solve for $\mathbf{x}[0]$ iff $\text{rank } \mathcal{O}_L = n$ (system is *observable*; use **(U2)**)
- Theorem:** System is observable iff $\text{rank } \mathcal{O}_{n-\text{rank}(\mathbf{C})} = n$
- Remark:** There is a delay $L \leq n - \text{rank}(\mathbf{C})$ before we can estimate the state $\mathbf{x}[0]$

Example 3: Observability

Example 2.4. Consider the pair $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $\mathbf{C} = [1 \quad 0]$ $\mathbf{x}$, with no inputs to the system. The observability matrix for this pair is

$$\mathcal{O}_{n-\text{rank}(\mathbf{C})} = \mathcal{O}_1 = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

which has rank 2, and thus the pair is observable. The initial state of the system can be recovered as follows:

$$\mathbf{y}[0 : 1] = \mathcal{O}_1 \mathbf{x}[0] \Rightarrow \mathcal{O}_1^{-1} \mathbf{y}[0 : 1] = \mathbf{x}[0].$$

Problem 2: Estimation of $\mathbf{u}$ from $\mathbf{y}$ and $\mathbf{x}[0]$

- Input estimation: estimate **unknown input** from **known input and state**
- Stack data over time horizon $[0, L]$:

$$\underbrace{\begin{bmatrix} \mathbf{y}[0] \\ \mathbf{y}[1] \\ \mathbf{y}[2] \\ \vdots \\ \mathbf{y}[L] \end{bmatrix}}_{\mathbf{y}[0:L]} = \underbrace{\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^L \end{bmatrix}}_{\mathcal{O}_L} \underbrace{\mathbf{x}[0]}_{\mathbf{x}[0]} + \underbrace{\begin{bmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CB} & \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CAB} & \mathbf{CB} & \mathbf{D} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{CA}^{L-1}\mathbf{B} & \mathbf{CA}^{L-2}\mathbf{B} & \mathbf{CA}^{L-3}\mathbf{B} & \cdots & \mathbf{D} \end{bmatrix}}_{\mathcal{J}_L} \underbrace{\begin{bmatrix} \mathbf{u}[0] \\ \mathbf{u}[1] \\ \mathbf{u}[2] \\ \vdots \\ \mathbf{u}[L] \end{bmatrix}}_{\mathbf{u}[0:L]}$$

with *observability matrix* $\mathcal{O}_L$ and *invertibility matrix* $\mathcal{J}_L$

- Rearranging we obtain $\mathbf{y}[0:L] - \mathcal{O}_L \mathbf{x}[0] = \mathcal{J}_L \mathbf{u}[0:L]$ and can uniquely solve for $\mathbf{u}[0]$ iff $\text{rank } \mathcal{J}_L = \text{rank } \mathcal{J}_{L-1} + m$ (system is *invertible* with delay L ; use **(U2)**)
- Theorem:** If system not invertible for $L \leq n - \text{nullity}(\mathbf{D}) + 1$, then it is never invertible
- Remark:** To determine $\mathbf{u}[1]$, shift the horizon $[0, L] \rightarrow [1, L+1]$
- Lemma:** System is invertible (for some L) iff $\text{rank} \left(\begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right) = n + m$ for *at least one* $z \in \mathbb{C}$

Example 4: Fault/Attack in F-8 Elevator Angle

Consider a model for the (sampled) linearized longitudinal dynamics of an F-8 aircraft [87]:

$$\underbrace{\begin{bmatrix} V[k+1] \\ \gamma[k+1] \\ \alpha[k+1] \\ q[k+1] \end{bmatrix}}_{\mathbf{x}[k+1]} = \begin{bmatrix} 0.9987 & -3.2178 & -4.4793 & -0.2220 \\ 0 & 1 & 0.1126 & 0.0057 \\ 0 & 0 & 0.8454 & 0.0897 \\ 0.0001 & -0.0001 & -0.8080 & 0.8942 \end{bmatrix} \underbrace{\begin{bmatrix} V[k] \\ \gamma[k] \\ \alpha[k] \\ q[k] \end{bmatrix}}_{\mathbf{x}[k]} + \begin{bmatrix} -0.0329 \\ 0.0131 \\ -0.0137 \\ -0.0092 \end{bmatrix} \mathbf{u}[k], \quad (2.4)$$

where $V[k]$ is the velocity of the aircraft, $\gamma[k]$ is the flight-path angle, $\alpha[k]$ is the angle-of-attack, and $q[k]$ is the pitch rate. The input to the aircraft is taken to be the deflection of the elevator flaps.

The output of the system will depend on the sensors that are installed on the aircraft. For example, if there is a sensor to measure the velocity and the pitch rate, the output would be given by

$$\mathbf{y}[k] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}[k].$$

Suppose there is a fault/attack in the elevator angle: $\tilde{\mathbf{u}}[k] = \mathbf{u}_d[k] + \mathbf{i}_u[k]$. Can we reconstruct the fault/attack using only $\mathbf{x}[0]$, $\mathbf{y}_2$, and $\mathbf{u}_d$?

Problem 3: Estimation of u and x from y

- Input and state estimation: estimate **unknown input and state** from **known output**
- Stack data over time horizon $[0, L]$:

$$\underbrace{\begin{bmatrix} y[0] \\ y[1] \\ y[2] \\ \vdots \\ y[L] \end{bmatrix}}_{y[0:L]} = \underbrace{\begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^L \end{bmatrix}}_{\mathcal{O}_L} \underbrace{x[0]}_{x[0]} + \underbrace{\begin{bmatrix} D & 0 & 0 & \dots & 0 \\ CB & D & 0 & \dots & 0 \\ CAB & CB & D & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{L-1}B & CA^{L-2}B & CA^{L-3}B & \dots & D \end{bmatrix}}_{\mathcal{J}_L} \underbrace{\begin{bmatrix} u[0] \\ u[1] \\ u[2] \\ \vdots \\ u[L] \end{bmatrix}}_{u[0:L]}$$

with *observability matrix* $\mathcal{O}_L$ and *invertibility matrix* $\mathcal{J}_L$

- We can uniquely solve for $x[0]$ iff $\text{rank} [\mathcal{O}_L \ \mathcal{J}_L] = \text{rank} \mathcal{J}_L + n$ (system is *strongly observable* with delay L ; use **(U2)**)
- Theorem:** If system not strongly observable for $L \leq n$, then it is never strongly observable
- Remark:** Once $x[k]$ and $x[k+1]$ are found, we solve for $u[k]$ using $\begin{bmatrix} x[k+1] - Ax[k] \\ y[k] \end{bmatrix} = \begin{bmatrix} B \\ D \end{bmatrix} u[k]$
- Lemma:** System is strongly observable (for some L) iff $\text{rank} \left(\begin{bmatrix} A - zI_n & B \\ C & D \end{bmatrix} \right) = n + m$ for all $z \in \mathbb{C}$ (transfer function matrix has *no zeros*, only poles)

Example 5: Strong Observability of F-8

Example 2.7. Consider again the F-8 from Example 2.6. To check whether one can recover the fault input $\mathbf{u}[k]$ can be recovered, regardless of the states of the system, we check whether the system is strongly observable. Specifically, for $L = n$, we have

$$\mathbf{y}[k] = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}[k]$$

$$\mathcal{O}_n = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.0001 & -0.0001 & -0.8080 & 0.8942 \\ 0.0001 & -0.0004 & -1.4058 & 0.7271 \\ 0.0002 & -0.0009 & -1.7765 & 0.5240 \\ 0.0002 & -0.0015 & -1.9261 & 0.3092 \end{bmatrix},$$

$$\mathcal{J}_n = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -0.0092 & 0 & 0 & 0 & 0 \\ 0.0028 & -0.0092 & 0 & 0 & 0 \\ 0.0125 & 0.0028 & -0.0092 & 0 & 0 \\ 0.0195 & 0.0125 & 0.0028 & -0.0092 & 0 \end{bmatrix}.$$

One can verify that $\text{rank}([\mathcal{O}_n \quad \mathcal{J}_n]) - \text{rank}(\mathcal{J}_n) = 1$, and thus the system is not strongly observable.

However, suppose that we also have a sensor that measures the velocity of the aircraft (in addition to the pitch rate). The $\mathbf{C}$ matrix would then become

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and one can verify that the system is strongly observable in this case. Specifically, one can recover the fault input $\mathbf{u}[k]$ from the output of the system $\mathbf{y}[k : k + n]$ without knowing the initial state of the system.

Problem 4: Asymptotic Estimation of $\mathbf{u}$ and $\mathbf{x}$ from $\mathbf{y}$

- If system is *not strongly observable*, we cannot exactly find state and input with a fixed delay using only outputs. *When and how can we asymptotically estimate state and input?*
- If system is invertible with delay L , there is a $\mathbf{P}$ such that $\mathbf{P}\mathcal{J}_L = [\mathbf{I}_m \quad 0 \quad \dots \quad 0]$
- We can write the (unknown) input as $\mathbf{P}\mathbf{y}[k : k + L] = \mathbf{P}\mathcal{O}_L\mathbf{x}[k] + \mathbf{u}[k]$
- Eliminate input in the system model

$$\begin{aligned}\mathbf{x}[k + 1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] \\ &= \mathbf{A}\mathbf{x}[k] - \mathbf{B}\mathbf{P}\mathcal{O}_L\mathbf{x}[k] + \mathbf{B}\mathbf{P}\mathbf{y}[k : k + L] \\ &= (\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L)\mathbf{x}[k] + \mathbf{B}\mathbf{P}\mathbf{y}[k : k + L]\end{aligned}$$

- Replace $\mathbf{x}$ with a state estimate $\hat{\mathbf{x}}$ leading to an *unknown input observer* (UIO):

$$\hat{\mathbf{x}}[k + 1] = (\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L)\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{P}\mathbf{y}[k : k + L]$$

- When do we have $\hat{\mathbf{x}}[k] \rightarrow \mathbf{x}[k]$ for any $\hat{\mathbf{x}}[0]$? Iff there is a $\mathbf{P}$ such that all eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L$ are inside the unit disc ($\rho(\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L) < 1$)! (System is *strongly detectable* with delay L)
- **Lemma:** System is strongly detectable (for some L) iff $\text{rank} \left(\begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right) = n + m$ for all $z \in \mathbb{C}$ such that $|z| \geq 1$ (transfer function matrix has *no zeros on or outside of the disc*. The system is *minimum phase*)
- **Remark:** Asymptotically estimate the attack using $\mathbf{P}\mathbf{y}[k : k + L] - \mathbf{P}\mathcal{O}_L\hat{\mathbf{x}}[k] = \hat{\mathbf{u}}[k]$

Comparison to Regular (Luenberger) State Observer

- Consider the system model

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k]$$

$$\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k]$$

with *known* input $\mathbf{u}[k]$

- For any filter gain $\mathbf{L}$, add output injection

$$\begin{aligned}\mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] + \underbrace{\mathbf{L}(\mathbf{y}[k] - \mathbf{C}\mathbf{x}[k])}_{\equiv 0} \\ &= (\mathbf{A} - \mathbf{L}\mathbf{C})\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{L}\mathbf{y}[k]\end{aligned}$$

- Replace $\mathbf{x}$ with a state estimate $\hat{\mathbf{x}}$ leading to the state observer

$$\hat{\mathbf{x}}[k+1] = (\mathbf{A} - \mathbf{L}\mathbf{C})\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{L}\mathbf{y}[k]$$

- When do we have $\hat{\mathbf{x}}[k] \rightarrow \mathbf{x}[k]$ for any $\hat{\mathbf{x}}[0]$? Iff there is an $\mathbf{L}$ such that all eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$ are inside the unit disc ($\rho(\mathbf{A} - \mathbf{L}\mathbf{C}) < 1$)! (Model $(\mathbf{A}, \mathbf{C})$ is *detectable*)

Problem 4: Asymptotic Estimation of $\mathbf{u}$ and $\mathbf{x}$ from $\mathbf{y}$

- To find an UIO in practice, first solve

$$\mathbf{P}\mathcal{J}_L = [\mathbf{I}_m \quad 0 \quad \dots \quad 0] \quad (\mathbf{P} = [\mathbf{I}_m \quad 0 \quad \dots \quad 0] \mathcal{J}_L^+)$$

$$\mathbf{N}\mathcal{J}_L = [0 \quad 0 \quad \dots \quad 0]$$

where $\mathbf{N}$ is a basis of the left nullspace of $\mathcal{J}_L$

- Note that for any filter gain $\mathbf{G}$ of appropriate dimensions

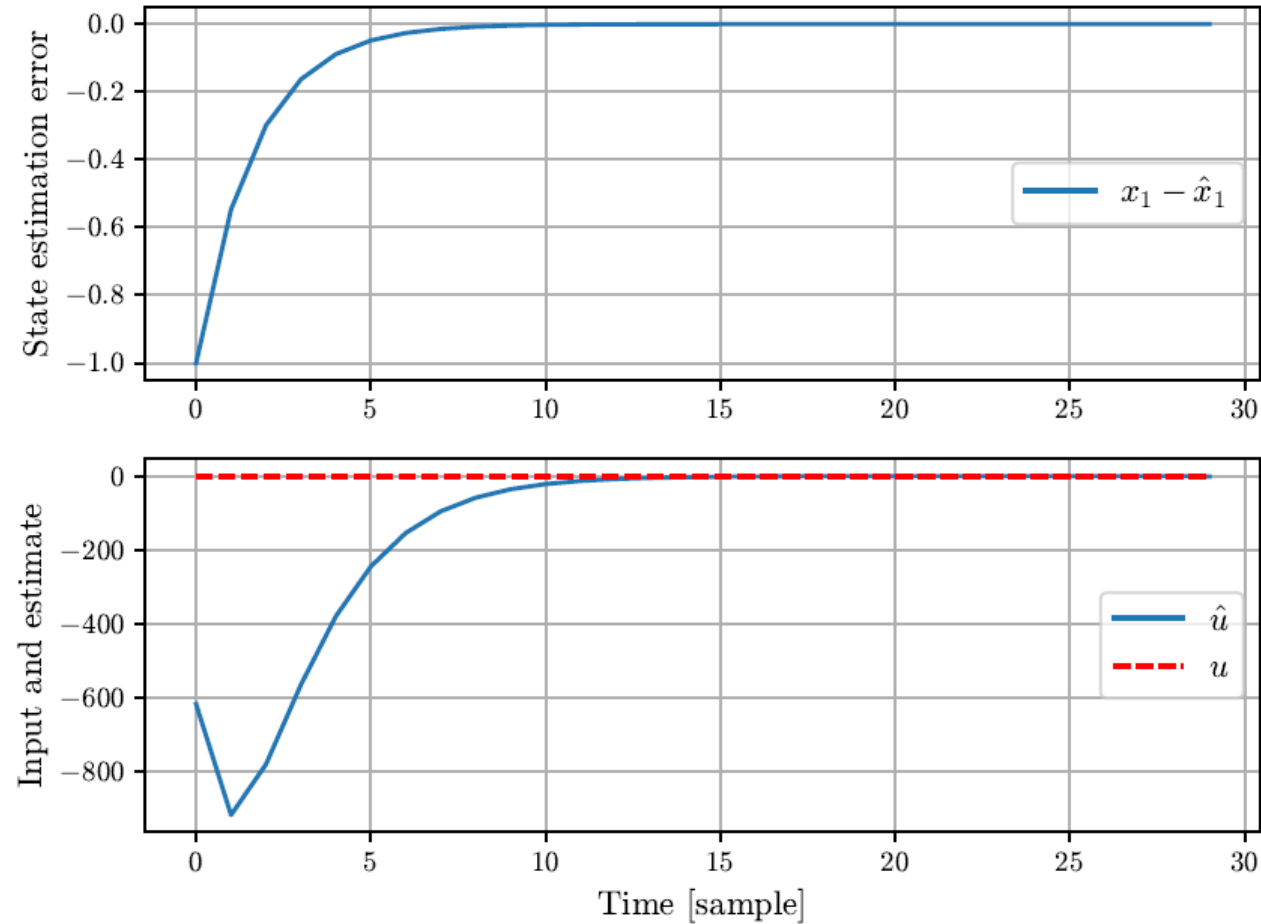
$$(\mathbf{B}\mathbf{P} + \mathbf{G}\mathbf{N})(\mathbf{y}[k : k + L] - \mathcal{O}_L\mathbf{x}[k]) = \mathbf{B}\mathbf{u}[k]$$

- Then an UIO is given by

$$\hat{\mathbf{x}}[k + 1] = (\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L - \mathbf{G}\mathbf{N}\mathcal{O}_L)\hat{\mathbf{x}}[k] + (\mathbf{B}\mathbf{P} + \mathbf{G}\mathbf{N})\mathbf{y}[k : k + L]$$

- If model $(\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L, \mathbf{N}\mathcal{O}_L)$ is *detectable* there is a gain $\mathbf{G}$ such that $\rho(\mathbf{A} - \mathbf{B}\mathbf{P}\mathcal{O}_L - \mathbf{G}\mathbf{N}\mathcal{O}_L) < 1$.
Equivalent(!) to *no zeros on or outside of the disc*
- Compare gain $\mathbf{G}$ with $\mathbf{L}$!

Example 6: Estimating Unknown Input and State in F-8 Airplane from Outputs (with delay $L = 1$)



[UIO_f8.py]

Example 7: Design of Unknown Input Observer

Consider a linear dynamical system

$$\begin{aligned}x[k+1] &= x[k] + 0.5f[k] \\ y[k] &= cx[k] + f[k],\end{aligned}\tag{3}$$

where $x \in \mathbb{R}$ is the state, $f \in \mathbb{R}$ is a potential fault, $y \in \mathbb{R}$ is the sensor measurement, and c is a constant sensor parameter.

Suppose we would like to equip the system with an unknown input observer (UIO) to detect and identify the potential fault f .

What constraints should we then consider on the parameter c when selecting an appropriate sensor for system (3)?

Outline

- DoS attack modeling and dynamic instability
- FDI attack modeling and detection
 - Unknown state and input estimation
 - **CUSUM**
 - PCA
- Main references:
 - [S. Sundaram, "Fault-Tolerant and Secure Control Systems," Lecture Notes, 2010: https://engineering.purdue.edu/~sundara2/misc/ft_control_lecture_notes.pdf]
 - [H. Sandberg, Lecture notes, EL2850, KTH, 2025]

What If There Is Stochastic Noise? Use CUSUM!

- What if measurements are corrupted by stochastic noise? Leads to a noisy residual signal

$$\mathbf{r}[k] := \hat{\mathbf{u}}[k] = \mathbf{u}[k] + \epsilon[k]$$

Suppose $\epsilon[k]$ is zero-mean i.i.d. (independent, identically distributed) stochastic noise

- A simple threshold $\eta > 0$ alarm test:

$$\|\mathbf{r}[k]\|_2^2 \underset{\text{No Alarm}}{\overset{\text{Alarm}}{\geq}} \eta$$

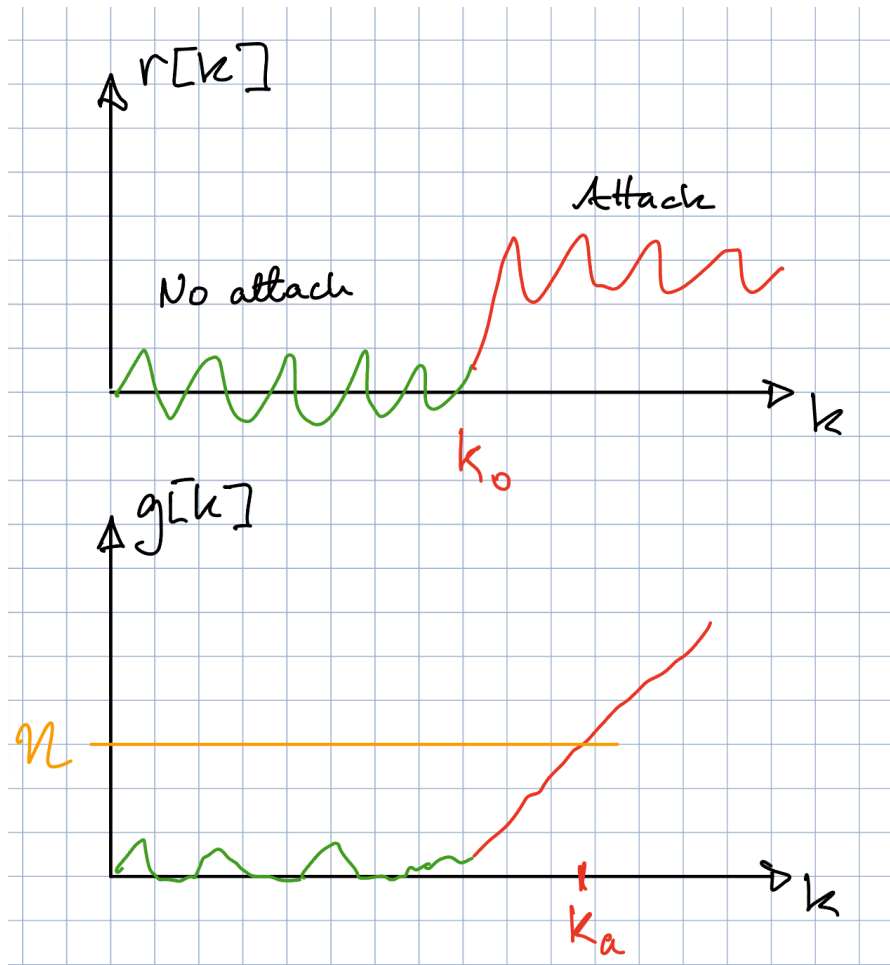
generally has a *bad trade-off between false and true alarm rates*. [**Stateless test**]

- Cumulative Sum (CUSUM) tests generally generate *far fewer false alarms* and still provide *short delay until true alarm* (sometimes optimal test structure):

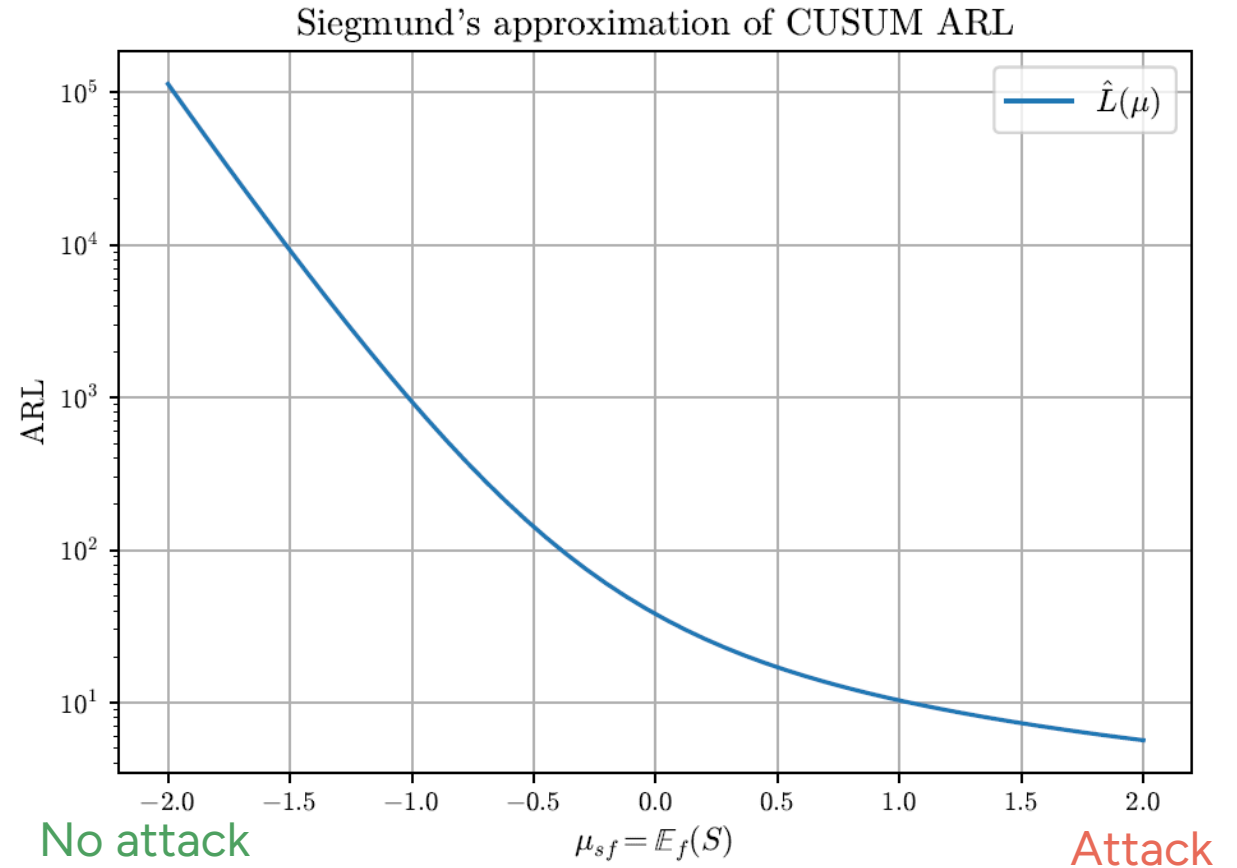
$$\begin{aligned} g[k] &= (g[k-1] + \|\mathbf{r}[k]\|_2^2 - \delta)^+ := \max\{0, g[k-1] + \|\mathbf{r}[k]\|_2^2 - \delta\} \\ g[-1] &= 0 \\ g[k] &\underset{\text{No Alarm}}{\overset{\text{Alarm}}{\geq}} \eta \end{aligned}$$

where $\delta > 0$ is a tunable bias. [**Stateful test**]

CUSUM Performance



ARL = Average Run Length (mean #samples to alarm)



Mean #samples between false alarms: $\sim 10^5$

Mean #samples alarm delay after onset attack: ~ 5

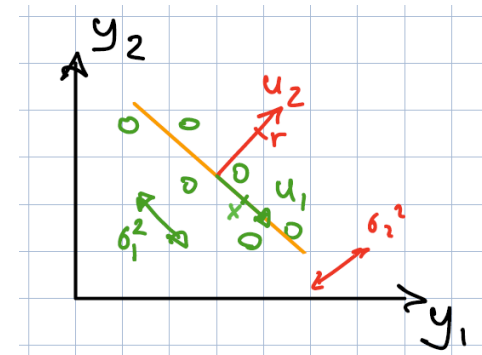
Outline

- DoS attack modeling and dynamic instability
- FDI attack modeling and detection
 - Unknown state and input estimation
 - CUSUM
 - **PCA**
- Main references:
 - [S. Sundaram, "Fault-Tolerant and Secure Control Systems," Lecture Notes, 2010: https://engineering.purdue.edu/~sundara2/misc/ft_control_lecture_notes.pdf]
 - [H. Sandberg, Lecture notes, EL2850, KTH, 2025]

What If There Is No Model? Consider PCA!

- Results so far have assumed model $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is known. What if we only have collected input/output data?
- **Indirect approach:** Estimate $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ first using system identification (subspace, ML, etc.)
- **Direct approach:** Large number of data-driven methods available (autoencoders, **PCA**, etc.)
- **Principal Component Analysis (PCA)** is a classical method from multivariate statistics
- Suppose high-dimensional sample data sequence $\mathbf{y}[0], \mathbf{y}[1], \dots, \mathbf{y}[L]$ (L and p large) collected
- Suppose nominal data (with no attack/anomalies) lie on a low-dimensional subspace in $\mathbb{R}^p$:
 1. Estimate nominal subspace of dimension $n < p$ that captures the maximum possible variance of the collected data
 2. As new data samples arrive, $\mathbf{y}[k], k > L$, compute distance to nominal subspace and *alarm if distance to subspace is large enough* (use stateless or stateful test)
- **Autoencoders:** Nonlinear generalization of PCA. Assume nominal data lies on low-dimensional (nonlinear) manifold instead of (linear) subspace. Less theory, lots of code available

PCA Formulas



- From samples $\mathbf{y}[0], \mathbf{y}[1], \dots, \mathbf{y}[L]$, first compute sample mean and center the data

$$\bar{\mathbf{y}} := \frac{1}{N} \sum_{k=0}^L \mathbf{y}[k] \quad \mathbf{z}[k] := \mathbf{y}[k] - \bar{\mathbf{y}}$$

- Compute sample covariance of data

$$\mathbf{Z} := [\mathbf{z}[0] \quad \mathbf{z}[1] \quad \dots \quad \mathbf{z}[L]] \in \mathbb{R}^{p \times N} \qquad \mathbf{\Sigma} := \frac{1}{L} \sum_{k=0}^L \mathbf{z}[k] \mathbf{z}[k]^\top = \frac{1}{L} \mathbf{Z} \mathbf{Z}^\top \in \mathbb{R}^{p \times p}$$

- Find the largest directions of variations in data ("principal components") using eigenvectors

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_p \end{bmatrix} \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_p^2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^\top \\ \mathbf{u}_2^\top \\ \vdots \\ \mathbf{u}_p^\top \end{bmatrix} \quad \sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_p^2$$

PCA Formulas

- If data mostly lies in a subspace of dimension n , then

$$\sigma_1^2 \geq \dots \geq \sigma_n^2 \gg \sigma_{n+1}^2 \geq \dots \geq \sigma_p^2$$

- Projected coordinates in subspace

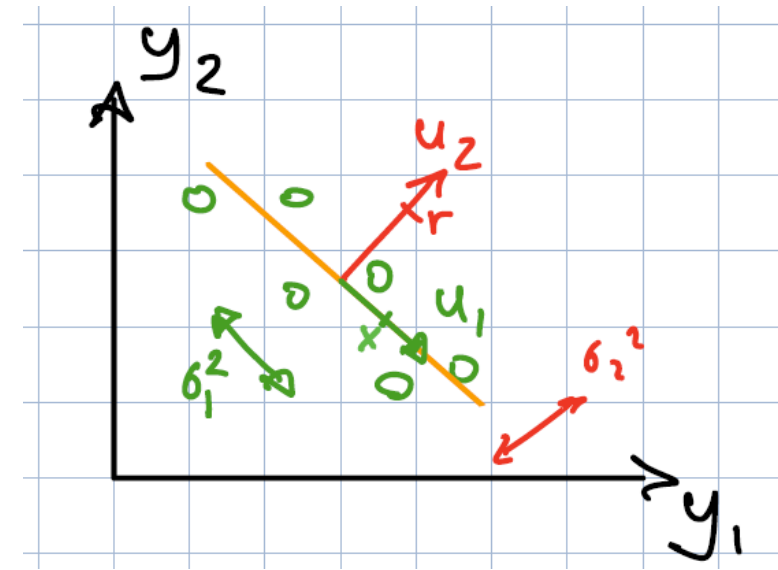
$$\mathbf{x}[k] := \mathbf{U}_{1:n}^\top \mathbf{z}[k] = \mathbf{U}_{1:n}^\top (\mathbf{y}[k] - \bar{\mathbf{y}}) \in \mathbb{R}^n \quad \mathbf{U}_{1:n}^\top := \begin{bmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p}$$

- For new data point, compute residual to check distance to subspace and alarm if large

$$\mathbf{r}[k] := \mathbf{U}_{n+1:p}^\top \mathbf{z}[k] = \mathbf{U}_{n+1:p}^\top (\mathbf{y}[k] - \bar{\mathbf{y}}) \in \mathbb{R}^{p-n}, \quad k > L \quad \mathbf{U}_{n+1:p}^\top := \begin{bmatrix} \mathbf{u}_{n+1}^\top \\ \vdots \\ \mathbf{u}_p^\top \end{bmatrix}$$

- Use CUSUM test to curtail number of false alarms

$$g[k] = (g[k-1] + \|\mathbf{r}[k]\|_2^2 - \delta)^+, \quad g[L] := 0$$



Autoencoders and Nonlinear PCA

- **Encoder:** $\mathbf{x}[k] = f_{\theta}(\mathbf{y}[k])$

where f_{θ} is (deep) Neural Network (NN) with many trained parameters θ

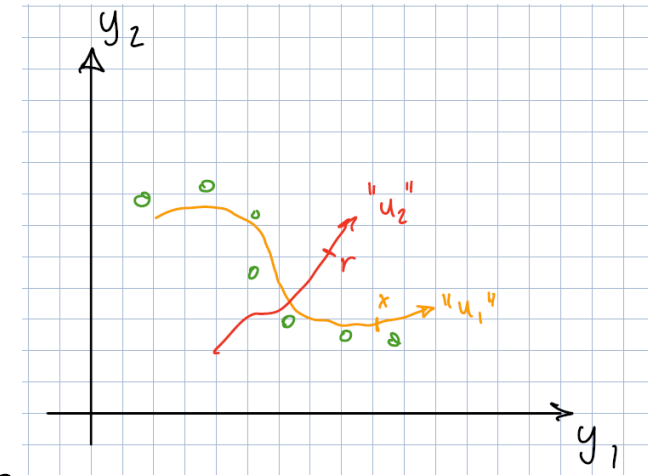
- **Decoder:** $\hat{\mathbf{y}}[k] = g_{\phi}(\mathbf{x}[k])$

where g_{ϕ} is (deep) NN with many trained parameters ϕ

- Parameters trained to minimize reconstruction loss

$$\min_{\theta, \phi} \frac{1}{N} \sum_{k=0}^L \text{loss}(\mathbf{y}[k], \hat{\mathbf{y}}[k]) \quad \text{loss}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$$

- For new data point, alarm if $\text{loss}(\mathbf{y}[k], \hat{\mathbf{y}}[k]) > \eta, k > L$, or use CUSUM test as earlier
- If maps f_{θ}, g_{ϕ} linear and loss quadratic: PCA is the optimal solution. *For nonlinear maps, few theoretical guarantees available*



Summary (Part II)

- DoS attack modeling and instability in control systems
- FDI attacks and detection methods
 - Linear algebra reveals which sensors are necessary for detecting specified attack/faults
 - *Dynamics helps* in estimating unknown state and attacks/faults!
 - Linear algebra provides constructive tools for reconstructing attacks/faults
 - Useful tools both for *risk management (offline)* and *operations (online)*
 - Four prototypical estimation/detection problems
- Noisy data: Use threshold tests. Stateful tests (CUSUM) provide (much) better design trade-offs than stateless tests (χ^2 -test, for example)
- Indirect vs. direct data-based methods: Powerful direct data-based *anomaly detection* methods (PCA, autoencoders, etc.) available. To *reconstruct* attacks/faults, some model knowledge generally necessary, and indirect methods are then preferable



Cyber-Physical Security in Energy Systems

Part II: Attack Modeling and Detection Methods

Henrik Sandberg (hsan@kth.se)

KTH EECS, Decision and Control Systems