# Data-Injection Attacks

Iñaki Esnaola, Samir M. Perlaza, Ke Sun

# Data-Injection Attacks

Iñaki Esnaola[1,3], Samir M. Perlaza[2,3], and Ke Sun[1]

1. Dept. of Automatic Control and Systems Eng., University of Sheffield, UK
2. INRIA, Centre de Recherche de Sophia Antipolis-Méditerranée, France
3. Department of Electrical Engineering, Princeton University, USA

email: esnaola@sheffield.ac.uk, samir.perlaza@inria.fr, ke.sun@sheffield.ac.uk

**Chapter 9** of: Advanced Data Analytics for Power Systems, A. Tajer, S. M. Perlaza and H. V. Poor, Eds., Cambridge University Press, Cambridge, UK, 2021, pp. 197-229

February 02, 2021

# Contents

# Chapter 9

# Data-Injection Attacks

## 9.1 Introduction

The pervasive deployment of sensing, monitoring, and data acquisition techniques in modern power systems enables the definition of functionalities and services that leverage accurate and real-time information about the system. This wealth of data supports network operators in the design of advanced control and management techniques that will inevitably change the operation of future power systems. An interesting side-effect of the data collection exercise that is starting to take place in power systems is that the unprecedented data analysis effort is shedding some light on the turbulent dynamics of power systems. While the underlying physical laws governing power systems are well understood, the large scale, distributed structure, and stochastic nature of the generation and consumption processes in the system results in a complex system. The large volumes of data about the state of the system are opening the door to modelling aspirations that were not feasible prior to the arrival of the smart grid paradigm.

The refinement of the models describing the power system operation will undoubtedly provide valuable insight to the network operator. However, that knowledge and the explanatory principles that it uncovers are also subject to be used in a malicious fashion. Access to statistics describing the state of the grid can inform malicious attackers by allowing them to pose the data-injection problem [1] problem within a probabilistic framework [2, 3]. By describing the processes taking place in the grid as a stochastic process, the network operator can incorporate the statistical description of the state variables in the state estimation procedure and pose it within a Bayesian estimation setting. Similarly, the attacker can exploit the stochastic description of the state variables by incorporating it to the attack construction in the form of prior knowledge about the state variables. Interestingly wether the network operator or the attacker benefit more from adding a stochastic description to the state variables does not have a simple answer and depends greatly on the parameters describing the power system.

In this chapter we review some of the basic attack constructions that exploit a stochastic description of the state variables. We pose the state estimation problem in a Bayesian setting

and cast the bad data detection procedure as a Bayesian hypothesis testing problem. This revised detection framework provides the benchmark for the attack detection problem that limits the achievable attack disruption. Indeed, the trade-off between the impact of the attack, in terms of disruption to the state estimator, and the probability of attack detection is analytically characterized within this Bayesian attack setting. We then generalize the attack construction by considering information-theoretic measures that place fundamental limits to a broad class of detection, estimation, and learning techniques. Because the attack constructions proposed in this chapter rely on the attacker having access to the statistical structure of the random process describing the state variables, we conclude by studying the impact of imperfect statistics on the attack performance. Specifically, we study the attack performance as a function of the size of the training data set that is available to the attacker to estimate the second-order statistics of the state variables.

## 9.2 System Model

### 9.2.1 Bayesian State Estimation

We model the state of the system as the vector of $n$ random variables $X^n$ taking values in $\mathbb{R}^n$ with distribution $P_{X^n}$. The random variable $X_i$ with $i = 1, 2, \ldots, n$, denotes the state variable $i$ of the power system, and therefore, each entry represents a different physical magnitude of the system that the network operator wishes to monitor. The prior knowledge that is available to the network operator is described by the probability distribution $P_{X^n}$. The knowledge of the distribution is a consequence of the modelling based on historical data acquired by the network operator. Assuming linearized system dynamics with $m$ measurements corrupted by additive white Gaussian noise (AWGN), the measurements are modelled as the vector of random variables $Y^m \in \mathbb{R}^m$ with distribution $P_{Y^m}$ given by

$$Y^n = \mathbf{H}X^n + Z^m, \tag{9.1}$$

where $\mathbf{H} \in \mathbb{R}^{m \times n}$ is the Jacobian of the linearized system dynamics around a given operating point and $Z^m \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is thermal white noise with power spectral density $\sigma^2$. While the operation point of the system induces a dynamic on the Jacobian matrix $\mathbf{H}$, in the following we assume that the time-scale over which the operation point changes is small compared to the time-scale at which the state estimator operates to produce the estimates. For that reason, in the following we assume that the Jacobian matrix is fixed and the only sources of uncertainty in the observation process originate from the stochasticity of the state variables and the additive noise corrupting the measurements.

The aim of the state estimator is to obtain an estimate $\hat{X}^n$ of the state vector $X^n$ from the system observations $Y^m$. In this chapter we adopt a linear estimation framework resulting in an estimate given by $\hat{X}^n = \mathbf{L}Y^m$, where $\mathbf{L} \in \mathbb{R}^{n \times m}$ is the linear estimation matrix determining the estimation procedure. In the case in which the operator knows the distribution $P_{X^n}$ of the underlying random process governing the state of the network, the estimation is performed by selecting the estimate that minimizes a given error cost function.

A common approach is to use the mean square error (MSE) as the error cost function. In this case, the network operator uses an estimator $\mathbf{M}$ that is the unique solution to the following optimization problem:

$$\mathbf{M} = \arg \min_{\mathbf{L} \in \mathbb{R}^{n \times m}} \mathbb{E} \left[ \frac{1}{n} \| X^n - \mathbf{L} Y^m \|_2^2 \right], \tag{9.2}$$

where the expectation is taken with respect to $X^n$ and $Z^m$.

Under the assumption that the network state vector $X^n$ follows an $n$-dimensional real Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}_{XX} \in \mathcal{S}_+^m$, i.e. $X^n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{XX})$, the minimum MSE (MMSE) estimate is given by

$$\hat{X}^n \triangleq \mathbb{E}[X^n | Y^m] = \mathbf{M} Y^m \tag{9.3}$$

where,

$$\mathbf{M} = \mathbf{\Sigma}_{XX} \mathbf{H}^\mathsf{T} (\mathbf{H} \mathbf{\Sigma}_{XX} \mathbf{H}^\mathsf{T} + \sigma^2 \mathbf{I})^{-1}. \tag{9.4}$$

### 9.2.2 Deterministic Attack Model

The aim of the attacker is to corrupt the estimate by altering the measurements. Data-injection attacks alter the measurements available to the operator by adding an attack vector to the measurements. The resulting observation model with the additive attack vector is given by

$$Y_a^m = \mathbf{H} X^m + Z^m + \mathbf{a}, \tag{9.5}$$

where $\mathbf{a}^m \in \mathbb{R}^m$ is the attack vector and $Y_a^m \in \mathbb{R}^m$ is the vector containing the compromised measurements [1]. Note that in this formulation, the attack vector does not have a probabilistic structure, i.e. the attack vector is deterministic. The random attack construction is considered later in the chapter.

The intention of the attacker can respond to diverse motivations, and therefore, attack construction strategy changes depending on the aim of the attacker. In this chapter, we study attacks that aim to maximize the monitoring disruption, i.e. attacks that obstruct the state estimation procedure with the aim of deviating the estimate as much as possible from the true state. In that sense, the attack problem is bound to the cost function used by the state estimator to obtain the estimate, as the attacker aims to maximize it while the estimator aims to minimize it. In the MMSE setting described in the preceding text, it follows that the the impact of the attack vector is obtained by noticing that the estimate when the attack vector is present is given by

$$\hat{X}_a^n = \mathbf{M}(\mathbf{H} X^n + Z^m) + \mathbf{M}\mathbf{a}. \tag{9.6}$$

The term $\mathbf{M}\mathbf{a}$ is referred to as the *Bayesian injection vector* introduced by the attack vector $\mathbf{a}$ and is denoted by

$$\mathbf{c} \triangleq \mathbf{M}\mathbf{a} = \mathbf{\Sigma}_{XX} \mathbf{H}^\mathsf{T} (\mathbf{H} \mathbf{\Sigma}_{XX} \mathbf{H}^\mathsf{T} + \sigma^2 \mathbf{I})^{-1} \mathbf{a}. \tag{9.7}$$

The *Bayesian injection vector* is a deterministic vector that corrupts the MMSE estimate of the operator resulting in

$$\hat{X}_a^n = \hat{X}^n + \mathbf{c}. \tag{9.8}$$

where $\hat{X}^n$ is given in (9.3).

### 9.2.3 Attack Detection

As a part of the grid management, a network operator systematically attempts to identify measurements that are not deemed of sufficient quality for the state estimator. In practice, this operation can be cast as a hypothesis testing problem with hypotheses

$$\begin{aligned} \mathcal{H}_0: &\quad \text{There is no attack,} \quad \text{versus} \\ \mathcal{H}_1: &\quad \text{Measurements are compromised.} \end{aligned} \tag{9.9}$$

Assuming the operator knows the distribution of the state variables, $P_{X^n}$, and the observation model (9.5), then it can obtain the joint distribution of the measurements and the state variables for both normal operation conditions and the case when an attack is present, i.e. $P_{X^nY^m}$ and $P_{X^nY_a^m}$, respectively.

Under the assumption that the state variables follow a multivariate Gaussian distribution $X^n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{XX})$ it follows that the vector of measurements $Y^n$ follows an $m$-dimensional real Gaussian random distribution with covariance matrix

$$\mathbf{\Sigma}_{YY} = \mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^\mathsf{T} + \sigma^2\mathbf{I}, \tag{9.10}$$

and mean $\mathbf{a}$ when there is an attack; or zero mean when there is no attack. Within this setting, the hypothesis testing problem described before is adapted to the attack detection problem by comparing the following hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad Y^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{YY}), \quad \text{versus} \\ \mathcal{H}_1: &\quad Y^m \sim \mathcal{N}(\mathbf{a}, \mathbf{\Sigma}_{YY}). \end{aligned} \tag{9.11}$$

A worst case scenario approach is assumed for the attackers, namely, the operator knows the attack vector, $\mathbf{a}$, used in the attack. However, the operator does not know a priori whether the grid is under attack or not, which accounts for the need of an attack detection strategy. That being the case, the optimal detection strategy for the operator is to perform a likelihood ratio test (LRT) $L(\mathbf{y}, \mathbf{a})$ with respect to the observations $\mathbf{y}$. Under the assumption that state variables follow a multivariate Gaussian distribution, the likelihood ratio can be calculated as

$$L(\mathbf{y}, \mathbf{a}) = \frac{f_{\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{yy}})}(\mathbf{y})}{f_{\mathcal{N}(\mathbf{a}, \mathbf{\Sigma}_{\mathbf{yy}})}(\mathbf{y})} = \exp\left(\frac{1}{2}\mathbf{a}^\mathsf{T}\mathbf{\Sigma}_{YY}^{-1}\mathbf{a} - \mathbf{a}^\mathsf{T}\mathbf{\Sigma}_{YY}^{-1}\mathbf{y}\right), \tag{9.12}$$

where $f_{\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})}$ is the probability density function of a multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Therefore, either hypothesis is accepted by evaluating the inequalities

$$L(\mathbf{y}, \mathbf{a}) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau, \tag{9.13}$$

where $\tau \in [0, \infty)$ is tuned to set the trade-off between the probability of detection and the probability of false alarm.

## 9.3 Centralized Deterministic Attacks

This section describes the construction of data-injection attacks in the case in which there is a unique attacker with access to all the measurements on the power system. This scenario is referred to as *centralized attacks* in order to highlight that there exists a unique entity deciding the data-injection vector $\mathbf{a} \in \mathbb{R}^m$ in (9.5). The difference between the scenario in which there exists a unique attacker or several (competing or cooperating) attackers is subtle and it is treated in Section 9.4.

Let $\mathcal{M} = \{1, \ldots, m\}$ denote the set of all $m$ sensors available to the network operator. A sensor is said to be compromised if the attacker is able to arbitrarily modify its output. Given a total energy budget $E > 0$ at the attacker, the set of all possible attacks that can be injected to the network can be explicitly described:

$$\mathcal{A} = \left\{ \mathbf{a} \in \mathbb{R}^m : \mathbf{a}^{\mathsf{T}} \mathbf{a} \leqslant E \right\}. \tag{9.14}$$

### 9.3.1 Attacks with Minimum Probability of Detection

The attacker chooses a vector $\mathbf{a} \in \mathcal{A}$ taking into account the trade-off between the probability of being detected and the distortion induced by the Bayesian injection vector given by (9.7). However, the choice of a particular data-injection vector is not trivial as the attacker does not have any information about the exact realizations of the vector of state variables $\mathbf{x}$ and the noise vector $\mathbf{z}$. A reasonable assumption on the knowledge of the attacker is to consider that it knows the structure of the power system and thus, it knows the matrix $\mathbf{H}$. It is also reasonable to assume that it knows the first and second moments of the state variables $X^n$ and noise $Z^m$ as this can be computed from historical data.

Under these knowledge assumptions, the probability that the network operator is unable to detect the attack vector $\mathbf{a}$ is

$$\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \mathsf{E}\left[ \mathbb{1}_{\{\mathsf{L}(\mathbf{y},\mathbf{a}) > \tau\}} \right], \tag{9.15}$$

where the expectation is taken over the joint probability distribution of state variables $X^n$ and the AWGN noise vector $Z^n$, and $\mathbb{1}_{\{.\}}$ denotes the indicator function. Note that under these assumptions, $Y^m$ is a random variable with Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{\Sigma}_{YY}$. Thus, the probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ of a vector $\mathbf{a}$ being a successful attack, i.e., a non-detected attack is given by [4]

$$\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \frac{1}{2} \mathsf{erfc}\left( \frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathsf{YY}}^{-1} \mathbf{a} + \log \tau}{\sqrt{2 \mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathsf{YY}}^{-1} \mathbf{a}}} \right). \tag{9.16}$$

Often, the knowledge of the threshold $\tau$ in (9.13) is not available to the attacker and thus, it cannot determine the exact probability of not being detected for a given attack

vector $\mathbf{a}$. However, the knowledge of whether $\tau > 1$ or $\tau \leqslant 1$ induces different behaviors on the attacker. The following propositions follow immediately from (9.16) and the properties of the complementary error function.

**Proposition 9.1** (Case $\tau \leqslant 1$)**.** *Let $\tau \leqslant 1$. Then, for all $\mathbf{a} \in \mathcal{A}$, $\mathsf{P_{ND}}(\mathbf{a}) < \mathsf{P_{ND}}((0, \ldots, 0))$ and the probability $\mathsf{P_{ND}}(\mathbf{a})$ is monotonically decreasing with $\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a}$.*

**Proposition 9.2** (Case $\tau > 1$)**.** *Let $\tau > 1$ and let also $\boldsymbol{\Sigma}_{YY} = \mathbf{U}_{YY}\boldsymbol{\Lambda}_{YY}\mathbf{U}_{YY}^\mathsf{T}$ be the singular value decomposition of $\boldsymbol{\Sigma}_{YY}$, with $\mathbf{U}_{YY}^\mathsf{T} = (\mathbf{u}_{YY,1}, \ldots, \mathbf{u}_{YY,m})$ and $\boldsymbol{\Lambda}_{YY} = diag\,(\lambda_{YY,1}, \ldots, \lambda_{YY,m})$ and $\lambda_{YY,1} \geqslant \lambda_{YY,2} \geqslant \ldots, \geqslant \lambda_{YY,m}$. Then, any vector of the form*

$$\mathbf{a} = \pm\sqrt{\lambda_{YY,k} 2\log\tau}\,\mathbf{u}_{YY,k}, \tag{9.17}$$

*with $k \in \{1, \ldots, m\}$, is a data-injection attack that satisfies for all $\mathbf{a}' \in \mathbb{R}^m$, $\mathsf{P_{ND}}(\mathbf{a}') \leqslant \mathsf{P_{ND}}(\mathbf{a})$.*

The proof of Proposition 9.1 and Proposition 9.2 follows.

*Proof.* Let $x = \mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a}$ and note that $x > 0$ due to the positive definiteness of $\boldsymbol{\Sigma}_{YY}$. Let also the function $g : \mathbb{R} \to \mathbb{R}$ be

$$g(x) = \frac{\frac{1}{2}x + \log\tau}{\sqrt{2x}}. \tag{9.18}$$

The first derivative of $g(x)$ is

$$g'(x) = \frac{1}{2\sqrt{2x}}\left(\frac{1}{2} - \frac{\log\tau}{x}\right). \tag{9.19}$$

Note that in the case in which $\log\tau \leqslant 0$ (or $\tau \leqslant 1$), then for all $x \in \mathbb{R}^+$, $g'(x) > 0$ and thus, $g$ is monotonically increasing with $x$. Since the complementary error function erfc is monotonically decreasing with its argument, the statement of Proposition 9.1 follows and completes its proof. In the case in which $\log\tau \geqslant 0$ (or $\tau > 1$), the solution to $g'(x) = 0$ is $x = 2\log\tau$ and it corresponds to a minimum of the function $g$. The maximum of $\frac{1}{2}\mathrm{erfc}(g(x))$ occurs at the minimum of $g(x)$ given that erfc is monotonically decreasing with its argument. Hence, the maximum of $\mathsf{P_{ND}}(\mathbf{a})$ occurs for the attack vectors satisfying:

$$\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} = 2\log\tau. \tag{9.20}$$

Solving for $\mathbf{a}$ in (9.20) yields (9.17) and this completes the proof of Proposition 9.2. □

The relevance of Proposition 9.1 is that it states that when $\tau \leqslant 1$, any non-zero data-injection attack vector possesses a non zero probability of being detected. Indeed, the highest probability $\mathsf{P_{ND}}(\mathbf{a})$ of not being detected is guaranteed by the null vector $\mathbf{a}$= (0, ..., 0), i.e., there is no attack. Alternatively, when $\tau > 1$ it follows from Proposition 9.2 that there always exists a non-zero vector that possesses maximum probability of not being detected.

However, in both cases, it is clear that the corresponding data-injection vectors which induce the highest probability of not being detected are not necessarily the same that inflige the largest damage to the network, i.e., maximize the excess distortion.

From this point of view, the attacker faces the trade-off between maximizing the excess distortion and minimizing the probability of being detected. Thus, the attack construction can be formulated as an optimization problem in which the solution $\mathbf{a}$ is a data-injection vector that maximizes the probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ of not being detected at the same time that it induces a distortion $\|\mathbf{c}\|_2^2 \geqslant \mathsf{D}_0$ into the estimate. In the case in which $\tau \leqslant 1$, it follows from Proposition 9.1 and (9.7) that this problem can be formulated as the following optimization problem:

$$\min_{\mathbf{a}\in\mathcal{A}} \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} \quad \text{s.t.} \quad \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} \geq \mathsf{D}_0. \tag{9.21}$$

The solution to the optimization problem in (9.21) is given by the following theorem.

**Theorem 9.1.** *Let* $\mathbf{G} = \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{G} = \mathbf{U}_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$, *with* $\mathbf{U} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},m})$ *a unitary matrix and* $\boldsymbol{\Sigma}_{\mathbf{G}} = diag\,(\lambda_{\mathbf{G},1}, \ldots, \lambda_{\mathbf{G},m})$ *a diagonal matrix with* $\lambda_{\mathbf{G},1} \geqslant \ldots \geqslant \lambda_{\mathbf{G},m}$. *Then, if* $\tau \leqslant 1$, *the attack vector* $\mathbf{a}$ *that maximizes the probability of not being detected* $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ *while inducing an excess distortion not less than* $\mathsf{D}_0$ *is*

$$\mathbf{a} \;=\; \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}\boldsymbol{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}. \tag{9.22}$$

*Moreover,* $\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \frac{1}{2}\mathrm{erfc}\left(\frac{\frac{\mathsf{D}_0}{2\lambda_{\mathbf{G},1}} + \log\tau}{\sqrt{\frac{2\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}}\right)$.

*Proof.* Consider the Lagrangian

$$L(\mathbf{a}) = \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} - \gamma\left(\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} - \mathsf{D}_0\right), \tag{9.23}$$

with $\gamma > 0$ a Lagrangian multiplier. Then, the necessary conditions for $\mathbf{a}$ to be a solution to the optimization problem (9.21) are:

$$\nabla_{\mathbf{a}} L(\mathbf{a}) = 2\left(\boldsymbol{\Sigma}_{YY}^{-1} - \gamma\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\right)\mathbf{a} = 0 \tag{9.24}$$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}L(\mathbf{a}) = \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} - \mathsf{D}_0 = 0. \tag{9.25}$$

Note that any

$$\mathbf{a}_i = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\boldsymbol{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} \text{ and} \tag{9.26}$$

$$\gamma_i = \lambda_{\mathbf{G},i}, \text{ with } 1 \leqslant i \leqslant \mathrm{rank}\,(\mathbf{G}), \tag{9.27}$$

satisfy $\gamma_i > 0$ and conditions (9.24) and (9.25). Hence, the set of vectors that satisfy the necessary conditions to be a solution of (9.21) is

$$\left\{ \mathbf{a}_i = \pm \sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}} \boldsymbol{\Sigma}_{YY}^{\frac{1}{2}} \mathbf{u}_{\mathbf{G},i} : 1 \leqslant i \leqslant \operatorname{rank}(\mathbf{G}) \right\}. \tag{9.28}$$

More importantly, any vector $\mathbf{a} \neq \mathbf{a}_i$, with $1 \leqslant i \leqslant \operatorname{rank}(\mathbf{G})$, does not satisfy the necessary conditions. Moreover,

$$\mathbf{a}_i^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a}_i = \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}} \geqslant \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}. \tag{9.29}$$

Therefore, $\mathbf{a} = \pm \sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}} \boldsymbol{\Sigma}_{YY}^{\frac{1}{2}} \mathbf{u}_{\mathbf{G},1}$ are the unique solutions to (9.21). This completes the proof. □

Interestingly, the construction of the data-injection attack $\mathbf{a}$ in (9.22) does not require the exact knowledge of $\tau$. That is, only knowing that $\tau \leqslant 1$ is enough to build the data-injection attack that has the highest probability of not being detected and induces a distortion of at least $\mathsf{D}_0$.

In the case in which $\tau > 1$, it is also possible to find the data-injection attack vector that induces a distortion not less than $\mathsf{D}_0$ and the maximum probability of not being detected. Such a vector is the solution to the following optimization problem.

$$\min_{\mathbf{a} \in \mathcal{A}} \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a} + \log \tau}{\sqrt{2 \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a}}} \text{ s.t. } \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{XX}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a} \geq \mathsf{D}_0. \tag{9.30}$$

The solution to the optimization problem in (9.30) is given by the following theorem.

**Theorem 9.2.** *Let* $\mathbf{G} = \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{H} \boldsymbol{\Sigma}_{XX}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{G} = \mathbf{U}_{\mathbf{G}} \boldsymbol{\Sigma}_{\mathbf{G}} \mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$, *with* $\mathbf{U}_{\mathbf{G}} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},m})$ *a unitary matrix and* $\boldsymbol{\Sigma}_{\mathbf{G}} = \operatorname{diag}(\lambda_{\mathbf{G},1}, \ldots, \lambda_{\mathbf{G},m})$ *a diagonal matrix with* $\lambda_{\mathbf{G},1} \geqslant \ldots \geqslant \lambda_{\mathbf{G},m}$. *Then, when* $\tau > 1$, *the attack vector* $\mathbf{a}$ *that maximizes the probability of not being detected* $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ *while producing an excess distortion not less than* $\mathsf{D}_0$ *is*

$$\mathbf{a} = \begin{cases} \pm \sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k^*}}} \boldsymbol{\Sigma}_{YY}^{\frac{1}{2}} \mathbf{u}_{\mathbf{G},k^*} & \text{if } \frac{\mathsf{D}_0}{2 \log \tau \lambda_{\mathbf{G},\operatorname{rank}\mathbf{G}}} \geqslant 1, \\ \pm \sqrt{2 \log \tau} \boldsymbol{\Sigma}_{YY}^{\frac{1}{2}} \mathbf{u}_{\mathbf{G},1} & \text{if } \frac{\mathsf{D}_0}{2 \log \tau \lambda_{\mathbf{G},\operatorname{rank}\mathbf{G}}} < 1 \end{cases}$$

*with*

$$k^* = \arg \min_{k \in \{1,\ldots,\operatorname{rank}\mathbf{G}\} : \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k}} > 2\log(\tau)} \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k}}. \tag{9.31}$$

*Proof.* The structure of the proof of Theorem 9.2 is similar to the proof of Theorem 9.1 and is omitted in this chapter. A complete proof can be found in [5]. □

### 9.3.2 Attacks with Maximum Distortion

In the previous subsection, the attacker constructs its data-injection vector $\mathbf{a}$ aiming to maximize the probability of non-detection $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ while guaranteeing a minimum distortion. However, this problem has a dual in which the objective is to maximize the distortion $\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a}$ while guaranteeing that the probability of not being detected remains always larger than a given threshold $L_0' \in [0, \frac{1}{2}]$. This problem can be formulated as the following optimization problem:

$$\max_{\mathbf{a}\in\mathcal{A}} \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} \;\text{ s.t. }\; \frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-1}\mathbf{a}}} \leq L_0, \tag{9.32}$$

with $L_0 = \operatorname{erfc}^{-1}(2L_0') \in [0, \infty)$.

The solution to the optimization problem in (9.32) is given by the following theorem.

**Theorem 9.3.** *Let the matrix* $\mathbf{G} = \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}_{XX}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{U}_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$, *with* $\mathbf{U} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},m})$ *a unitary matrix and* $\boldsymbol{\Sigma}_{\mathbf{G}} = diag(\lambda_{\mathbf{G},1}, \ldots, \lambda_{\mathbf{G},m})$ *a diagonal matrix with* $\lambda_{\mathbf{G},1} \geqslant \ldots \geqslant \lambda_{\mathbf{G},m}$. *Then, the attack vector* $\mathbf{a}$ *that maximizes the excess distortion* $\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{a}$ *with a probability of not being detected that does not go below* $L_0 \in [0, \frac{1}{2}]$ *is*

$$\mathbf{a} = \pm\left(\sqrt{2}L_0 + \sqrt{2L_0^2 - 2\log\tau}\right)\boldsymbol{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}, \tag{9.33}$$

*when a solution exists.*

*Proof.* The structure of the proof of Theorem 9.3 is similar to the proof of Theorem 9.1 and is omitted in this chapter. A complete proof can be found in [5]. □

## 9.4 Decentralized Deterministic Attacks

Let $\mathcal{K} = \{1, \ldots, K\}$ be the set of attackers that can potentially perform a data injection attack on the network, e.g., a decentralized vector attack. Let also $\mathcal{C}_k \in \{1, 2, \ldots, m\}$ be the set of sensors that attacker $k \in \mathcal{K}$ can control. Assume that $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are proper sets and form a partition of the set $\mathcal{M}$ of all sensors. The set $\mathcal{A}_k$ of data attack vectors $\mathbf{a}_k = (a_{k,1}, a_{k,2}, \ldots, a_{k,m})$ that can be injected into the network by attacker $k \in \mathcal{K}$ is of the form

$$\mathcal{A}_k = \{\mathbf{a}_k \in \mathbb{R}^m : \mathbf{a}_{k,j} = 0 \text{ for all } j \notin \mathcal{C}_k, \mathbf{a}_k^{\mathsf{T}}\mathbf{a}_k \leq E_k\}. \tag{9.34}$$

The constant $E_k < \infty$ represents the energy budget of attacker $k$. Let the set of all possible sums of the elements of $\mathcal{A}_i$ and $\mathcal{A}_j$ be denoted by $\mathcal{A}_i \oplus \mathcal{A}_j$. That is, for all $\mathbf{a} \in \mathcal{A}_i \oplus \mathcal{A}_j$, there exists a pair of vectors $(\mathbf{a}_i, \mathbf{a}_j) \in \mathcal{A}_i \times \mathcal{A}_j$ such that $\mathbf{a} = \mathbf{a}_i + \mathbf{a}_j$. Using this notation, let the set of all possible data-injection attacks be denoted by

$$\mathcal{A} = \mathcal{A}_1 \oplus \mathcal{A}_2 \oplus \ldots \oplus \mathcal{A}_K, \tag{9.35}$$

and the set of complementary data-injection attacks with respect to attacker $k$ be denoted by

$$\mathcal{A}_{-k} = \mathcal{A}_1 \oplus \ldots \oplus \mathcal{A}_{k-1} \oplus \mathcal{A}_{k+1} \oplus \ldots \oplus \mathcal{A}_K. \tag{9.36}$$

Given the individual data injection vectors $\mathbf{a}_i \in \mathcal{A}_i$, with $i \in \{1, \ldots, K\}$, the global attack vector $\mathbf{a}$ is

$$\mathbf{a} = \sum_{i=1}^{K} \mathbf{a}_k \in \mathcal{A}. \tag{9.37}$$

The aim of attacker $k$ is to corrupt the measurements obtained by the set of meters $\mathcal{C}_k$ by injecting an error vector $\mathbf{a}_k \in \mathcal{A}_k$ that maximizes the damage to the network, e.g., the excess distortion, while avoiding the detection of the global data-injection vector $\mathbf{a}$. Clearly, all attackers have the same interest but they control different sets of measurements, i.e., $\mathcal{C}_i \neq \mathcal{C}_k$, for a any pair $(i,k) \in \mathcal{K}^2$. For modeling this behavior, attackers use the utility function $\phi : \mathbb{R}^m \to \mathbb{R}$, to determine whether a data-injection vector $\mathbf{a}_k \in \mathcal{A}_k$ is more beneficial than another $\mathbf{a}'_k \in \mathcal{A}_k$ given the complementary attack vector

$$\mathbf{a}_{-k} = \sum_{i \in \{1,\ldots,K\} \setminus \{k\}} \mathbf{a}_i \in \mathcal{A}_{-k} \tag{9.38}$$

adopted by all the other attackers. The function $\phi$ is chosen considering the fact that an attack is said to be successful if it induces a non-zero distortion and it is not detected. Alternatively, if the attack is detected no damage is induced into the network as the operator discards the measurements and no estimation is performed. Hence, given a global attack $\mathbf{a}$, the distortion induced into the measurements is $\mathbb{1}_{\{L(Y_a^m, \mathbf{a}) > \tau\}} \mathbf{x}_a^{\mathsf{T}} \mathbf{x}_a$. However, attackers are not able to know the exact state of the network $\mathbf{x}$ and the realization of the noise $\mathbf{z}$ before launching the attack. Thus, it appears natural to exploit the knowledge of the first and second moments of both the state variables $\mathbf{x}$ and noise $\mathbf{z}$ and consider as a metric the expected distortion $\phi(\mathbf{a})$ that can be induced by the attack vector $\mathbf{a}$:

$$\phi(\mathbf{a}) = \mathsf{E}\left[\left(\mathbb{1}_{\{\mathsf{L}(\mathsf{Y}_{\mathsf{a}}^{\mathsf{m}}, \mathbf{a}) > \tau\}}\right) \mathbf{c}^{\mathsf{T}} \mathbf{c}\right], \tag{9.39}$$

$$= \mathsf{P}_{\mathsf{ND}}(\mathbf{a})\, \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathsf{YY}}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\mathsf{XX}}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathsf{YY}}^{-1} \mathbf{a}, \tag{9.40}$$

where $\mathbf{c}$ is in (9.7) and the expectation is taken over the distribution of state variables $X^n$ and the noise $Z^m$. Note that under this assumptions of global knowledge, this model considers the worst case scenario for the network operator. Indeed, the result presented in this section corresponds to the case in which the attackers inflict the most harm onto the state estimator.

### 9.4.1 Game Formulation

The benefit $\phi(\mathbf{a})$ obtained by attacker $k$ does not only depend on its own data-injection vector $\mathbf{a}_k$, but also on the data-injection vectors $\mathbf{a}_{-k}$ of all the other attackers. This becomes clear from the construction of the global data-injection vector $\mathbf{a}$ in (9.37), the excess

distortion $\mathbf{x}_a$ in (9.7) and the probability of not being detected $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ in (9.16). Therefore, the interaction of all attackers in the network can be described by a game in normal form

$$\mathcal{G} = \left(\mathcal{K}, \{\mathcal{A}_k\}_{k\in\mathcal{K}}, \phi\right). \tag{9.41}$$

Each attacker is a player in the game $\mathcal{G}$ and it is identified by an index from the set $\mathcal{K}$. The actions player $k$ might adopt are data-injection vectors $\mathbf{a}_k$ in the set $\mathcal{A}_k$ in (9.34). The underlying assumption in the following of this section is that, given a vector of data-injection attacks $\mathbf{a}_{-k}$, player $k$ aims to adopt a data-injection vector $\mathbf{a}_k$ such that the expected excess distortion $\phi(\mathbf{a}_k + \mathbf{a}_{-k})$ is maximized. That is,

$$\mathbf{a}_k \in \mathrm{BR}_k\left(\mathbf{a}_{-k}\right), \tag{9.42}$$

where the correspondence $\mathrm{BR}_k : \mathcal{A}_{-k} \to 2^{\mathcal{A}_k}$ is the best response correspondence, i.e.,

$$\mathrm{BR}_k\left(\mathbf{a}_{-k}\right) = \arg\max_{\mathbf{a}_k\in\mathcal{A}_k} \phi\left(\mathbf{a}_k + \mathbf{a}_{-k}\right). \tag{9.43}$$

The notation $2^{\mathcal{A}_k}$ represents the set of all possible subsets of $\mathcal{A}_k$. Note that $\mathrm{BR}_k\left(\mathbf{a}_{-k}\right) \subseteq \mathcal{A}_k$ is the set of data-injection attack vectors that are optimal given that the other attackers have adopted the data-injection vector $\mathbf{a}_{-k}$. In this setting, each attacker tampers with a subset $\mathcal{C}_k$ of all sensors $\mathcal{C} = \{1, 2, \ldots, m\}$, as opposed to the centralized case in which there exists a single attacker that is able to tampers with all sensors in $\mathcal{C}$.

A game solution that is particularly relevant for this analysis is the NE [6].

**Definition 9.1** (Nash Equilibrium)**.** *The data-injection vector* $\mathbf{a}$ *is an NE of the game* $\mathcal{G}$ *if and only if it is a solution of the fix point equation*

$$\mathbf{a} = \mathrm{BR}\left(\mathbf{a}\right), \tag{9.44}$$

*with* $\mathrm{BR} : \mathcal{A} \to 2^{\mathcal{A}}$ *being the global best-response correspondence, i.e.,*

$$\mathrm{BR}\left(\mathbf{a}\right) = \mathrm{BR}_1\left(\mathbf{a}_{-1}\right) + \ldots + \mathrm{BR}_K\left(\mathbf{a}_{-K}\right). \tag{9.45}$$

Essentially, at an NE, attackers obtain the maximum benefit given the data-injection vector adopted by all the other attackers. This implies that an NE is an operating point at which attackers achieve the highest expected distortion induced over the measurements. More importantly, any unilateral deviation from an equilibrium data-injection vector $\mathbf{a}$ does not lead to an improvement of the average excess distortion. Note that this formulation does not say anything about the exact distortion induced by an attack but the average distortion. This is mainly because the attack is chosen under the uncertainty of the state vector $X^n$ and the noise term $Z^m$.

The following proposition highlights an important property of the game $\mathcal{G}$ in (9.41).

**Proposition 9.3.** *The game* $\mathcal{G}$ *in* (9.41) *is a potential game.*

*Proof.* The proof follows immediately from the observation that all the players have the same utility function $\phi$ [7]. Thus, the function $\phi$ is a potential of the game $\mathcal{G}$ in (9.41) and any maximum of the potential function is an NE of the game $\mathcal{G}$. □

In general, potential games [7] possess numerous properties that are inherited by the game $\mathcal{G}$ in (9.41). These properties are detailed by the following propositions

**Proposition 9.4.** *The game $\mathcal{G}$ possesses at least one NE.*

*Proof.* Note that $\phi$ is continuous in $\mathcal{A}$ and $\mathcal{A}$ is a convex and closed set; therefore, there always exists a maximum of the potential function $\phi$ in $\mathcal{A}$. Finally from Lemma 4.3 in [7], it follows that such a maximum corresponds to an NE. □

### 9.4.2 Achievability of an NE

The attackers are said to play a sequential best response dynamic (BRD) if the attackers can sequentially decide their own data-injection vector $\mathbf{a}_k$ from their sets of best responses following a round-robin (increasing) order. Denote by $\mathbf{a}_k^{(t)} \in \mathcal{A}$ the choice of attacker $k$ during round $t \in \mathbb{N}$ and assume that attackers are able to observe all the other attackers' data-injection vectors. Under these assumptions, the BRD can be defined as follows.

**Definition 9.2** (Best Response Dynamics)**.** *The players of the game $\mathcal{G}$ are said to play best response dynamics if there exists a round-robin order of the elements of $\mathcal{K}$ in which at each round $t \in \mathbb{N}$, the following holds:*

$$\mathbf{a}_k^{(t)} \in \mathrm{BR}_k \left( \mathbf{a}_1^{(t)} + \ldots + \mathbf{a}_{k-1}^{(t)} + \mathbf{a}_{k+1}^{(t-1)} + \ldots + \mathbf{a}_K^{(t-1)} \right). \tag{9.46}$$

From the properties of potential games (Lemma 4.2 in [7]), the following proposition follows.

**Lemma 9.1** (Achievability of NE attacks)**.** *Any BRD in the game $\mathcal{G}$ converges to a data-injection attack vector that is an NE.*

The relevance of Lemma 9.1 is that it establishes that if attackers can communicate in at least a round-robin fashion, they are always able to attack the network with a data-injection vector that maximizes the average excess distortion. Note that there might exists several NEs (local maxima of $\phi$) and there is no guarantee that attackers will converge to the best NE, i.e., a global maximum of $\phi$. It is important to note that under the assumption that there exists a unique maximum, which is not the case for the game $\mathcal{G}$ (see Theorem 9.4), all attackers are able to calculate such a global maximum and no communications is required among the attackers. Nonetheless, the game $\mathcal{G}$ always possesses at least two NEs, which enforces the use of a sequential BRD to converge to an NE.

### 9.4.3 Cardinality of the set of NEs

Let $\mathcal{A}_{\text{NE}}$ be the set of all data-injection attacks that form NEs. The following theorem bounds the number of NEs in the game.

**Theorem 9.4.** *The cardinality of the set $\mathcal{A}_{\text{NE}}$ of NE of the game $\mathcal{G}$ satisfies*

$$2 \leqslant |\mathcal{A}_{\text{NE}}| \leqslant C \cdot \text{rank}(\mathbf{H}) \tag{9.47}$$

*where $C < \infty$ is a constant that depends on $\tau$.*

*Proof.* The lower bound follows from the symmetry of the utility function given in (9.39), i.e. $\phi(\mathbf{a}) = \phi(-\mathbf{a})$, and the existence of at least one NE claimed in Proposition 9.4.

To prove the upper bound the number of stationary points of the utility function is evaluated. This is equivalent to the cardinality of the set

$$\mathcal{S} = \{\mathbf{a} \in \mathbb{R}^m : \nabla_{\mathbf{a}} \phi(\mathbf{a}) = \mathbf{0}\}, \tag{9.48}$$

which satisfies $\mathcal{A}_{NE} \subseteq \mathcal{S}$. Calculating the gradient with respect to the attack vector yields

$$\nabla_{\mathbf{a}} \phi(\mathbf{a}) = \left(\alpha(\mathbf{a}) \mathbf{M}^{\mathsf{T}} \mathbf{M} - \beta(\mathbf{a}) \boldsymbol{\Sigma}_{YY}^{-1}\right) \mathbf{a}, \tag{9.49}$$

where

$$\alpha(\mathbf{a}) \triangleq \text{erfc}\left( \frac{1}{\sqrt{2}} \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a} + \log \tau}{\left(\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a}\right)^{\frac{1}{2}}} \right) \tag{9.50}$$

and

$$\beta(\mathbf{a}) \triangleq \frac{\mathbf{a}^{\mathsf{T}} \mathbf{M}^{\mathsf{T}} \mathbf{M} \mathbf{a}}{\sqrt{2\pi} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a}} \left( \frac{1}{2} - \frac{\log \tau}{\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a}} \right) \exp\left( - \left( \frac{1}{\sqrt{2}} \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a} + \log \tau}{\left(\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} \mathbf{a}\right)^{\frac{1}{2}}} \right)^2 \right). \tag{9.51}$$

Define $\delta(\mathbf{a}) \triangleq \frac{\beta(\mathbf{a})}{\alpha(\mathbf{a})}$ and note that combining (9.4) with (9.49) gives the following condition for the stationary points:

$$\left(\mathbf{H} \boldsymbol{\Sigma}_{XX}^2 \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1} - \delta(\mathbf{a}) \mathbf{I}\right) \mathbf{a} = \mathbf{0}. \tag{9.52}$$

Note that the number of linearly independent attack vectors that are a solution of the linear system in (9.52) is given by

$$R \triangleq \text{rank}\left(\mathbf{H} \boldsymbol{\Sigma}_{XX}^2 \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-1}\right) \tag{9.53}$$

$$= \text{rank}\left(\mathbf{H}\right). \tag{9.54}$$

where (9.54) follows from the fact that $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$ are positive definite. Define the eigenvalue decomposition

$$\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{H} \boldsymbol{\Sigma}_{XX}^2 \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^{\mathsf{T}} \tag{9.55}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the ordered eigenvalues $\{\lambda_i\}_{i=1}^m$ matching the order of of the eigenvectors in $\mathbf{U}$. As a result of (9.53) there are $r$ eigenvalues, $\lambda_k$, which are different from zero and $m-r$ diagonal elements of $\mathbf{\Lambda}$ which are zero. Combining this decomposition with some algebraic manipulation, the condition for stationary points in (9.52) can be recast as

$$\mathbf{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{U}\left(\mathbf{\Lambda}-\delta(\mathbf{a})\mathbf{I}\right)\mathbf{U}^{\mathsf{T}}\mathbf{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{a}=\mathbf{0}. \tag{9.56}$$

Let $w\in\mathbb{R}$ be a scaling parameter and observe that the attack vectors that satisfy $\mathbf{a}=w\mathbf{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{U}\mathbf{e}_k$ and $\delta(\mathbf{a})=\lambda_k$ for $k=1,\dots,r$ are solutions of (9.56). Note that the critical points associated to zero eigenvalues are not NE. Indeed, the eigenvectors associated to zero eigenvalues yield zero utility. Since the utility function is strictly positive, these critical points are minima of the utility function and can be discarded when counting the number of NE. Therefore, the set in (9.48) can be rewritten based on the condition in (9.56) as

$$\mathcal{S}=\bigcup_{k=1}^{R}\mathcal{S}_k, \tag{9.57}$$

where

$$\mathcal{S}_k=\{\mathbf{a}\in\mathbb{R}^m:\mathbf{a}=w\mathbf{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{U}\mathbf{e}_k \text{ and } \delta(\mathbf{a})=\lambda_k\}. \tag{9.58}$$

There are $r$ linearly independent solutions of (9.56) but for each linearly independent solution there can be several scaling parameters, $w$, which satisfy $\delta(\mathbf{a})=\lambda_k$. For that reason, $|\mathcal{S}_k|$ is determined by the number of scaling parameters that satisfy $\delta(\mathbf{a})=\lambda_k$. To that end, define $\delta':\mathbb{R}\to\mathbb{R}$ as $\delta'(w)\overset{\Delta}{=}\delta(w\mathbf{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{U}\mathbf{e}_k)$. It is easy to check that $\delta'(w)=\lambda_k$ has a finite number of solutions for $k=1,\dots,r$. Hence, for all $k$ there exists a constant $C_k$ such that $|\mathcal{S}_k|\leq C_k$ which yields the upper bound

$$|\mathcal{S}|\leq\sum_{i=1}^{R}|\mathcal{S}_k|\leq\sum_{i=1}^{R}C_k\leq\max_k C_k R. \tag{9.59}$$

Noticing that the there is a finite number of solutions of $\delta'(w)=\lambda_k$ and that they depend only on $\tau$ yields the upper bound.

□

## 9.5 Information-Theoretic Attacks

Modern sensing infrastructure is moving toward increasing the number of measurements that the operator acquires, e.g. phasor measurement units exhibit temporal resolutions in the order of miliseconds while supervisory control and data acquisition (SCADA) systems traditionally operate with a temporal resolution in the order of seconds. As a result, attack constructions that do not change within the same temporal scale at which measurements are reported do not exploit all the *degrees of freedom* that are available to the attacker.

Indeed, an attacker can choose to change the attack vector with every measurement vector that is reported to the network operator. However, the deterministic attack construction changes when the Jacobian measurement matrix changes, i.e. with the operation point of the system. Thus, in the deterministic attack case, the attack construction changes at the same rate that the Jacobian measurement matrix changes and, therefore, the dynamics of the state variables define the update cadency of the attack vector.

In this section, we study the case in which the attacker constructs the attack vector as a random process that corrupts the measurements. By endowing the attack vector with a probabilistic structure we provide the attacker with an attack construction strategy that generates attack vector realizations over time and that achieve a determined objective on average. In view of this, the task of the attacker in this case is to devise the optimal distribution for the attack vectors. In the following, we pose the attack construction problem within an information-theoretic framework and characterize the attacks that simultaneously minimize the mutual information and the probability of detection.

### 9.5.1 Random Attack Model

We consider an additive attack model as in (9.5) but with the distinction that the attack is a random process. The resulting vector of compromised measurements is given by

$$Y_A^m = \mathbf{H}X^m + Z^m + A^m, \tag{9.60}$$

where $A^m \in \mathbb{R}^m$ is the vector of random variables introduced by the attacker and $Y_A^m \in \mathbb{R}^m$ is the vector containing the compromised measurements. The attack vector of random variables is described by the distribution $P_{A^m}$ which is the determined by the attacker. We assume that the attacker has no access to the realizations of the state variables, and therefore, it holds that $P_{A^m X^n} = P_{A^m} P_{X^n}$ where $P_{A^m X^n}$ denotes the joint distribution of $A^m$ and $X^n$.

Similarly to the deterministic attack case, we adopt a multivariate Gaussian framework for the state variables such that $X^n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{XX})$. Moreover, we limit the attack vector distribution to the set of zero-mean multivariate Gaussian distributions, i.e. $A^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{AA})$ where $\mathbf{\Sigma}_{AA} \in \mathcal{S}_+^m$ is the covariance matrix of the attack distribution. The rationale for choosing a Gaussian distribution for the attack vector follows from the fact that for the measurement model in (9.60) the additive attack distribution that minimizes the mutual information between the vector of state variables and the compromised measurements is Gaussian [8]. As we will see later, minimizing this mutual information is central to the proposed information-theoretic attack construction and indeed one of the objectives of the attacker. Because of the Gaussianity of the attack distribution, the vector of compromised measurements is distributed as

$$Y_A^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{Y_A Y_A}), \tag{9.61}$$

where $\mathbf{\Sigma}_{Y_A Y_A} = \mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^\mathsf{T} + \sigma^2\mathbf{I} + \mathbf{\Sigma}_{AA}$ is the covariance matrix of the distribution of the compromised measurements. Note that while in the case of deterministic attacks the effect

of the attack vector was captured by shifting the mean of the measurement vector, in the random attack case the attack changes the structure of the second order moments of the measurements. Interestingly, the Gaussian attack construction implies that knowledge of the second order moments of the state variables and the variance of the AWGN introduced by the measurement process suffices to construct the attack. This assumption significantly reduces the difficulty of the attack construction.

The operator of the power system makes use of the acquired measurements to detect the attack. The detection problem is cast as a hypothesis testing problem with hypotheses

$$\begin{aligned} &\mathcal{H}_0 : Y^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{YY}), \quad \text{versus} \\ &\mathcal{H}_1 : Y^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{Y_A Y_A}). \end{aligned} \tag{9.62}$$

The null hypothesis $\mathcal{H}_0$ describes the case in which the power system is not compromised, while the alternative hypothesis $\mathcal{H}_1$ describes the case in which the power system is under attack.

Two types of error are considered in hypothesis testing problems, Type I error is the probability of a "true negative" event; and Type II error is the probability of a "false alarm" event. The Neyman-Pearson lemma [9] states that for a fixed probability of Type I error, the likelihood ratio test (LRT) achieves the minimum Type II error when compared with any other test with an equal or smaller Type I error. Consequently, the LRT is chosen to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$ based on the available measurements. The LRT between $\mathcal{H}_0$ and $\mathcal{H}_1$ takes following form:

$$L(\mathbf{y}) \stackrel{\Delta}{=} \frac{f_{Y_A^m}(\mathbf{y})}{f_{Y^m}(\mathbf{y})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau, \tag{9.63}$$

where $\mathbf{y} \in \mathbb{R}^m$ is a realization of the vector of random variables modelling the measurements, $f_{Y_A^m}$ and $f_{Y^m}$ denote the probability density functions (p.d.f.'s) of $Y_A^m$ and $Y^m$, respectively, and $\tau$ is the decision threshold set by the operator to meet the false alarm constraint.

### 9.5.2 Information-Theoretic Setting

The aim of the attacker is twofold. Firstly, it aims to disrupt the state estimation process by corrupting the measurements in such a way that the network operator acquires the least amount of knowledge about the state of the system. Secondly, the attacker aspires to remain stealthy and corrupt the measurements without being detected by the network operator. In the following we propose to information-theoretic measures that provide quantitative metrics for the objectives of the attacker.

The data-integrity of the measurements is measured in terms of the mutual information between the state variables and the measurements. The mutual information between two random variables is a measure of the amount of information that each random variable contains about the other random variable. By adding the attack vector to the measurements the attacker aims to reduce the mutual information which ultimately results in a loss of information about the state by the network operator. Specifically, the attacker aims to minimize $I(X^n; Y_A^m)$. In view of this, it seems reasonable to consider a Gaussian distribution

for the attack vector as the minimum mutual information for the observation model in (9.5) is achieved by additive Gaussian noise.

The probability of attack detection is determined by the detection threshold $\tau$ set by the operator for the LRT and the distribution induced by the attack on the vector of compromised measurements. An analytical expression of the probability of attack detection can be described in closed-form as a function of the distributions describing the measurements under both hypotheses. However, the expression is involved in general and it is not straightforward to incorporate it into an analytical formulation of the attack construction. For that reason, we instead consider the asymptotic performance of the LRT to evaluate the detection performance of the operator. The Chernoff-Stein lemma [10] characterizes the asymptotic exponent of the probability of detection when the number of observations of measurement vectors grows to infinity. In our setting, the Chernoff-Stein lemma states that for any LRT and $\epsilon \in (0, 1/2)$, it holds that

$$\lim_{T\to\infty} \frac{1}{T} \log \beta_T^{\epsilon} = -D(P_{Y_A^m} || P_{Y^m}), \tag{9.64}$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence, $\beta_T^{\epsilon}$ is the minimum Type II error such that the Type I error $\alpha$ satisfies $\alpha < \epsilon$, and $T$ is the number of $m$-dimensional measurement vectors that are available for the LRT detection procedure. As a result, minimizing the asymptotic probability of false alarm given an upper bound on the probability of misdetection is equivalent to minimizing $D(P_{Y_A^m}||P_{Y^m})$, where $P_{Y_A^m}$ and $P_{Y^m}$ denote the probability distributions of $Y_A^m$ and $Y^m$, respectively.

The purpose of the attacker is to disrupt the normal state estimation procedure by minimizing the information that the operator acquires about the state variables, while guaranteeing that the probability of attack detection is sufficiently small, and therefore, remain stealthy.

### 9.5.3 Generalized Stealth Attacks

When the two information-theoretic objectives are considered by the attacker, in [11], a stealthy attack construction is proposed by combining two objectives in one cost function, i.e.,

$$I(X^n; Y_A^m) + D(P_{Y_A^m}||P_{Y^m}) = D(P_{X^n Y_A^m}||P_{X^n} P_{Y^m}), \tag{9.65}$$

where $P_{X^n Y_A^m}$ is the joint distribution of $X^n$ and $Y_A^m$. The resulting optimization problem to construct the attack is given by

$$\min_{A^m} D(P_{X^n Y_A^m}||P_{X^n} P_{Y^m}). \tag{9.66}$$

Therein, it is shown that (9.66) is a convex optimization problem and the covariance matrix of the optimal Gaussian attack is $\mathbf{\Sigma}_{AA} = \mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}}$. However, numerical simulations on IEEE test system show that the attack construction proposed in the preceding text yields large values of probability of detection in practical settings.

To control the probability of attack detection of the attack, the preceding construction is generalized in [12] by introducing a parameter that weights the detection term in the cost function. The resulting optimization problem is given by

$$\min_{A^m} I(X^n; Y_A^m) + \lambda D(P_{Y_A^m} || P_{Y^m}), \tag{9.67}$$

where $\lambda \geq 1$ governs the weight given to each objective in the cost function. It is interesting to note that for the case in which $\lambda = 1$ the proposed cost function boils down to the effective secrecy proposed in [13] and the attack construction in (9.67) coincides with that in [11]. For $\lambda > 1$, the attacker adopts a conservative approach and prioritizes remaining undetected over minimizing the amount of information acquired by the operator. By increasing the value of $\lambda$ the attacker decreases the probability of detection at the expense of increasing the amount of information acquired by the operator using the measurements.

The attack construction in (9.67) is formulated in a general setting. The following propositions particularize the KL divergence and MI to our multivariate Gaussian setting.

**Proposition 9.5.** [10] *The KL divergence between m-dimensional multivariate Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{Y_A Y_A})$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{YY})$ is given by*

$$D(P_{Y_A^m} || P_{Y^m}) = \frac{1}{2}\left(\log \frac{|\mathbf{\Sigma}_{YY}|}{|\mathbf{\Sigma}_{Y_A Y_A}|} - m + \operatorname{tr}\left(\mathbf{\Sigma}_{YY}^{-1} \mathbf{\Sigma}_{Y_A Y_A}\right)\right). \tag{9.68}$$

**Proposition 9.6.** [10] *The mutual information between the vectors of random variables $X^n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{XX})$ and $Y_A^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{Y_A Y_A})$ is given by*

$$I(X^n; Y_A^m) = \frac{1}{2} \log \frac{|\mathbf{\Sigma}_{XX}||\mathbf{\Sigma}_{Y_A Y_A}|}{|\mathbf{\Sigma}|}, \tag{9.69}$$

*where $\mathbf{\Sigma}$ is the covariance matrix of the joint distribution of $(X^n, Y_A^m)$.*

Substituting (9.68) and (9.69) in (9.67) we can now pose the Gaussian attack construction as the following optimization problem:

$$\min_{\mathbf{\Sigma}_{AA} \in \mathcal{S}_+^m} -(\lambda - 1) \log |\mathbf{\Sigma}_{YY} + \mathbf{\Sigma}_{AA}| - \log |\mathbf{\Sigma}_{AA} + \sigma^2 \mathbf{I}| + \lambda \operatorname{tr}(\mathbf{\Sigma}_{YY}^{-1} \mathbf{\Sigma}_{AA}). \tag{9.70}$$

We now proceed to solve the optimization problem in the preceding text. First, note that the optimization domain $\mathcal{S}_+^m$ is a convex set. The following proposition characterizes the convexity of the cost function.

**Proposition 9.7.** *Let $\lambda \geq 1$. Then the cost function in the optimization problem in (9.70) is convex.*

*Proof.* Note that the term $-\log |\mathbf{\Sigma}_{AA} + \sigma^2 \mathbf{I}|$ is a convex function on $\mathbf{\Sigma}_{AA} \in \mathcal{S}_+^m$ [14]. Additionally, $-(\lambda - 1) \log |\mathbf{\Sigma}_{YY} + \mathbf{\Sigma}_{AA}|$ is a convex function on $\mathbf{\Sigma}_{AA} \in \mathcal{S}_+^m$ when $\lambda \geq 1$. Since the trace operator is a linear operator and the sum of convex functions is convex, it follows that the cost function in (9.70) is convex on $\mathbf{\Sigma}_{AA} \in \mathcal{S}_+^m$. □

**Theorem 9.5.** *Let $\lambda \geq 1$. Then the solution to the optimization problem in (9.70) is*

$$\boldsymbol{\Sigma}_{AA}^{\star} = \frac{1}{\lambda}\mathbf{H}\boldsymbol{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}}. \tag{9.71}$$

*Proof.* Denote the cost function in (9.70) by $f(\boldsymbol{\Sigma}_{AA})$. Taking the derivative of the cost function with respect to $\boldsymbol{\Sigma}_{AA}$ yields

$$\begin{aligned}\frac{\partial f(\boldsymbol{\Sigma}_{AA})}{\partial \boldsymbol{\Sigma}_{AA}} =& -2(\lambda-1)(\boldsymbol{\Sigma}_{YY}+\boldsymbol{\Sigma}_{AA})^{-1} - 2(\boldsymbol{\Sigma}_{AA}+\sigma^2\mathbf{I}_M)^{-1} + 2\lambda\boldsymbol{\Sigma}_{YY}^{-1} - \lambda\mathrm{diag}(\boldsymbol{\Sigma}_{YY}^{-1}) \\ &+ (\lambda-1)\mathrm{diag}\left((\boldsymbol{\Sigma}_{YY}+\boldsymbol{\Sigma}_{AA})^{-1}\right) + \mathrm{diag}\left((\boldsymbol{\Sigma}_{AA}+\sigma^2\mathbf{I})^{-1}\right). \end{aligned} \tag{9.72}$$

Note that the only critical point is $\boldsymbol{\Sigma}_{AA}^{\star} = \frac{1}{\lambda}\mathbf{H}\boldsymbol{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}}$. Theorem 9.5 follows immediately from combining this result with Proposition 9.7. □

**Corollary 9.1.** *The mutual information between the vector of state variables and the vector of compromised measurements induced by the optimal attack construction is given by*

$$I(X^n; Y_A^m) = \frac{1}{2}\log\left|\mathbf{H}\boldsymbol{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}}\left(\sigma^2\mathbf{I} + \frac{1}{\lambda}\mathbf{H}\boldsymbol{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}}\right)^{-1} + \mathbf{I}\right|. \tag{9.73}$$

Theorem 9.5 shows that the generalized stealth attacks share the same structure of the stealth attacks in [11] up to a scaling factor determined by $\lambda$. The solution in Theorem 9.5 holds for the case in which $\lambda \geq 1$, and therefore, lacks full generality. However, the case in which $\lambda < 1$ yields unreasonably high probability of detection [11] which indicates that the proposed attack construction is indeed of practical interest in a wide range of state estimation settings.

The resulting attack construction is remarkably simple to implement provided that the information about the system is available to the attacker. Indeed, the attacker only requires access to the linearized Jacobian measurement matrix $\mathbf{H}$ and the second order statistics of the state variables, but the variance of the noise introduced by the sensors is not necessary. To obtain the Jacobian, a malicious attacker needs to know the topology of the grid, the admittances of the branches, and the operation point of the system. The second order statistics of the state variables on the other hand, can be estimated using historical data. In [11] it is shown that the attack construction with a sample covariance matrix of the state variables obtained with historical data is asymptotically optimal when the size of the training data grows to infinity.

It is interesting to note that the mutual information in (9.73) increases monotonically with $\lambda$ and that it asymptotically converges to $I(X^n; Y^m)$, i.e. the case in which there is no attack. While the evaluation of the mutual information as shown in Corollary 9.1 is straightforward, the computation of the associated probability of detection yields involved expressions that do not provide much insight. For that reason, the probability of detection of optimal attacks is treated in the following section.

### 9.5.4 Probability of Detection of Generalized Stealth Attacks

The asymptotic probability of detection of the generalized stealth attacks is governed by the KL divergence as described in (9.64). However in the non-asymptotic case, determining the probability of detection is difficult, and therefore, choosing a value of $\lambda$ that provides the desired probability of detection is a challenging task. In this section we first provide a closed-form expression of the probability of detection by direct evaluation and show that the expression does not provide any practical insight over the choice of $\lambda$ that achieves the desired detection performance. That being the case, we then provide an upper bound on the probability of detection, which, in turn, provides a lower bound on the value of $\lambda$ that achieves the desired probability of detection.

#### Direct Evaluation of the Probability of Detection

Detection based on the LRT with threshold $\tau$ yields a probability of detection given by

$$\mathsf{P}_\mathsf{D} \triangleq \mathbb{E}\left[\mathbb{1}_{\left\{L(Y_A^m)\geq\tau\right\}}\right]. \tag{9.74}$$

The following proposition particularizes the above expression to the optimal attack construction described in Section 9.5.3.

**Lemma 9.2.** *The probability of detection of the LRT in (9.63) for the attack construction in (9.71) is given by*

$$\mathsf{P}_\mathsf{D}(\lambda) = \mathbb{P}\left[(U^p)^\mathsf{T}\boldsymbol{\Delta}U^p \geq \lambda\left(2\log\tau + \log\left|\mathbf{I}+\lambda^{-1}\boldsymbol{\Delta}\right|\right)\right], \tag{9.75}$$

*where $p = rank(\mathbf{H}\boldsymbol{\Sigma}_{XX}\mathbf{H}^\mathsf{T})$, $U^p \in \mathbb{R}^p$ is a vector of random variables with distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$, and $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with entries given by $(\boldsymbol{\Delta})_{i,i} = \lambda_i(\mathbf{H}\boldsymbol{\Sigma}_{XX}\mathbf{H}^\mathsf{T})\lambda_i(\boldsymbol{\Sigma}_{YY}^{-1})$, where $\lambda_i(\mathbf{A})$ with $i = 1,\ldots,p$ denotes the $i$-th eigenvalue of matrix $\mathbf{A}$ in descending order.*

*Proof.* The probability of detection of the stealth attack is,

$$\mathsf{P}_\mathsf{D}(\lambda) = \int_{\mathcal{S}} \mathrm{d}P_{Y_A^m} \tag{9.76}$$

$$= \frac{1}{(2\pi)^{\frac{m}{2}}\left|\boldsymbol{\Sigma}_{Y_AY_A}\right|^{\frac{1}{2}}}\int_{\mathcal{S}} \exp\left\{-\frac{1}{2}\mathbf{y}^\mathsf{T}\boldsymbol{\Sigma}_{Y_AY_A}^{-1}\mathbf{y}\right\}\mathrm{d}\mathbf{y}, \tag{9.77}$$

where

$$\mathcal{S} = \{\mathbf{y}\in\mathbb{R}^m : L(\mathbf{y}) \geq \tau\}. \tag{9.78}$$

Algebraic manipulation yields the following equivalent description of the integration domain:

$$\mathcal{S} = \left\{\mathbf{y}\in\mathbb{R}^m : \mathbf{y}^\mathsf{T}\boldsymbol{\Delta}_0\mathbf{y} \geq 2\log\tau + \log\left|\mathbf{I}+\boldsymbol{\Sigma}_{AA}\boldsymbol{\Sigma}_{YY}^{-1}\right|\right\}, \tag{9.79}$$

with $\boldsymbol{\Delta}_0 \triangleq \boldsymbol{\Sigma}_{YY}^{-1} - \boldsymbol{\Sigma}_{Y_AY_A}^{-1}$. Let $\boldsymbol{\Sigma}_{YY} = \mathbf{U}_{YY}\boldsymbol{\Lambda}_{YY}\mathbf{U}_{YY}^\mathsf{T}$ where $\boldsymbol{\Lambda}_{YY} \in \mathbb{R}^{m\times m}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}_{YY}$ in descending order and $\mathbf{U}_{YY} \in \mathbb{R}^{m\times m}$ is a

unitary matrix whose columns are the eigenvectors of $\mathbf{\Sigma}_{YY}$ ordered matching the order of the eigenvalues. Applying the change of variable $\mathbf{y}_1 \triangleq \mathbf{U}_{YY}\mathbf{y}$ in (9.77) results in

$$\mathsf{P}_{\mathsf{D}}(\lambda) = \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{\Sigma}_{Y_A Y_A}|^{\frac{1}{2}}} \int_{\mathcal{S}_1} \exp\left\{-\frac{1}{2}\mathbf{y}_1^{\mathsf{T}} \mathbf{\Lambda}_{Y_A Y_A}^{-1} \mathbf{y}_1\right\} \mathrm{d}\mathbf{y}_1, \tag{9.80}$$

where $\mathbf{\Lambda}_{Y_A Y_A} \in \mathbb{R}^{m\times m}$ denotes the diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}_{Y_A Y_A}$ in descending order. Noticing that $\mathbf{\Sigma}_{YY}$, $\mathbf{\Sigma}_{AA}$ and $\mathbf{\Sigma}_{Y_A Y_A}$ are also diagonalized by $\mathbf{U}_{YY}$, the integration domain $\mathcal{S}_1$ is given by

$$\mathcal{S}_1 = \left\{\mathbf{y}_1 \in \mathbb{R}^m : \mathbf{y}_1^{\mathsf{T}} \mathbf{\Delta}_1 \mathbf{y}_1 \geq 2\log\tau + \log|\mathbf{I} + \mathbf{\Lambda}_{AA}\mathbf{\Lambda}_{YY}^{-1}|\right\}, \tag{9.81}$$

where $\mathbf{\Delta}_1 \triangleq \mathbf{\Lambda}_{YY}^{-1} - \mathbf{\Lambda}_{Y_A Y_A}^{-1}$ with $\mathbf{\Lambda}_{AA}$ denoting the diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}_{AA}$ in descending order. Further applying the change of variable $\mathbf{y}_2 \triangleq \mathbf{\Lambda}_{Y_A Y_A}^{-\frac{1}{2}} \mathbf{y}_1$ in (9.80) results in

$$\mathsf{P}_{\mathsf{D}}(\lambda) = \frac{1}{\sqrt{(2\pi)^m}} \int_{\mathcal{S}_2} \exp\{-\frac{1}{2}\mathbf{y}_2^{\mathsf{T}}\mathbf{y}_2\} \mathrm{d}\mathbf{y}_2, \tag{9.82}$$

with the transformed integration domain given by

$$\mathcal{S}_2 = \left\{\mathbf{y}_2 \in \mathbb{R}^m : \mathbf{y}_2^{\mathsf{T}} \mathbf{\Delta}_2 \mathbf{y}_2 \geq 2\log\tau + \log|\mathbf{I} + \mathbf{\Delta}_2|\right\}, \tag{9.83}$$

with

$$\mathbf{\Delta}_2 \triangleq \mathbf{\Lambda}_{AA}\mathbf{\Lambda}_{YY}^{-1}. \tag{9.84}$$

Setting $\mathbf{\Delta} \triangleq \lambda\mathbf{\Delta}_2$ and noticing that $\mathrm{rank}(\mathbf{\Delta}) = \mathrm{rank}(\mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}})$ concludes the proof. □

Notice that the left-hand term $(U^p)^{\mathsf{T}}\mathbf{\Delta}U^p$ in (9.75) is a weighted sum of independent $\chi^2$ distributed random variables with one degree of freedom where the weights are determined by the diagonal entries of $\mathbf{\Delta}$ which depend on the second order statistics of the state variables, the Jacobian measurement matrix, and the variance of the noise; i.e. the attacker has no control over this term. The right-hand side contains in addition $\lambda$ and $\tau$, and therefore, the probability of attack detection is described as a function of the parameter $\lambda$. However, characterizing the distribution of the resulting random variable is not practical since there is no closed-form expression for the distribution of a positively weighted sum of independent $\chi^2$ random variables with one degree of freedom [15]. Usually, some moment matching approximation approaches such as the Lindsay-Pilla-Basak method [16] are utilized to solve this problem but the resulting expressions are complex and the relation of the probability of detection with $\lambda$ is difficult to describe analytically following this course of action. In the following an upper bound on the probability of attack detection is derived. The upper bound is then used to provide a simple lower bound on the value $\lambda$ that achieves the desired probability of detection.

**Upper Bound on the Probability of Detection**

The following theorem provides a sufficient condition for $\lambda$ to achieve a desired probability of attack detection.

**Theorem 9.6.** *Let $\tau > 1$ be the decision threshold of the LRT. For any $t > 0$ and $\lambda \geq \max(\lambda^{\star}(t), 1)$ then the probability of attack detection satisfies*

$$\mathsf{P}_{\mathsf{D}}(\lambda) \leq e^{-t}, \tag{9.85}$$

*where $\lambda^{*}(t)$ is the only positive solution of $\lambda$ satisfying*

$$2\lambda \log \tau - \frac{1}{2\lambda}\mathrm{tr}(\boldsymbol{\Delta}^2) - 2\sqrt{\mathrm{tr}(\boldsymbol{\Delta}^2)t} - 2||\boldsymbol{\Delta}||_{\infty} t = 0. \tag{9.86}$$

*and $||\cdot||_{\infty}$ is the infinity norm.*

*Proof.* We start with the result of Lemma 9.2 which gives

$$\mathsf{P}_{\mathsf{D}}(\lambda) = \mathbb{P}\left[(U^p)^{\mathsf{T}}\boldsymbol{\Delta}U^p \geq \lambda\left(2\log\tau + \log\left|\mathbf{I} + \lambda^{-1}\boldsymbol{\Delta}\right|\right)\right]. \tag{9.87}$$

We now proceed to expand the term $\log\left|\mathbf{I} + \lambda^{-1}\boldsymbol{\Delta}\right|$ using a Taylor series expansion resulting in

$$\log\left|\mathbf{I} + \lambda^{-1}\boldsymbol{\Delta}\right| = \sum_{i=1}^{p} \log\left(1 + \lambda^{-1}(\boldsymbol{\Delta})_{i,i}\right) \tag{9.88}$$

$$= \sum_{i=1}^{p}\left(\sum_{j=1}^{\infty}\left(\frac{\left(\lambda^{-1}(\boldsymbol{\Delta})_{i,i}\right)^{2j-1}}{2j-1} - \frac{\left(\lambda^{-1}(\boldsymbol{\Delta})_{i,i}\right)^{2j}}{2j}\right)\right). \tag{9.89}$$

Because $(\boldsymbol{\Delta})_{i,i} \leq 1, \text{for } i = 1, \ldots, p$, and $\lambda \geq 1$, then

$$\frac{\left(\lambda^{-1}(\boldsymbol{\Delta})_{i,i}\right)^{2j-1}}{2j-1} - \frac{\left(\lambda^{-1}(\boldsymbol{\Delta})_{i,i}\right)^{2j}}{2j} \geq 0, \text{ for } j \in \mathbb{Z}^{+}. \tag{9.90}$$

Thus, (9.89) is lower bounded by the second order Taylor expansion, i.e.,

$$\log|\mathbf{I} + \boldsymbol{\Delta}| \geq \sum_{i=1}^{p}\left(\lambda^{-1}(\boldsymbol{\Delta})_{i,i} - \frac{\left(\lambda^{-1}(\boldsymbol{\Delta})_{i,i}\right)^2}{2}\right) \tag{9.91}$$

$$= \frac{1}{\lambda}\mathrm{tr}(\boldsymbol{\Delta}) - \frac{1}{2\lambda^2}\mathrm{tr}(\boldsymbol{\Delta}^2). \tag{9.92}$$

Substituting (9.92) in (9.87) yields

$$\mathsf{P}_{\mathsf{D}}(\lambda) \leq \mathbb{P}\left[(U^p)^{\mathsf{T}}\boldsymbol{\Delta}U^p \geq \mathrm{tr}(\boldsymbol{\Delta}) + 2\lambda\log\tau - \frac{1}{2\lambda}\mathrm{tr}(\boldsymbol{\Delta}^2)\right]. \tag{9.93}$$

Note that $\mathbb{E}\left[(U^p)^\mathsf{T}\mathbf{\Delta}U^p\right] = \mathrm{tr}(\mathbf{\Delta})$, and therefore, evaluating the probability in (9.93) is equivalent to evaluating the probability of $(U^p)^\mathsf{T}\mathbf{\Delta}U^p$ deviating $2\lambda\log\tau - \frac{1}{2\lambda}\mathrm{tr}(\mathbf{\Delta}^2)$ from the mean. In view of this, the right-hand side in (9.93) is upper bounded by [17, 18]

$$\mathsf{P}_\mathsf{D}(\lambda) \leq \mathbb{P}\left[(U^p)^\mathsf{T}\mathbf{\Delta}U^p \geq \mathrm{tr}(\mathbf{\Delta}) + 2\sqrt{\mathrm{tr}(\mathbf{\Delta}^2)t} + 2||\mathbf{\Delta}||_\infty t\right] \leq e^{-t}, \tag{9.94}$$

for $t > 0$ satisfying

$$2\lambda\log\tau - \frac{1}{2\lambda}\mathrm{tr}(\mathbf{\Delta}^2) \geq 2\sqrt{\mathrm{tr}(\mathbf{\Delta}^2)t} + 2||\mathbf{\Delta}||_\infty t. \tag{9.95}$$

The expression in (9.95) is satisfied with equality for two values of $\lambda$, one is strictly negative and the other one is strictly positive denoted by $\lambda^\star(t)$, when $\tau > 1$. The result follows by noticing that the left-hand term of (9.95) increases monotonically for $\lambda > 0$ and choosing $\lambda \geq \max\left(\lambda^\star(t), 1\right)$. This concludes the proof. □

It is interesting to note that for large values of $\lambda$ the probability of detection decreases exponentially fast with $\lambda$. We will later show in the numerical results that the regime in which the exponentially fast decrease kicks in does not align with the saturation of the mutual information loss induced by the attack.

### 9.5.5 Numerical Evaluation of Stealth Attacks

We evaluate the performance of stealth attacks in practical state estimation settings. n particular, the IEEE 14-Bus, 30-Bus and 118-Bus test systems are considered in the simulation. In state estimation with linearized dynamics, the Jacobian measurement matrix is determined by the operation point. We assume a DC state estimation scenario [19, 20], and thus, we set the resistances of the branches to 0 and the bus voltage magnitude to 1.0 per unit. Note that in this setting it is sufficient to specify the network topology, the branch reactances, real power flow, and the power injection values to fully characterize the system. Specifically, we use the IEEE test system framework provided by MATPOWER [21]. We choose the bus voltage angle to be the state variables, and use the power injection and the power flows in both directions as the measurements.

As stated in Section 9.5.4, there is no closed-form expression for the distribution of a positively weighted sum of independent $\chi^2$ random variables, which is required to calculate the probability of detection of the generalized stealth attacks as shown in Lemma 9.2. For that reason, we use the Lindsay–Pilla–Basak method and the MOMENTCHI2 package [22] to numerically evaluate the probability of attack detection.

The covariance matrix of the state variables is modelled as a Toeplitz matrix with exponential decay parameter $\rho$, where the exponential decay parameter $\rho$ determines the correlation strength between different entries of the state variable vector. The performance of the generalized stealth attack is a function of weight given to the detection term in the attack construction cost function, i.e. $\lambda$, the correlation strength between state variables,

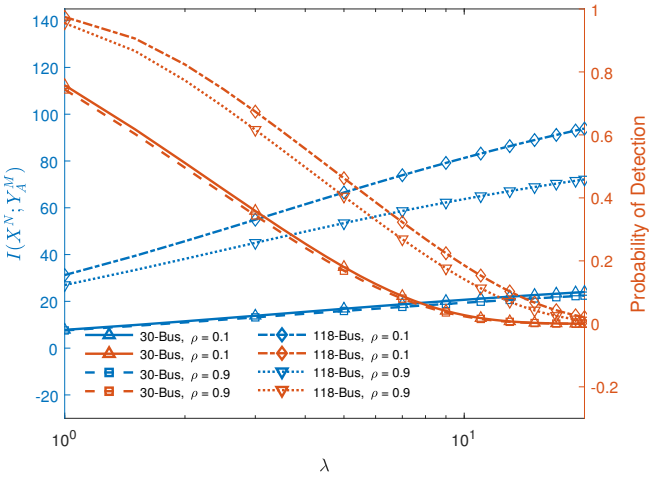


Figure 9.1: Performance of the generalized stealth attack in terms of mutual information and probability of detection for different values of $\lambda$ and system size when $\rho = 0.1$, $\rho = 0.9$, SNR $= 10$ dB and $\tau = 2$.

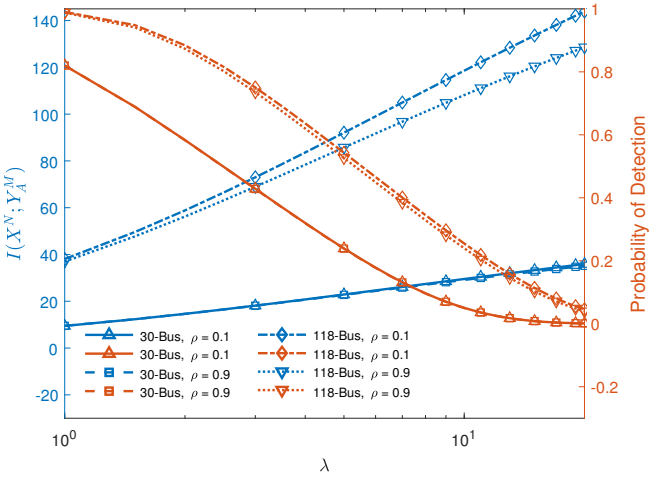


Figure 9.2: Performance of the generalized stealth attack in terms of mutual information and probability of detection for different values of $\lambda$ and system size when $\rho = 0.1$, $\rho = 0.9$, SNR $= 20$ dB and $\tau = 2$.

i.e. $\rho$, and the Signal-to-Noise Ratio (SNR) of the power system which is defined as

$$\text{SNR} \triangleq 10 \log_{10} \left( \frac{\text{tr}(\mathbf{H} \boldsymbol{\Sigma}_{XX} \mathbf{H}^{\mathsf{T}})}{m\sigma^2} \right). \tag{9.96}$$

Fig. 9.1 and Fig. 9.2 depict the performance of the optimal attack construction for different values of $\lambda$ and $\rho$ with SNR $= 10$ dB and SNR $= 20$ dB, respectively, when $\tau = 2$.

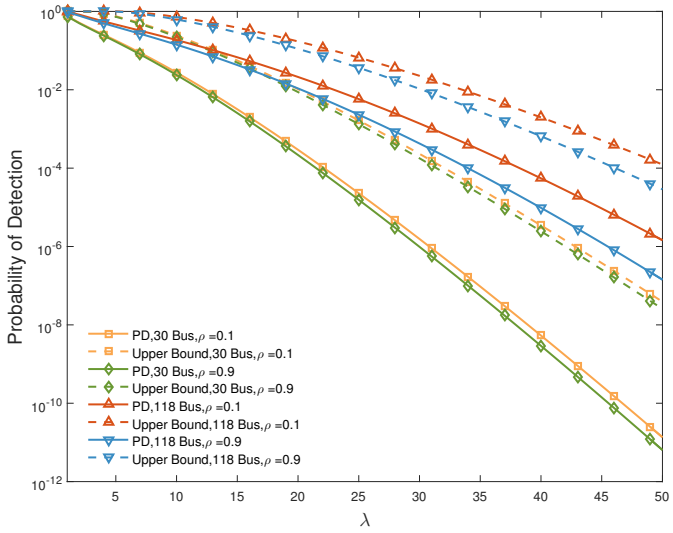


Figure 9.3: Upper bound on probability of detection given in Theorem 9.6 for different values of $\lambda$ when $\rho = 0.1$ or 0.9, SNR $= 10$ dB, and $\tau = 2$.

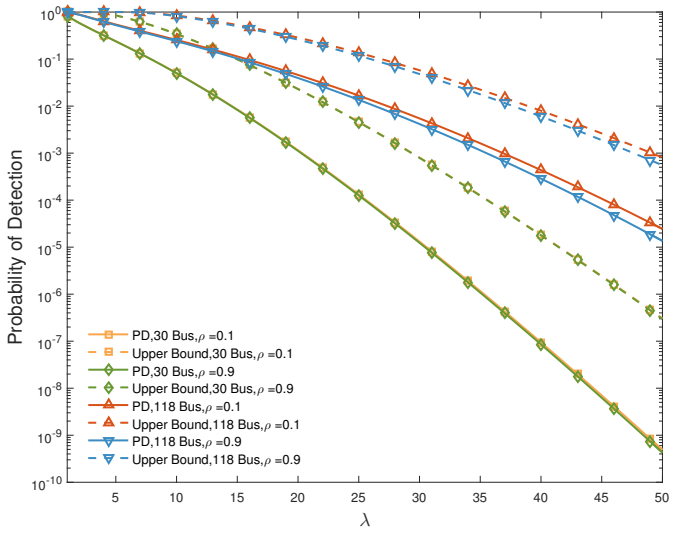


Figure 9.4: Upper bound on probability of detection given in Theorem 9.6 for different values of $\lambda$ when $\rho = 0.1$ or 0.9, SNR $= 20$ dB, and $\tau = 2$.

As expected, larger values of the parameter $\lambda$ yield smaller values of the probability of attack detection while increasing the mutual information between the state variables vector and the compromised measurement vector. We observe that the probability of detection decreases approximately linearly for moderate values of $\lambda$. On the other hand, Theorem 9.6 states that for large values of $\lambda$ the probability of detection decreases exponentially fast to

zero. However, for the range of values of $\lambda$ in which the decrease of probability of detection is approximately linear, there is no significant reduction on the rate of growth of mutual information. In view of this, the attacker needs to choose the value of $\lambda$ carefully as the convergence of the mutual information to the asymptote $I(X^n; Y^m)$ is slower than that of the probability of detection to zero.

The comparison between the 30-Bus and 118-Bus systems shows that for the smaller size system the probability of detection decreases faster to zero while the rate of growth of mutual information is smaller than that on the larger system. This suggests that the choice of $\lambda$ is particularly critical in large size systems as smaller size systems exhibit a more robust attack performance for different values of $\lambda$. The effect of the correlation between the state variables is significantly more noticeable for the 118-bus system. While there is a performance gain for the 30-bus system in terms of both mutual information and probability of detection due to the high correlation between the state variables, the improvement is more noteworthy for the 118-bus case. Remarkably, the difference in terms of mutual information between the case in which $\rho = 0.1$ and $\rho = 0.9$ increases as $\lambda$ increases which indicates that the cost in terms of mutual information of reducing the probability of detection is large in the small values of correlation.

The performance of the upper bound given by Theorem 9.6 on the probability of detection for different values of $\lambda$ and $\rho$ when $\tau = 2$ and SNR $= 10$ dB is shown in Fig. 9.3. Similarly, Fig. 9.4 depicts the upper bound with the same parameters but with SNR $= 20$ dB. As shown by Theorem 9.6 the bound decreases exponentially fast for large values of $\lambda$. Still, there is a significant gap to the probability of attack detection evaluated numerically. This is partially due to the fact that our bound is based on the concentration inequality in [17] which introduces a gap of more than an order of magnitude. Interestingly, the gap decreases when the value of $\rho$ increases although the change is not significant. More importantly, the bound is tighter for lower values of SNR for both 30-bus and 118-bus systems.

## 9.6 Attack Construction with Estimated State Variable Statistics

### 9.6.1 Learning the Second-Order Statistics of the State Variables

The stealth attack construction proposed in the preceding text requires perfect knowledge of the covariance matrix of the state variables and the linearized Jacobian measurement matrix. In [23], the performance of the attack when the second-order statistics are not perfectly known by the attacker but the linearized Jacobian measurement matrix is known. Therein, the partial knowledge is modelled by assuming that the attacker has access to a sample covariance matrix of the state variables. Specifically, the training data consisting of $k$ state variable realizations $\{\mathbf{x}_i^n\}_{i=1}^k$ is available to the attacker. That being the case the attacker computes the unbiased estimate of the covariance matrix of the state variables given by

$$\mathbf{S}_{XX} = \frac{1}{k-1}\sum_{i=1}^{k} \mathbf{x}_i^n (\mathbf{x}_i^n)^{\mathsf{T}}. \tag{9.97}$$

The stealth attack constructed using the sample covariance matrix follows a multivariate Gaussian distribution given by

$$\tilde{A}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\tilde{A}\tilde{A}}), \tag{9.98}$$

where $\mathbf{\Sigma}_{\tilde{A}\tilde{A}} = \mathbf{H}\mathbf{S}_{XX}\mathbf{H}^{\mathsf{T}}$.

Because the sample covariance matrix in (9.97) is a random matrix with central Wishart distribution given by

$$\mathbf{S}_{XX} \sim \frac{1}{k-1} W_n(k-1, \mathbf{\Sigma}_{XX}), \tag{9.99}$$

the ergodic counterpart of the cost function in (9.65) is defined in terms of the conditional KL divergence given by

$$\mathbb{E}_{\mathbf{S}_{XX}}\left[D\left(P_{X^n Y_A^m|\mathbf{S}_{XX}} \| P_{X^n} P_{Y^m}\right)\right]. \tag{9.100}$$

The ergodic cost function characterizes the expected performance of the attack averaged over the realizations of training data. Note that the performance using the sample covariance matrix is suboptimal [11] and that the ergodic performance converges asymptotically to the optimal attack construction when the size of the training data set increases.

### 9.6.2 Ergodic Stealth Attack Performance

In this section, we analytically characterize the ergodic attack performance defined in (9.100) by providing an upper bound using random matrix theory tools. Before introducing the upper bound, some auxiliary results on the expected value of the extreme eigenvalues of Wishart random matrices are presented below.

#### Auxiliary Results in Random Matrix Theory

**Lemma 9.3.** *Let $\mathbf{Z}_l$ be an $(k-1)\times l$ matrix whose entries are independent standard normal random variables, then*

$$\text{var}\left(s_{max}(\mathbf{Z}_l)\right) \leq 1, \tag{9.101}$$

*where* var $(\cdot)$ *denotes the variance and $s_{max}(\mathbf{Z}_l)$ is the maximum singular value of $\mathbf{Z}_l$.*

*Proof.* Note that $s_{max}(\mathbf{Z}_l)$ is a 1-Lipschitz function of matrix $\mathbf{Z}_l$, the maximum singular value of $\mathbf{Z}_l$ is concentrated around the mean [24, Proposition 5.34] given by $\mathbb{E}[s_{max}(\mathbf{Z}_l)]$. Then for $t \geq 0$, it holds that

$$\mathbb{P}[|s_{max}(\mathbf{Z}_l) - \mathbb{E}[s_{max}(\mathbf{Z}_l)]| > t] \leq 2\exp\{-t^2/2\} \tag{9.102}$$

$$\leq \exp\{1 - t^2/2\}. \tag{9.103}$$

Therefore $s_{max}(\mathbf{Z}_l)$ is a sub-gaussian random variable with variance proxy $\sigma_p^2 \leq 1$. The lemma follows from the fact that $\text{var}\left(s_{max}(\mathbf{Z}_l)\right) \leq \sigma_p^2$. □

**Lemma 9.4.** *Let $\mathbf{W}_l$ denote a central Wishart matrix distributed as $\frac{1}{k-1}W_l(k-1,\mathbf{I})$, then the non-asymptotic expected value of the extreme eigenvalues of $\mathbf{W}_l$ is bounded by*

$$\left(1-\sqrt{l/(k-1)}\right)^2 \leq \mathbb{E}[\lambda_{min}(\mathbf{W}_l)] \tag{9.104}$$

*and*

$$\mathbb{E}[\lambda_{max}(\mathbf{W}_l)] \leq \left(1+\sqrt{l/(k-1)}\right)^2 + 1/(k-1), \tag{9.105}$$

*where $\lambda_{min}(\mathbf{W}_l)$ and $\lambda_{max}(\mathbf{W}_l)$ denote the minimum eigenvalue and maximum eigenvalue of $\mathbf{W}_l$, respectively.*

*Proof.* Note that [24, Theorem 5.32]

$$\sqrt{k-1}-\sqrt{l} \leq \mathbb{E}[s_{min}(\mathbf{Z}_l)] \tag{9.106}$$

and

$$\sqrt{k-1}+\sqrt{l} \geq \mathbb{E}[s_{max}(\mathbf{Z}_l)], \tag{9.107}$$

where $s_{min}(\mathbf{Z}_l)$ is the minimum singular value of $\mathbf{Z}_l$. Given the fact that $\mathbf{W}_l = \frac{1}{k-1}\mathbf{Z}_l^{\mathsf{T}}\mathbf{Z}_l$, then it holds that

$$\mathbb{E}[\lambda_{min}(\mathbf{W}_l)] = \frac{\mathbb{E}\left[s_{min}(\mathbf{Z}_l)^2\right]}{k-1} \geq \frac{\mathbb{E}\left[s_{min}(\mathbf{Z}_l)\right]^2}{k-1} \tag{9.108}$$

and

$$\mathbb{E}[\lambda_{max}(\mathbf{W}_l)] = \frac{\mathbb{E}\left[s_{max}(\mathbf{Z}_l)^2\right]}{k-1} \leq \frac{\mathbb{E}\left[s_{max}(\mathbf{Z}_l)\right]^2+1}{k-1}, \tag{9.109}$$

where (9.109) follows from Lemma 9.3. Combining (9.106) with (9.108), and (9.107) with (9.109), respectively, yields the lemma. □

Recall the cost function describing the attack performance given in (9.100) can be written in terms of the covariance matrix $\mathbf{\Sigma}_{\tilde{A}\tilde{A}}$ in the multivariate Gaussian case with imperfect second-order statistics. The ergodic cost function that results from averaging the cost over the training data yields

$$\mathbb{E}_{\mathbf{S}_{XX}}\left[D\left(P_{X^nY_A^m|\mathbf{S}_{XX}}\|P_{X^n}P_{Y^m}\right)\right] = \frac{1}{2}\mathbb{E}\left[\mathrm{tr}(\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{\tilde{A}\tilde{A}})-\log|\mathbf{\Sigma}_{\tilde{A}\tilde{A}}+\sigma^2\mathbf{I}|-\log|\mathbf{\Sigma}_{YY}^{-1}|\right] \tag{9.110}$$

$$= \frac{1}{2}\Big(\mathrm{tr}\big(\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{AA}^{\star}\big)-\log|\mathbf{\Sigma}_{YY}^{-1}|-\mathbb{E}\big[\log|\mathbf{\Sigma}_{\tilde{A}\tilde{A}}+\sigma^2\mathbf{I}|\big]\Big). \tag{9.111}$$

The assessment of the ergodic attack performance boils down to evaluating the last term in (9.110). Closed form expressions for this term are provided in [25] for the same case considered in this paper. However, the resulting expressions are involved and are only computable for small dimensional settings. For systems with a large number of dimensions the expressions are computationally prohibitive. To circumvent this challenge we propose a lower bound on the term that yields an upper bound on the ergodic attack performance. Before presenting the main result we provide the following auxiliary convex optimization result.

**Lemma 9.5.** *Let $\mathbf{W}_p$ denote a central Wishart matrix distributed as $\frac{1}{k-1}W_p(k-1,\mathbf{I})$ and let $\mathbf{B}=\mathrm{diag}(b_1,\ldots,b_p)$ denote a positive definite diagonal matrix. Then*

$$\mathbb{E}\left[\log\left|\mathbf{B}+\mathbf{W}_p^{-1}\right|\right]\geq\sum_{i=1}^{p}\log\left(b_i+1/x_i^{\star}\right), \tag{9.112}$$

*where $x_i^{\star}$ is the solution to the convex optimization problem given by*

$$\min_{\{x_i\}_{i=1}^p}\ \sum_{i=1}^{p}\log\left(b_i+1/x_i\right) \tag{9.113}$$

$$s.t.\ \ \sum_{i=1}^{p}x_i=p \tag{9.114}$$

$$\max\left(x_i\right)\leq\left(1+\sqrt{p/(k-1)}\right)^2+1/(k-1) \tag{9.115}$$

$$\min\left(x_i\right)\geq\left(1-\sqrt{p/(k-1)}\right)^2. \tag{9.116}$$

*Proof.* Note that

$$\mathbb{E}\left[\log\left|\mathbf{B}+\mathbf{W}_p^{-1}\right|\right]=\sum_{i=1}^{p}\mathbb{E}\left[\log\left(b_i+\frac{1}{\lambda_i(\mathbf{W}_p)}\right)\right] \tag{9.117}$$

$$\geq\sum_{i=1}^{p}\log\left(b_i+\frac{1}{\mathbb{E}[\lambda_i(\mathbf{W}_p)]}\right), \tag{9.118}$$

where in (9.117), $\lambda_i(\mathbf{W}_p)$ is the $i$-th eigenvalue of $\mathbf{W}_p$ in decreasing order; (9.118) follows from Jensen's inequality due to the convexity of $\log\left(b_i+\frac{1}{x}\right)$ for $x>0$. Constraint (9.114) follows from the fact that $\mathbb{E}[\mathrm{trace}(\mathbf{W}_p)]=p$, and constraints (9.115) and (9.116) follow from Lemma 9.4. This completes the proof. □

### Upper Bound on the Ergodic Stealth Attack Performance

The following theorem provides a lower bound for the last term in (9.110), and therefore, it enables us to upper bound the ergodic stealth attack performance.

**Theorem 9.7.** *Let $\mathbf{\Sigma}_{\tilde{A}\tilde{A}}=\mathbf{H}\mathbf{S}_{XX}\mathbf{H}^{\mathsf{T}}$ with $\mathbf{S}_{XX}$ distributed as $\frac{1}{k-1}W_n(k-1,\mathbf{\Sigma}_{XX})$ and denote by $\mathbf{\Lambda}_p=\mathrm{diag}(\lambda_1,\ldots,\lambda_p)$ the diagonal matrix containing the nonzero eigenvalues in decreasing order. Then*

$$\mathbb{E}\big[\log|\mathbf{\Sigma}_{\tilde{A}\tilde{A}}+\sigma^2\mathbf{I}|\big]\geq\left(\sum_{i=0}^{p-1}\psi(k-1-i)\right)-p\log(k-1)+\sum_{i=1}^{p}\log\left(\frac{\lambda_i}{\sigma^2}+\frac{1}{\lambda_i^{\star}}\right)+2m\log\sigma, \tag{9.119}$$

*where $\psi(\cdot)$ is the Euler digamma function, $p=\mathrm{rank}(\mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^{\mathsf{T}})$, and $\{\lambda_i^{\star}\}_{i=1}^p$ is the solution to the optimization problem given by (9.113) - (9.116) with $b_i=\frac{\lambda_i}{\sigma^2}$, for $i=1,\ldots,p$.*

*Proof.* We proceed by noticing that

$$\mathbb{E}\big[\log|\mathbf{\Sigma}_{\tilde{A}\tilde{A}}+\sigma^2\mathbf{I}|\big] = \mathbb{E}\left[\log\left|\frac{1}{(k-1)\sigma^2}\mathbf{Z}_m^{\mathrm{T}}\mathbf{\Lambda}\mathbf{Z}_m+\mathbf{I}\right|\right]+2m\log\sigma \tag{9.120}$$

$$= \mathbb{E}\left[\log\left|\frac{\mathbf{\Lambda}_p}{\sigma^2}\frac{\mathbf{Z}_p^{\mathrm{T}}\mathbf{Z}_p}{k-1}+\mathbf{I}\right|\right]+2m\log\sigma \tag{9.121}$$

$$= \mathbb{E}\left[\log\left|\frac{\mathbf{Z}_p^{\mathrm{T}}\mathbf{Z}_p}{k-1}\right|+\log\left|\frac{\mathbf{\Lambda}_p}{\sigma^2}+\left(\frac{\mathbf{Z}_p^{\mathrm{T}}\mathbf{Z}_p}{k-1}\right)^{-1}\right|\right]+2m\log\sigma \tag{9.122}$$

$$\geq \left(\sum_{i=0}^{p-1}\psi(k-1-i)\right)-p\log(k-1)+\sum_{i=1}^{p}\log\left(\frac{\lambda_i}{\sigma^2}+\frac{1}{\lambda_i^\star}\right)+2m\log\sigma, \tag{9.123}$$

where in (9.120), $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^{\mathrm{T}}$ in decreasing order; (9.121) follows from the fact that $p=\mathrm{rank}(\mathbf{H}\mathbf{\Sigma}_{XX}\mathbf{H}^{\mathrm{T}})$; (9.123) follows from [26, Theorem 2.11] and Lemma 9.5. This completes the proof. □

**Theorem 9.8.** *The ergodic attack performance given in (9.110) is upper bounded by*

$$\mathbb{E}\left[f(\mathbf{\Sigma}_{\tilde{A}\tilde{A}})\right] \leq \frac{1}{2}\Bigg(\mathrm{trace}\big(\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{AA}^{\star}\big)-\log\big|\mathbf{\Sigma}_{YY}^{-1}\big|-2m\log\sigma \tag{9.124}$$

$$-\Big(\sum_{i=0}^{p-1}\psi(k-1-i)\Big)+p\log(k-1) \tag{9.125}$$

$$-\sum_{i=1}^{p}\log\left(\frac{\lambda_i}{\sigma^2}+\frac{1}{\lambda_i^\star}\right)\Bigg). \tag{9.126}$$

*Proof.* The proof follows immediately from combing Theorem 9.7 with (9.110). □

Fig.9.5 depicts the upper bound in Theorem 9.8 as a function of number of samples for $\rho=0.1$ and $\rho=0.8$ when SNR = 20 dB. Interestingly, the upper bound in Theorem 9.8 is tight for large values of the training data set size for all values of the exponential decay parameter determining the correlation.

## 9.7 Conclusions

We have cast the state estimation problem in a Bayesian setting and shown that the attacker can construct data-injection attacks that exploit prior knowledge about the state variables. In particular, we have focused in multivariate Gaussian random processes to describe the state variables and proposed two attack construction strategies: determinis- tic attacks and random attacks.

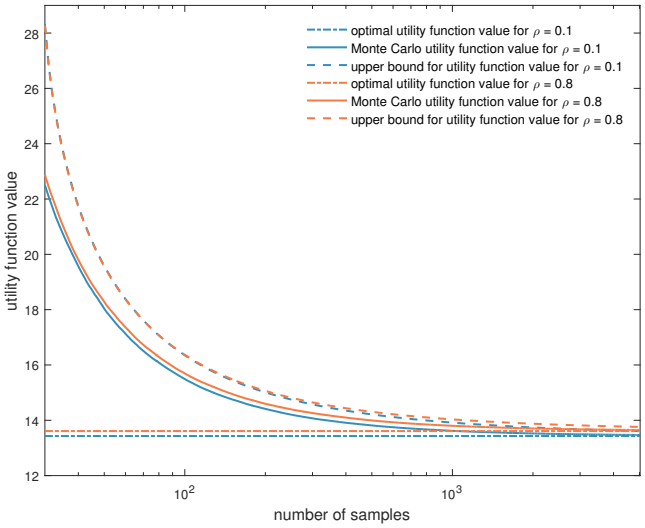


Figure 9.5: Performance of the upper bound in Theorem 9.8 as a function of number of sample for $\rho = 0.1$ and $\rho = 0.8$ when SNR $= 20$ dB.

The deterministic attack is specified by the power system and the statistical structure of the state variables. The attack problem is cast as a multiobjective optimization problem in which the attacker aims to simultaneously minimize the MSE distortion induced by the injection vector and the probability of the attack being detected using a LRT. Within this setting, we have characterized the tradeoff between the achievable distortion and probability of detection by deriving optimal centralized attack constructions for a given distortion and probability of detection pair. We have then extended the investi- gation to decentralized scenarios in which several attackers construct their respective attack without coordination. In this setting, we have posed the interaction between the attackers in a game-theoretic setting. We show that the proposed utility function results in a setting that can be described as a potential game that allows us to claim the existence of an NE and the convergence of BRD to an NE.

The random attack produces different attack vectors for each set of measurements that are reported to the state estimator. The attack vectors are generated by sampling a defined attack vector distribution that yields attack vector realizations to be added to the measurements. The attack aims to disrupt the state estimation process by minimizing the mutual information between the state variables and the altered measurements while minimizing the probability of detection. The rationale for posing the attack construction in information-theoretic terms stems from the fundamental character that information measures grant to the attack vector. By minimizing the mutual information, the attacker limits the performance of a wide range of estimation, detection, and learning options for the operator. We conclude the chapter by analyzing the impact of imperfect second- order statistics about the state variables in the attack performance. In particular, we consider the case in which the

attacker has access to a limited set of training state variable observations that are used to produce the sample covariance matrix of the state variables. Using random matrix theory tools we provide an upper bound on the ergodic attack performance.

This work was supported in part by the European Commission under Marie Skodowska–Curie Individual Fellowship No. 659316 and in part by the Agence Nationale de la Recherche (ANR, France) under Grant ANR-15-NMED-0009-03 and the China Scholarship Council (CSC, China).

# Bibliography


[1] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proc. ACM Conf. on Computer and Communications Security*, Chicago, IL, USA, Nov. 2009, pp. 21–32.

[2] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.

[3] I. Esnaola, S. M. Perlaza, H. V. Poor, and O. Kosut, "Maximum distortion attacks in electricity grids," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 2007–2015, Jul. 2016.

[4] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1994.

[5] I. Esnaola, S. M. Perlaza, H. V. Poor, and O. Kosut, "Decentralized maximum distortion mmse attacks in electricity grids," *INRIA, Lyon, Tech. Rep. 466*, Sep. 2015.

[6] J. F. Nash, "Equilibrium points in n-person games," *Proc. National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48–49, Jan. 1950.

[7] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124–143, May 1996.

[8] I. Shomorony and A. S. Avestimehr, "Worst-case additive noise in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3833–3847, Jun. 2013.

[9] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," in *Breakthroughs in Statistics*, Springer Series in Statistics, pp. 73–108. Springer New York, 1992.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Nov. 2012.

[11] K. Sun, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Information-theoretic attacks in the smart grid," in *Proc. IEEE Int. Conf. on Smart Grid Comm.*, Dresden, Germany, Oct. 2017, pp. 455–460.

[12] K. Sun, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Stealth attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1276–1285, Mar. 2020.

[13] J. Hou and G. Kramer, "Effective secrecy: Reliability, confusion and stealth," in *Proc. IEEE Int. Symp. on Information Theory*, Honolulu, HI, USA, Jun. 2014, pp. 601–605.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Mar. 2004.

[15] D. A. Bodenham and N. M. Adams, "A comparison of efficient approximations for a weighted sum of chi-squared random variables," *Stat Comput*, vol. 26, no. 4, pp. 917–928, Jul. 2016.

[16] Bruce G. Lindsay, Ramani S. Pilla, and Prasanta Basak, "Moment-based approximations of distributions using mixtures: Theory and applications," *Ann. Inst. Stat. Math.*, vol. 52, no. 2, pp. 215–230, Jun. 2000.

[17] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000.

[18] D. Hsu, S.M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electron. Commun. in Probab.*, vol. 17, no. 52, pp. 1–6, 2012.

[19] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*, CRC Press, Mar. 2004.

[20] J. J. Grainger and W. D. Stevenson, *Power System Analysis*, McGraw-Hill, 1994.

[21] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.

[22] D. Bodenham, *Momentchi2: Moment-Matching Methods for Weighted Sums of Chi-Squared Random Variables.* (2016) [Online]. Available: https://cran.r-project.org/web/packages/momentchi2/index.html.

[23] K. Sun, I. Esnaola, A. M. Tulino, and H. V. Poor, "Learning requirements for stealth attacks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Brighton, United Kingdom, 2019, pp. 8102–8106.

[24] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds., chapter 5, pp. 210–268. Cambridge University Press, Cambridge, UK, 2012.

[25] G. Alfano, A. M. Tulino, A. Lozano, and S. Verdú, "Capacity of MIMO channels with one-sided correlation," in *Proc. of IEEE 8th Int. Symp. on Spread Spectrum Techniques and Applications*, Sydney, Australia, Aug 2004.

[26] A. M. Tulino and S. Verdú, *Random Matrix Theory and Wireless Communications*, Now Publishers Inc, 2004.