Proceedings of Machine Learning Research 157, 2021 ACML 2021

Exposing Cyber-Physical System Weaknesses by Implicitly Learning their Underlying Models

Napoleon Costilla-Enriquez ncenriqu@asu.edu
Arizona State University, Tempe, AZ, US

Yang Weng yang.weng@asu.edu
Arizona State University, Tempe, AZ, US

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Cyber-Physical Systems (CPS) plays a critical role in today's social life, especially with occasional pandemic events. With more reliance on the cyber operation of infrastructures, it is important to understand attacking mechanisms in CPS for potential solutions and defenses, where False Data Injection Attack (FDIA) is an important class. FDIA methods in the literature require the mathematical CPS model and state variable values to create an efficient attack vector, unrealistic for many attackers in the real world. Also, they do not have performance guarantee. This paper shows that it is possible to deploy a FDIA without having the CPS model and state variables information. Additionally, we prove a theoretic bound for the proposed method. Specifically, we design a scheme that learns an implicit CPS model to create tampered sensor measurements to deploy an attack based only on historical data. The proposed framework utilizes a Wasserstein generative adversarial network with two regularization terms to create such tampered measurements also known as adversarial examples. To build an attack with confidence, we present a proof based on convergence in distribution and Lipschitz norm to show that our method captures the real observed measurement distribution. This means that our model learns the complex underlying processes from the CPSs. We demonstrate the robustness and universality of our proposed framework based on two diversified adversarial examples with different systems, domains, and datasets.

Keywords: Cyber-physical systems, False data injection attacks, Wasserstein GAN, Adversarial examples, Performance guarantee.

1. Introduction

The world changes dramatically with the data revolution worldwide and with more individuals working independently and remotely. Such a change requires better cyber security analysis, especially for the industrial sector with unprecedented integration of new real-time monitoring, sensing, control, and communication devices. Therefore, there is a heavy reliance on data-driven monitoring and control for modern Cyber-Physical Systems (CPSs), e.g., power systems, autonomous automobile systems, robotic systems, among others (Jazdi, 2014). However, such a data-driven outlook makes CPSs more vulnerable to attacks with potential dire consequences. Known as adversarial attacks (Kurakin et al., 2018), many studies show that outputs from classifiers can be easily changed with imperceptible changes to the input in the computer science domain (Carlini and Wagner, 2017; Szegedy et al.,

© 2021 N. Costilla-Enriquez & Y. Weng.

2013). In the CPS context, this is known as a False Data Injection Attack (FDIA), where an attacker intercepts and maliciously changes the system measurements to cause harm in the real world. For instance, a cyber-attack in a power system could cause a system operator to take wrong control actions causing a blackout. Similarly, a cyber-attack could cause fatal autopilot crashes in autonomous automobile systems (Banks et al., 2018).

To better protect various CPS systems, it is important to understand attack mechanisms with cutting-edge methods. While these cyber-attacks can cause dire consequences, such as cascading failure, they are hard to be deployed practically due to unrealistic settings or assumptions. For example, many FDIA attacks assume that the attacker has access to the entire CPSs information (e.g., topology, parameters, state estimator model, and estimated states, etc.) (Hug and Giampapa, 2012; Wang et al., 2020). But, many CPS systems, such as power systems, have their own code to avoid information leakage, e.g., build a local internet independent of the world wide web (www). There are also many security training and protocols to prevent system information from being available to the outsider. Therefore, there are approaches to reduce the required system information, e.g., only need partial system information (Hug and Giampapa, 2012). While these approaches work, they are still model-based FDIAs.

To resolve the issue, there are works independent of the system model. For example, Yu and Chin (2015) used principal component analysis, and Mohammadpourfard et al. (2020); Ahmadian et al. (2018) used a generative adversarial network to avoid system knowledge. However, these approaches do not present any convergence guarantee, which is an important metric to evaluate an attack's success. They also use linear system models, which is inaccurate, making attack detection easier from the defender side. Therefore, if an attacker has no access to the CPS model and information, can the attacker still initiate an attack confidently and without linear approximation? For example, it is quite practical for an attacker to intercept and alter sensor measurements. Although this idea of accessing only sensor measurements leads to a model-free framework, it is difficult to truly capture some properties of the system model to bypass tests, such as the Chi-squared test (Abur and Exposito, 2004).

Therefore, in this work we show that it is possible to construct an implicit data-driven CPS model that can mimic desired physics properties by learning measurement distribution. Specifically, we target to adapt Generative Adversarial Networks (GANs), where the Wasserstein Generative Adversarial Network (WGAN) is with the strongest mathematical properties for deriving our performance guarantee (Zhang et al., 2018). WGAN is a Deep Learning (DL) model that can learn the implicit training data distribution (e.g., sensor system measurements) so that we can sample from it and generate new data from that same distribution (Goodfellow et al., 2014a; Arjovsky et al., 2017). In our work, we will further integrate WGAN without needing system information so that the generative DL model in WGAN implicitly learns the underlying model that helps to pass system information-based tests.

To learn the system model explicitly and provide a performance guarantee, we design a novel WGAN architecture with a regularization term that incentivizes the generative network to create tampered measurements that will cause different estimated states than the real ones. To validate our proposed framework's contributions, we carry out simulations extensively on two diversified CPSs: the classic toy cart-pole system (Barto et al., 1983), and two power system grids (Zimmerman et al., 2011). Also, we compare our proposed model-free approach with the model-based method to deploy a FDIA proposed by Hug and Giampapa (2012). These results show that the proposed method achieves an effective FDIA without knowing the underlying CPS system information.

The contributions of this paper are: (1) We propose an innovative model-free framework to carry out an FDIA on CPSs. (2) We present a mathematical guarantee to show that our WGAN framework will converge to the true sensor system's measurement distribution, thus passing the defenses to detect bad measurement data. (3) We validate that our framework can effectively tamper measurements to deploy an FDIA. (4) We validate that the proposed model scales well to large CPSs.

2. Related Work

Adversarial attacks. Deep Convolutional Neural Networks (CNNs) models are highly susceptible to adversarial examples (Szegedy et al., 2013). Research on this topic has been done mainly for image classifiers and face recognition systems. These adversarial attacks or adversarial samples have drawn the community's attention. A lot of work to create adversarial samples has been done, including techniques such as the L-BFGS attack (Szegedy et al., 2013), FGSM (Goodfellow et al., 2014b), and the CW attack (Carlini and Wagner, 2016). In specific for classifiers, if the attacker knows the set of classes, Y, and the set of valid inputs, X, to the classifier.
False data injection in power systems. The work by Liu et al. (2011) firstly surprised the power and energy community by showing that a simple attack can be unobservable to existing tools in the field. There are various attack categorieswhich attracted great attention for the research community. Past work in the power systems area shows the vulnerability of the state estimation functionality (Mohammadpourfard et al., 2018).

To deploy FDIA, many papers have a very strong assumptions about the attacker's knowledge, which are unrealistic (Liu et al., 2011; Hug and Giampapa, 2012): (1) the attacker has access to the mathematical system model used to carry out the state estimation; (2) the attackers can intercept and alter the measurements used to obtain the estimated states in the system; and (3) the attacker can obtain the estimated states of the system. Under these strong assumptions, the attacker will be able to launch a perfect FDIA (Wang et al., 2020). To alleviate the strong assumption, Hug and Giampapa (2012) proposed an FDIA that can be deployed by having only partial system information. There are other types of FDIA. For example, a *generalized* FDIA evades residual-based bad data detection by exploiting small measurement errors tolerated by state estimation algorithms (Liang et al., 2017). However, all the methods above rely on system information.

3. Problem Formulation for State Estimation in CPS

The problem of inferring the state variables $\mathbf{s} = (s_1, \ldots, s_n)$ from a set of measurements $\mathbf{m} = (m_1, \ldots, m_k)$ is known as the State Estimation (SE) problem (Wood et al., 2013), where n is the number of state variables in the system, and k is the number of measurements. Mathematically, we can describe the problem as $\mathbf{m} = \mathbf{h}(\mathbf{s}) + \mathbf{e}$, where $\mathbf{h}$ is the physical (often non-linear) relationship between state variables and measurements, and $\mathbf{e}$ is a vector that represents white noise from the measurements. In practice, measurements are collected and used to obtain the estimated states, $\hat{\mathbf{s}}$, by solving the following optimization problem (Tarali and Abur, 2012): $\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \left(\mathbf{m} - \mathbf{h}\left(\mathbf{s}\right)\right)^T \mathbf{W}^{-1} \left(\mathbf{m} - \mathbf{h}\left(\mathbf{s}\right)\right)$.

The SE problem is commonly solved with the Weighted Least Squares (WLS) method. The vector of measurements, $\mathbf{m}$, might contain bad or wrong data due to telecommunication failures or meter errors (Abur and Exposito, 2004; Wang et al., 2020). To estimate the states with confidence, the SE possesses a Bad Data Detector (BDD) module or stage to detect and filter suspicious data.

Bad Data Detector (BDD): This BDD is based on the Chi-squared error test, also known as residual error test. Given that the measurement errors are assumed to follow a Gaussian distribution $e_i \sim \mathcal{N}(0, \sigma_i)$ (Abur and Exposito, 2004) (where σ_i is the standard deviation of the i-th measurement), the squared measurement residual error $\mathbf{r} = \|\mathbf{m} - \hat{\mathbf{m}}\|_2^2$ follows a Chi-square distribution χ_k, where k represents the number or independent variables in the CPS, and $\hat{\mathbf{m}} = \mathbf{h}\left(\hat{\mathbf{s}}\right)$ is the vector of estimated measurements. Then, the presence of errors in the measurements can be detected with the Chi-square test (or residual error test) (Abur and Exposito, 2004). This test works as follows. (i) Select the detection confidence probability p (e.g., 0.95), and compute its associated threshold value $\tau = \chi^2_{k,p}$ with $p = \Pr\left(\mathrm{J}\left(\hat{\mathbf{s}}\right) \leq \chi^2_{\mathrm{k,p}}\right)$. (ii) Compute the normalized measurement error $J(\hat{\mathbf{s}}) = \sum_{i=1}^{m} \left(m_i - h_i(\hat{s}_i)\right)^2 / \sigma_i^2$. (iii) If $J\left(\hat{\mathbf{m}}\right) \geq \tau$ holds, bad data will be suspected, else the measurements are assumed to be free of bad data.

4. Model-Free FDIA with a WGAN

Conventional model-based FDIA approaches rely upon the CPS model to deploy an FDIA (Liu et al., 2011; Hug and Giampapa, 2012). These methods need either all or partial knowledge of the targeted system to construct attack vectors successfully. These requirements are highly likely to be unrealistic under real-world settings. For instance, the owners of an autonomous automobile system will keep private the details of their CPS model. In this context, the attacker would not be able to deploy a model-based FDIA.

However, there are newer model-free, data driven methods (e.g., Yu and Chin 2015; Mohammadpourfard et al. 2020; Ahmadian et al. 2018). If the attacker chooses this option, he will not have a guarantee that the FDIA will be undetected. In this section, however, we show that for an attacker, it is possible to design a data driven FDIA without any underlying CPS knowledge, i.g., we learn an implicit version of the CPS model via data distribution. Specifically, after learning this distribution, the data driven mechanism will produce tampered measurements that look real but will pass the test by the BDD. The data driven approach is based on a WGAN with two regularization terms. First, the measurement distribution is learned with the WGAN, $\mathbf{m} + \mathbf{e}$. Second, a proxy of the unknown SE model is embedded into the WGAN as a regularization term, $\mathbf{h}\left(\mathbf{m}\right)$. This proxy will learn inferred states from the underlying system. Including these inferred states for attack model help to control the quality of data generation. Finally, a regularization term is added to maximize the attack impact. To ensure the confidence to system operators, in the next section, we will also give a proof of guarantee to show that the BDD will not notice such an attack.

4.1. Learning Measurement Distribution

In this subsection, we show how it is possible to implicitly learn the measurement distribution, $\mathbf{m} = \mathbf{h}\left(\mathbf{s}\right) + \mathbf{e}$, with a generative adversarial network. Goodfellow et al. (2014a) introduced the idea of generative adversarial networks, which revolutionized the machine learning (ML) field. GAN is a framework to teach a Deep Learning (DL) model the implicit training data distribution so that we can sample from it and generate new data from that same distribution. GANs are made of two distinct models, a generator and a discriminator. Formally, the minimax objective of the GAN is

$$\min_{G} \max_{D} \quad \mathbb{E}_{\mathbf{m}\sim\mathbb{P}_r}\mathbb{E}_{\boldsymbol{\lambda}\sim\mathbb{P}_\lambda}\left[\log D\left(\mathbf{m}\right) + \log\left(1 - D\left(G(\boldsymbol{\lambda})\right)\right)\right], \tag{1}$$

where D is a discriminative network, G is a generative network, $\mathbb{P}_r$ is the real data distribution, and $\boldsymbol{\lambda}$ is the latent space, which it is sampled from an independent distribution $\mathbb{P}_\lambda$; that is, $\boldsymbol{\lambda} \sim \mathbb{P}_\lambda$ (usually a Gaussian distribution).

However, GANs have some issues, such as vanishing gradient and the lack of guarantee to convergence. Arjovsky et al. (2017) presented the Wasserstein GAN (WGAN) that solves these issues. Also, WGANs possess stronger mathematical guarantees. For example, Zhang et al. (2018) proved that (under mild assumptions) the generator in the WGAN will converge to the true data distribution $\mathbb{P}_r$. Therefore, in this work, we will use this type of WGAN. The minimax objective of the WGAN is

$$\min_{G} \max_{D\in\mathcal{D}} \quad \mathbb{E}_{\mathbf{m}\sim\mathbb{P}_r}\mathbb{E}_{\boldsymbol{\lambda}\sim\mathbb{P}_\lambda}\left[D\left(\mathbf{m}\right) - D\left(G(\boldsymbol{\lambda})\right)\right], \tag{2}$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions (Arjovsky et al., 2017). In specific, the generator G learns the real distribution $\mathbb{P}_r$. In our context, this real distribution is the set of historical intercepted measurements $\mathcal{M} = \left\{\mathbf{m}_i \in \mathbb{R}^k\right\}_{i=1}^{L}$ (where L is the number of elements in the dataset), where $\mathbf{m}_i = \mathbf{h}\left(\mathbf{s}_i\right) + \mathbf{e}_i$. In other words, G is implicitly learns to generate samples, $\tilde{\mathbf{m}} = G(\boldsymbol{\lambda})$, from the underlying model $\mathbf{m} = \mathbf{h}\left(\mathbf{s}\right) + \mathbf{e}$.

4.2. Embedding a State Estimator Proxy

To gain trust from the power system operator, the created tampered measurements, $\tilde{\mathbf{m}} = G(\boldsymbol{\lambda})$, must pass the residual error test. This residual error for the tampered measurements is given as

$$\tilde{\mathbf{r}} = \left\|\tilde{\mathbf{m}} - \mathbf{h}\left(\tilde{\mathbf{m}}\right)\right\|^2, \tag{3}$$

where $\hat{\tilde{\mathbf{m}}} = \mathbf{h}\left(\tilde{\mathbf{m}}\right)$ is the vector of estimated tampered or fake measurements. The smaller the residual error $\tilde{\mathbf{r}}$, the bigger the probability of passing the test for a given tampered measurement, $\tilde{\mathbf{m}}$. In other words, a given vector of tampered measurements, $\tilde{\mathbf{m}}$, should produce a similar estimated vector, $\hat{\tilde{\mathbf{m}}} = \mathbf{h}\left(\tilde{\mathbf{m}}\right)$. This state estimator function $\mathbf{h}\left(\cdot\right)$ is unknown, but given its properties, it is possible to learn it from data and create a proxy to impose the same behaviour in the tampered measurements.

The residual error expression in eq. (3) resembles the loss function from an autoencoder (AE). Thus an AE model is a natural option to learn a proxy model of the unknown state estimator function $\mathbf{h}\left(\cdot\right)$. An autoencoder is a neural network that aims to produce or replicate its input to its output (Goodfellow et al., 2016). Mathematically, the autoencoder is represented as a function, that is, $\mathrm{AE}\left(\cdot\right)$, and it is trained with the squared loss function, that is $\left\|\mathbf{m}-\mathrm{AE}\left(\mathbf{m}\right)\right\|^{2}$. A trained AE with real measurements will learn the unknown function $\mathbf{h}\left(\cdot\right)$ that will minimize the residual error in eq. (3). Once the autoencoder is trained (denoted as AE^{*}), the AE loss function can be embedded into eq. (2) to incentive the generation of tampered measurements that will produce similar estimated measurements, and thus lowering the residual error. This can be done by adding the regularization term $\|\tilde{\mathbf{m}}-\mathrm{AE}^{*}\left(\tilde{\mathbf{m}}\right)\|_{2}^{2}$ in eq. (2):

$$\min_{G}\max_{D\in\mathcal{D}}\quad\mathbb{E}_{\mathbf{m}\sim\mathbb{P}_{r}}\mathbb{E}_{\boldsymbol{\lambda}\sim\mathbb{P}_{\lambda}}\Big[D\left(\mathbf{m}\right)-D\left(\tilde{\mathbf{m}}\right)+\|\tilde{\mathbf{m}}-\mathrm{AE}^{*}\left(\tilde{\mathbf{m}}\right)\|^{2}\Big].\tag{4}$$

4.3. Maximizing the FDIA Impact

The WGAN in eq. (4) implicitly learns the underlying model that generates the observed data (Goodfellow et al., 2014a; Arjovsky et al., 2017). To train a WGAN with eq. (4) the generator takes as input a random signal and maps it to the true data distribution space; that is, $\boldsymbol{\lambda}\sim\mathbb{P}_{\lambda}$, where $\mathbb{P}_{\lambda}$ is usually a Gaussian distribution. This means that we do not have any kind of control over the created fake measurements. To successfully deploy an attack on a CPS, these fake measurements, produced by our WGAN, must create different states from the actual ones. The attacker can only see and modify observed measurements. Thus, the only alternative to alter the unobservable states is to attempt to change intercepted measurements as much as possible. To accomplish this, we need to gain control over the generated measurements.

If we want to gain control over the generated measurements, rather than using a random distribution $\mathbb{P}_{\lambda}$ as latent space to feed our generator, we use the CPS measurements as input to the generator, that is, $\mathbb{P}_{\lambda}=\mathbb{P}_{r}$. In specific, we are conditioning the WGAN with respect to the actual measurement vector $\mathbf{m}$. This is desirable because it gives us control over the generated fake samples.

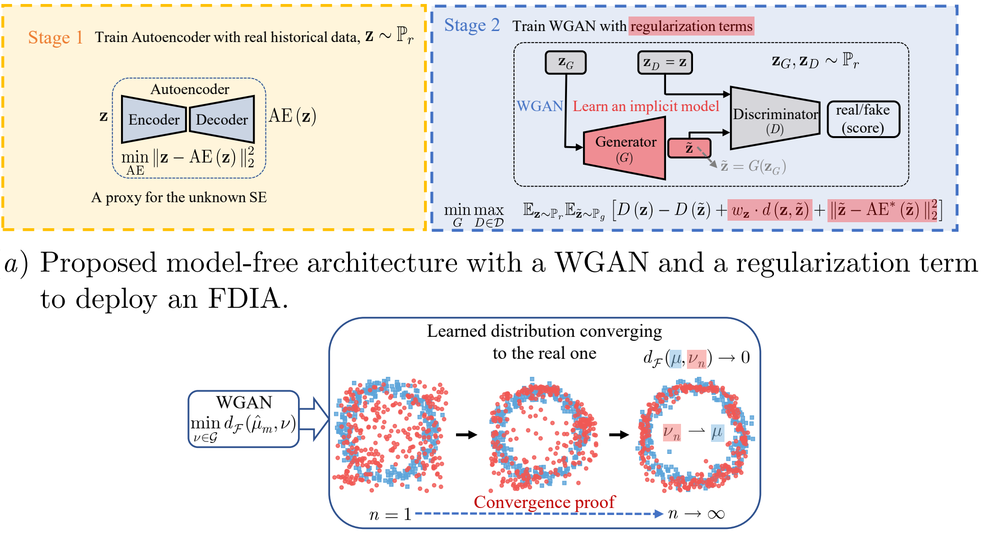
To successfully deploy an FDIA, we want to incentive the generator to construct measurements that will produce different measurements as the ones that receives as input. This will provoke, with high likelihood, that the SE will produce erroneous estimated states, which is the main objective in a FDIA. To accomplish this, we can incentive the model to generate such fake measurements with the regularization term $w_{\mathbf{m}}\cdot d\left(\mathbf{m},\tilde{\mathbf{m}}\right)$ in eq. (5), where $\tilde{\mathbf{m}}=G\left(\mathbf{m}\right)$, $d\left(\cdot\right)$ is a distance function (e.g., mean squared or mean absolute distance), $d\left(\mathbf{m},\tilde{\mathbf{m}}\right)$ represents the distance between the original measurement and the generated one, and $w_{\mathbf{m}}$ is a hyper-parameter that represents the weight of this distance. This regularization term incentives the WGAN to produce a tampered measurement vector $\tilde{\mathbf{m}}$ that will generate completely wrong estimated measurements.

$$\min_{G}\max_{D\in\mathcal{D}}\quad\mathbb{E}_{\mathbf{m}\sim\mathbb{P}_{r}}\mathbb{E}_{\tilde{\mathbf{m}}\sim\mathbb{P}_{g}}\Big[D\left(\mathbf{m}\right)-D\left(\tilde{\mathbf{m}}\right)+\|\tilde{\mathbf{m}}-\mathrm{AE}^{*}\left(\tilde{\mathbf{m}}\right)\|_{2}^{2}+w_{\mathbf{m}}\cdot d\left(\mathbf{m},\tilde{\mathbf{m}}\right)\Big].\tag{5}$$

Training the WGAN with regularization terms adds complexity to the training process. If the regularization term becomes too large with respect to the original WGAN loss, the generator will struggle to learn the right distribution. If the regularization term, on the other hand, is too small, it will not have any effect on the training process; thus, the regularization term will not fulfill its purpose. To solve this issue, a dynamic weight is introduced to control the size of $d(\mathbf{m}, \tilde{\mathbf{m}})$ throughout the training phase. This weight should be comparable to loss term associated with the fake sample, $D(\tilde{\mathbf{m}})$. We can express this as $|D(\tilde{\mathbf{m}})| = w_{\mathbf{m}} \cdot d(\mathbf{m}, \tilde{\mathbf{m}})$. Then, the result of such dynamic weight $w_{\mathbf{m}}$ is described in eq. (6) where $t > 1$ is the iteration number in the training phase. This dynamic weight adapts during training making easier and faster the model to converge.

$$w_{\mathbf{m}}^{(t)} = \frac{1}{2} \cdot \left| \frac{D\left(\tilde{\mathbf{m}}^{(t-1)}\right)}{d\left(\mathbf{m}^{(t-1)}, \tilde{\mathbf{m}}^{(t-1)}\right)} \right| \tag{6}$$

To summarize, our proposed architecture is shown in Fig. 1(*a*), where it can be seen that it is composed of two stages. First, an autoecoder is trained with historical measurement data. Second, the WGAN is trained with the same data and the two regularization terms: (1) one to maximize the impact of the attack and (2) another to incentive the WGAN to produce measurements that will pass the residual error test.


Stage 1
Train Autoencoder with real historical data, $\mathbf{z} \sim \mathbb{P}_r$
Autoencoder
$\mathbf{z}$
Encoder
Decoder
$\mathrm{AE}(\mathbf{z})$
$\min_{\mathrm{AE}} \|\mathbf{z} - \mathrm{AE}(\mathbf{z})\|_2^2$
A proxy for the unknown SE
Stage 2
Train WGAN with regularization terms
$\mathbf{z}_G$
$\mathbf{z}_D = \mathbf{z}$
$\mathbf{z}_G, \mathbf{z}_D \sim \mathbb{P}_r$
WGAN
Learn an implicit model
Discriminator (D)
real/fake (score)
Generator (G)
$\tilde{\mathbf{z}}$
$\tilde{\mathbf{z}} = G(\mathbf{z}_G)$
$\min_G \max_{D \in \mathcal{D}} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_r}\mathbb{E}_{\tilde{\mathbf{z}}\sim\mathbb{P}_g}\left[D(\mathbf{z}) - D(\tilde{\mathbf{z}}) + w_{\mathbf{z}} \cdot d(\mathbf{z}, \tilde{\mathbf{z}}) + \|\tilde{\mathbf{z}} - \mathrm{AE}^*(\tilde{\mathbf{z}})\|_2^2\right]$
WGAN
$\min_{\nu \in \mathcal{G}} d_{\mathcal{F}}(\hat{\mu}_m, \nu)$
Learned distribution converging to the real one
$d_{\mathcal{F}}(\mu, \nu_n) \to 0$
$\nu_n \to \mu$
Convergence proof
$n = 1$
$n \to \infty$

(*a*) Proposed model-free architecture with a WGAN and a regularization term to deploy an FDIA.

(*b*) Intuition for the WGAN convergence proof to the true observed distribution.

Figure 1: Architecture and convergence of our proposed method.

5. WGAN convergence with proposed regularization terms

In the last section, we presented a framework to create fake system measurements to deploy an FDIA. This framework consists of training a WGAN with a regularization term to control the output. This approach's main advantage is that we do not need confidential or inaccessible system information to build our model: We only need a set of historical measurements. However, to successfully deploy an FDIA without relying upon the CPS model and parameters, we need to be confident that our learned model will produce measurements that look legit so that the residual error test does not detect them. To show that our proposed framework truly captures the underlying measurement distribution, we present mathematical proof that certifies the WGAN convergence to the measurement distribution, thus creating fake measurements that will pass the residual error test.

Generative adversarial networks can be understood as minimizing a moment matching loss defined by a set of discriminator functions (Zhang et al., 2018), mathematically

$$\min_{\nu \in \mathcal{G}} \left\{ \begin{array}{l} d_{\mathcal{F}}\left(\hat{\mu}_m, \nu\right) := \\ \sup\limits_{f \in \mathcal{F}} \left\{ \mathbb{E}_{x \sim \hat{\mu}_m} \mathbb{E}_{\tilde{x} \sim \nu} \left[f\left(x\right) - f\left(\tilde{x}\right) + w_d \cdot d\left(x, \tilde{x}\right) \right] \right\} \end{array} \right\}, \tag{7}$$

where $\hat{\mu}_m$ is the empirical measure of the observed data (e.g., the CPS measurement distribution), and $\mathcal{F}$ and $\mathcal{G}$ are the sets of discriminators and generators, respectively. The practical WGANs take $\mathcal{F}$ as a parametric function class, that is, $\mathcal{F}_{nn} = \{f_\theta(x) : \theta \in \Theta\}$ where $f_\theta(x)$ is a neural network indexed by parameters θ that take values in $\Theta \subset \mathbb{R}^p$.

Notation and definitions. X denotes a subset of $\mathbb{R}^d$. For each continuous function $f : X \to \mathbb{R}$, we define the maximum norm as $\|f\|_\infty = \sup_{x \in X} |f(x)|$, and the Lipschitz norm $\|f\|_{Lip} = \sup\{|f(x) - f(y)| / \|x - y\| : x, y \in X, x \neq y\}$, and the bounded Lipschitz (BL) norm $\|f\|_{BL} = \max\left\{\|f\|_{Lip}, \|f\|_\infty\right\}$. The set of continuous functions on X is denoted by $C(X)$, and the Banach space of bounded continuous functions is $C_b(X) = \{f \in C(X) : \|f\|_\infty < \infty\}$.

Weak convergence. If $\mathcal{F}$ is discriminative, then $d_{\mathcal{F}}(\mu, \nu) = 0$ implies $\mu = \nu$. This means that the learned distribution is the same as the observed one. In reality, we cannot strictly get $d_{\mathcal{F}}(\mu, \nu) = 0$. Rather, we have $d_{\mathcal{F}}(\mu, \nu) \to 0$ for a sequence of ν_n and want to establish the weak convergence $\nu \rightharpoonup \mu$.

Theorem 1 *Let (X, d_X) be any metric space. If span$\mathcal{F}$ is dense in $C_b(X)$, we have $\lim_{n\to\infty} d_{\mathcal{F}}(\mu, \nu_n) = 0$ implies that the learned distribution ν_n weakly converges to the real observed distribution μ.*

In our context, the observed distribution μ corresponds to the set of CPS measurements $\mathcal{M}$. Fig. 1(b) gives the intuition for the convergence proof. First, the untrained generative network creates random samples (in red) without any structure, as shown in the left of Fig. 1(b). Then, after training the fake samples are converging to the true ring distribution (in blue), as shown in the right of Fig. 1(b). In other words, the learned distribution ν_n (in red) converges to the real one μ (in blue) as $n \to \infty$, which means that the WGAN is learning to create samples that look as taken from the true observed distribution μ.

Proof Given a function $g \in C_b(X)$, we say that g is approximated by $\mathcal{F}$ with error decay function $\epsilon(r)$ if for any $r \geq 0$, there exists $f_r \in span\mathcal{F}$ with $\|f_r\|_{\mathcal{F},1} \leq r$ such that $\|f - f_r\|_\infty \leq \epsilon(r)$. We note that $\epsilon(r)$ is a non-increasing function with respect to r. We know that the closure of $span\mathcal{F}$ is equal to the space of bounded continuous functions $C_b(X)$, that is, $cl(span\mathcal{F}) = C_b(X)$. Then, we have $\lim_{r\to\infty}\epsilon(r) = 0$. Now denote $r_n := d_F(\mu, \nu_n)^{-\frac{1}{2}}$, $f_n := f_{r_n}$ and $w_d = 1/r_n$. We have $|\mathbb{E}_\mu g - \mathbb{E}_{\nu_n} g| + w_d \cdot d(x, \tilde{x}) \leq |\mathbb{E}_\mu g - \mathbb{E}_\mu f_n| + |\mathbb{E}_\nu g - \mathbb{E}_\nu f_n| + |\mathbb{E}_\mu f_n - E_{\nu_n} f_n| + w_d \cdot d(x, \tilde{x}) \leq 2\epsilon(r_n) + r_n d_{\mathcal{F}}(\mu, \nu_n) + w_d \cdot d(x, \tilde{x}) = 2\epsilon(r_n) + 1/r_n + w_d \cdot d(x, \tilde{x})$. If $\lim_{r\to\infty} d_F(\mu, \nu_n) = 0$, we have $\lim_{r\to\infty} r_n = \infty$. Given that $\lim_{r\to\infty}\epsilon(r) = 0$, we prove that $\lim_{n\to\infty} |\mathbb{E}_\mu g - \mathbb{E}_{\nu_n} g| + w_d \cdot d(x, \tilde{x}) = 0$. Since this holds true for any $g \in C_b(X)$, we conclude that ν_n weakly converges to μ. If $\mathcal{F} \subseteq BL_C(X)$ for some $C > 0$, we have $d_{\mathcal{F}}(\mu, \nu) \leq C d_{BL}(\mu, \nu)$ for any μ, ν. Because the bounded Lipschitz distance metrizes the weak convergence, we obtain that $\nu_n \to \mu$ implies $d_{BL}(\mu, \nu_n) \to 0$, and $d_{\mathcal{F}}(\mu, \nu_n) \rightharpoonup 0$. ■

Theorem 1 guarantee us that the learned distribution ν by the WGAN (in eqs. (5) and (7)) will converge to the observed one μ. In other words, our model will create fake measurements that will pass the residual error test to deploy a FDIA in a CPS. This means that the WGAN captures the underlying system's interactions that produce the observed measurements. This idea is depicted in Fig. 1(b).

6. Experiments

This part will show how we deploy FDIAs on CPSs with our proposed WGAN framework without knowing their mathematical or physical model. To show the contributions and generality of our approach, we carry out extensive experiments on different types of CPSs. First, we train a WGAN with historical CPS measurements and demonstrate that the output of the WGAN converges to the true distribution of observed system measurements. We will show that the fake measurements will pass the residual error test. Second, we show that the trained WGAN creates different measurements (and therefore states) from the actual ones. This will show that the regularization term works and it is maximizing the FDIA impact. Finally, we show that our proposed framework is more reliable than the model-based ones by showing that our WGAN produces more realistic measurements. This implies that our model is capturing the underlying CPS model.

6.1. Generate Data for Evaluation

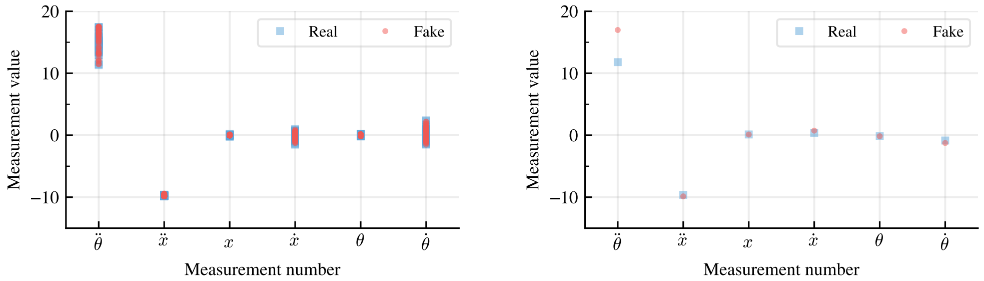
Cart-pole system. We simulate the physical system with the corrected nonlinear differential equations for the angular pole acceleration $\ddot{\theta}$ and cart acceleration $\ddot{x}$ for parameters reported by Florian (2007). This is a toy example with four state variables: cart position x, cart velocity $\dot{x}$, pole angle θ, and pole angular velocity $\dot{\theta}$. For the measurements, we suppose that our measurement vector is composed for six measurements as follows $\mathbf{m} = \left(\ddot{\theta}, \ddot{x}, x, \dot{x}, \theta, \dot{\theta}\right)$. We generate 1000 simulations by applying a random force F in the system. Note that we do not try to control the cart-pole; we only generate data from the physical model. We generate a set of real measurements $\mathcal{M} = \left\{\mathbf{m}^{(i)} \in \mathbb{R}^6\right\}_{i=1}^{1000}$, where the i-th measurement is $\mathbf{m}^{(i)} = \left(\ddot{\theta}^{(i)}, \ddot{x}^{(i)}, x^{(i)}, \dot{x}^{(i)}, \theta^{(i)}, \dot{\theta}^{(i)}\right)$.

Power Systems. We obtain the power systems' measurements by solving $8,760$ times (one year of hourly data) the AC power flow under different load conditions using MATPOWER (Zimmerman et al., 2011). We simulate the load fluctuation by multiplying each busload with the normalized load from the test power network RTS-GMLC (Preston and Barrows, 2018), which includes hourly load conditions for one year. This data generation approach will give us rich data variety with the power system under different load conditions at different seasons and hours. In specific, we present the IEEE 9-, 14-, 57-, 118-, and 300-bus test cases for the power system networks, and we consider all the active and reactive power flow measurements through transmission lines.

6.2. Validate on Learning the CPS Measurement Distribution

Cart-pole system. Fig. 2(a) shows 50 real samples and 50 fake ones. The fake samples overlap the real ones, which means that the generator learned the underlying cart-pole system measurement distribution. Furthermore, these fake samples pass the residual error test.

Power Systems. This part tests if the learned distribution by the WGAN converges to the true underlying power system measurement distribution. Fig. 3 shows 50 measurement samples from the real dataset and 50 created fake measurements for the 9-bus test case. We can see in Fig. 3(a) the generated fake measurements compared with real measurements from the training dataset; we can see that the fake measurements overlap the real ones, but they are not the same. In the same Figure, we can see that all the created fake measurements lie within the range of historical measurements that the WGAN has seen in the training process. Fig. 3(b) shows the associated probability to pass the Chi-squared test, that is $1-p$. This probability is associated to residual error of a specific set of measurements for the real (in blue) and fake (in red) measurements from Fig. 3(a). We can see in Fig. 3(b) that the real measurements have a 100% probability of passing this test; this is expected because these are perfect measurements, and the residual error is zero. The fake measurements' residual errors are not zero. Therefore, the associated probabilities are less than 100%, but all the fake measurements passed the residual error test. This means that the power system operator will accept these fake measurements.



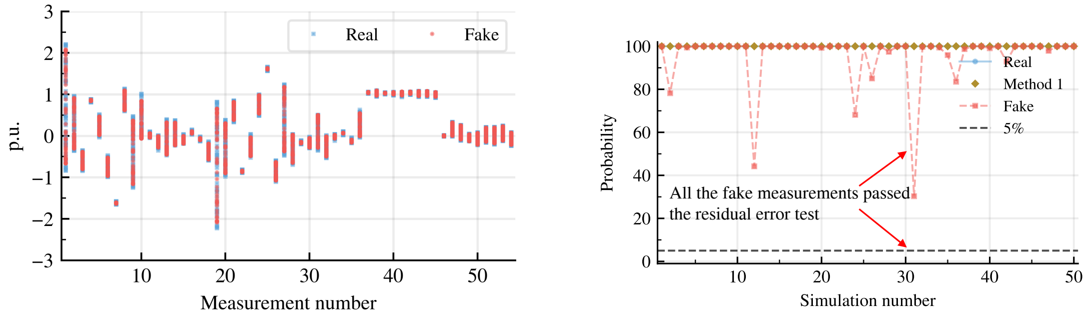
(a) Real and fake samples for the cart-pole system.

(b) The fake measurements pass the residual error test.

Figure 2: Results for the cart-pole system.

6.3. Validate the Deployment of FDIAs on CPSs

Deploying a FDIA in the Cart-pole System We train the WGAN and we use the hyper-parameters from Arjovsky et al. (2017): $n_{critic} = 5$, generator and discriminator learning rate $\alpha = 0.00005$, clipping parameter $c = 0.01$, batch size $b = 64$, and Adam adaptive learning algorithm (Kingma and Ba, 2014). Fig. 2(b) shows one specific real (in blue) and fake sample (in red). In the same Figure, we can see that the fake and real samples are similar but not the same. This means that the regularization term is working. The generator created this sample to maximize the impact of the FDIA. In other words, the created fake samples (1) look real and pass the residual test error, and (2) they are



(a) Real vs. fake measurements and states. Note that the fake measurements overlap the real ones, which means that the model learned to generate measurements that look real.

(b) Probability of fooling the state estimator. All the created fake measurements pass the 5% threshold $(1 - p)$ to fool the state estimator.

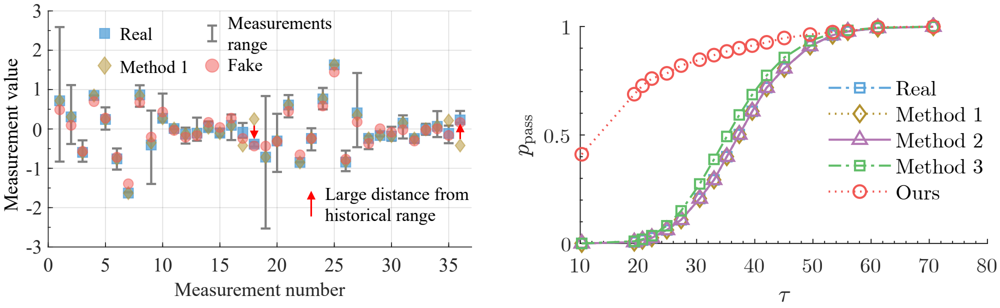
Figure 3: Real vs fake measurements and their associated probabilities of passing the residual error test for the 9-bus test case.

different from the real ones and have different associated states. These two traits indicate a successful FDIA carried out on the cart-plot system by our proposed approach.

Deploying a FDIA in Power Systems. Now, we will show how the fake measurements can be used to carry out a FDIA. We train the WGAN with the same parameters and procedure from the previous experiment. Fig. 4(a) shows a specific instance from Fig. 3(a) of a real measurement vector (in blue) and a created fake one (in red) for the 9-bus test network. The fake measurements are within the historical range from the dataset and look similar to the real ones. However, they produce significant changes in voltage magnitudes v and voltage angles δ with respect to the real states, as shown in Fig. 5(a). Furthermore, the fake measurements pass the residual error test, which means that the control center will not notice the FDIA. If the power system operator does not know the true power system's states, he will take wrong corrective actions that will cause issues in the network; for example, the transmission lines could overflow, leading to a blackout.

6.3.1. Comparison with Model-based FDIA.

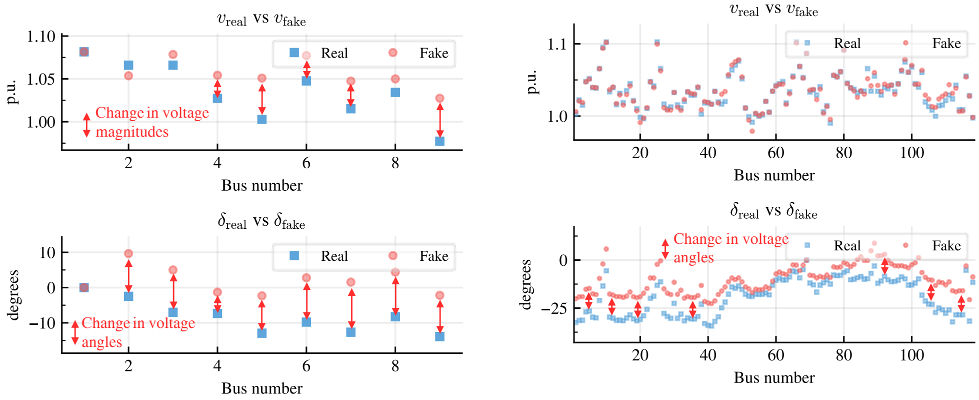
To assess the advantages and differences between our proposed model-free FDIA framework, we compare it against the model-based FDIA presented in Hug and Giampapa (2012)—we will refer to this FDIA as Method 1. We tamper 50 measurements with Method 1, and we present their associated probabilities (in brown) in Fig. 3(b). As we expected, the probabilities associated with the tampered measurements by Method 1 are 100%. This is because it has access to all the power system information and creates tampered measurements with zero residual error. However, tampering measurements to achieve zero residual error comes with a cost: The attacker has to alter some measurements to unrealistic values. Thus, the operator will realize about this odd, out-of-range value in the system.



(a) Comparison of the tampered measurements by the model-based Method 1 (Hug and Giampapa, 2012) with our model-free approach for the 9-bus test case.

(b) Comparison of probability of passing the residual error test with different decision thresholds τ for tampered measurements with different methods.

Figure 4: Example of a real vs a fake measurement for the 9-bus test cast.



(a) Results for the 9-bus test case.

(b) Results for the 118-bus test case.

Figure 5: Example of a real vs a fake measurement for the 9- and 118-bus test cases. Note that the fake measurements pass the residual error test.

To prove the last point, we perform the following experiment. We use Method 1 to tamper the state $v_5 = 1.05$ p.u. Fig. 4(a) shows the real measurements (in blue), the created tampered measurements by our proposed framework (in red), the created tampered measurements by Method 1, and the historical measurement range from the historical data from our dataset. In the same Figure, we see that the created measurements by our approach are within or very close to the historical range. In contrast, some tampered measurements by Method 1 are far away from the historical real measurements. In specific, we see in Fig. 4(a) that measurements 18 and 36 show a large distance from the historical range.

The key observation from the last point is: Even though Method 1 produces zero residual error measurements, these measurements will still look suspicious. The power system operator would realize that the tampered measurements 18 and 36 are outliers with respect to the historical ones, as shown in Fig. 4(a). The fake measurements obtained with our proposed framework do not have a zero error residual but lie within the historical range; thus, making them less suspicious for the power system operator.

Sensitivity Analysis. We also carried out a sensitivity analysis for different confidence values p. In this sensitivity analysis, we compare our method against three techniques in the literature: Method 1 introduced in Hug and Giampapa (2012), Method 2 from Chin et al. (2017), and Method 3 proposed in Liu and Li (2017).

To compare these methods, we made $1,000$ simulations with the same procedure previously described, and we tamper the real noisy measurements with our proposed approach and Methods 1, 2, and 3. For a given confidence value p, we compute its corresponding threshold $\tau = \chi^2_{k,p}$, and obtain the probability of each measurement to pass the residual error test for the specified threshold, that is, $\Pr\left(J(\mathbf{z}) \geq \tau\right)$. We repeat this process for each simulation and each aforementioned method, and we obtain the success rate of passing the residual error test. This is the probability of the simulations to pass the error test, and we call it p_{pass}. We do again this experiment for several values $p \in (0,1)$, and the result is shown in Fig. 4(b). We can see that as the threshold τ increases, the probability to pass the residual error test p_{pass} increases as well. Given that Methods 1 and 2 (in brown and purple, respectively) tampered the measurements such that the residual error is the same as the real one (in blue), they (almost) follow perfectly the real curve. Method 3 (in green) is close to the real curve, but just a little bit above. Our proposed approach (in red) consistently has larger probabilities of passing the residual error test: Even larger than the real measurements. Note that we trained our model with noisy measurements, and the method did not have access to the underlying power system model.

6.4. Validate Scalability and Sensitivity for CPS Systems

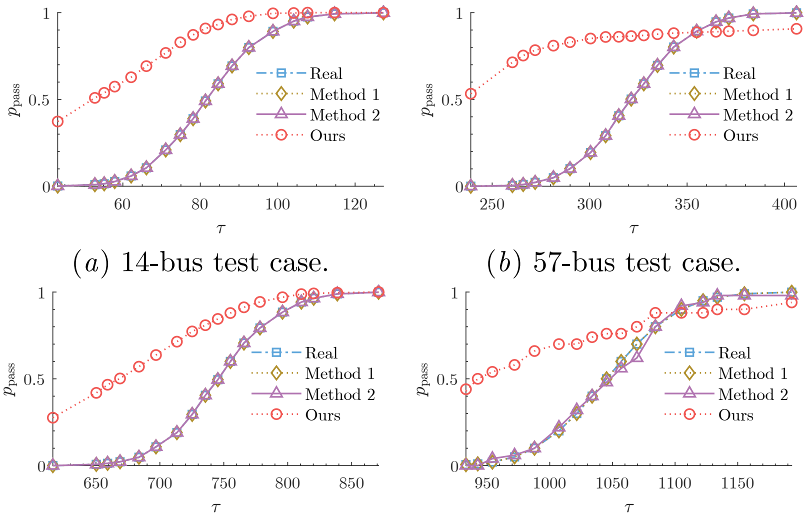
Finally, we show that our approach scales to bigger power system networks. To demonstrate it, we test our model-free FDIA on the IEEE 118-bus network, which has 744 measurements and 236 states. Fig. 5(b) shows an example of the real and fake states, where we can see that the created fake measurements provoke significant changes in the voltage angles with respect to the original ones. In addition, these created fake tampered measurements pass the residual error test. This means that the FDIA was successfully deployed.

Also, a sensitivity analysis, like the one in the previous section, is carried out for the IEEE 14-, 57-, 118, and 300-bus test cases, and the results are shown in Fig. 6. In the same Figure, we can see that our FDIA method outperforms the ones proposed in the literature.

Ablation study. We assess the impact of including the AE in our proposed scheme. We carried out $1,000$ simulations for each power network with and without an AE based on a confidence value $p = 0.95$. Table 1 shows the results where we can see that the addition of AE always improves the success rate of passing the residual error test for all the test cases.

Table 1: Impact of including an autoencoder.

Case	Success Rate (%)				
	9-bus	14-bus	57-bus	118-bus	300-bus
AE	95.5	95.7	89.3	97	91
No AE	63.7	81.1	61	54.33	70.6



(*a*) 14-bus test case.

(*b*) 57-bus test case.

(*c*) 118-bus test case.

(*d*) 300-bus test case.

Figure 6: CDF comparison for many test cases.

7. Conclusion

We presented an architecture to create adversarial sample measurement to carry out an FDIA in a CPS without knowing the underlying system information such as mathematical model or parameters. The architecture is framed into an optimization framework that considers the WGAN loss function and regularization terms to control the attack measurement vectors. We validated our proposed framework with two different CPS: the classic cart-pole system and two power system networks. We created fake measurements to create a bad data injection attack without access to the physical model details. These fake measurements passed the residual error test to detect bad data and gave completely wrong estimated state variables and measurements, which would compromise the reliability and operation of these CPSs. This work proves that it is not required for an attacker to access the underlying physical model.

References

Ali Abur and Antonio Gomez Exposito. *Power system state estimation: theory and implementation*. CRC press, 2004.

Saeed Ahmadian, Heidar Malki, and Zhu Han. Cyber attacks on smart energy grids using generative adverserial networks. In *IEEE Global Conference on Signal and Information Processing*, pages 942–946, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

Victoria A Banks, Katherine L Plant, and Neville A Stanton. Driver error or designer error: Using the perceptual cycle model to explore the circumstances surrounding the fatal tesla crash on 7th may 2016. *Safety science*, 108:278–285, 2018.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983.

Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

Wen-Long Chin, Chun-Hung Lee, and Tao Jiang. Blind false data attacks against ac state estimation based on geometric approach in smart grid communications. *IEEE Transactions on Smart Grid*, 9(6):6298–6306, 2017.

Razvan V Florian. Correct equations for the dynamics of the cart-pole system. *Center for Cognitive and Neural Studies, Romania*, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014a.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning (Adaptive Computation and Machine Learning series)*. The MIT Press, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

G. Hug and J. A. Giampapa. Vulnerability assessment of ac state estimation with respect to false data injection cyber-attacks. *IEEE Transactions on Smart Grid*, 3(3):1362–1370, 2012. doi: 10.1109/TSG.2012.2195338.

Nasser Jazdi. Cyber physical systems in the context of industry 4.0. In *International conference on automation, quality and testing, robotics*, pages 1–4. IEEE, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018.

G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, 8(4):1630–1638, 2017. doi: 10.1109/TSG.2015.2495133.

X. Liu and Z. Li. False data attacks against ac state estimation with incomplete network information. *IEEE Transactions on Smart Grid*, 8(5):2239–2248, 2017. doi: 10.1109/TSG.2016.2521178.

Yao Liu, Peng Ning, and Michael K Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(1):1–33, 2011.

M. Mohammadpourfard, A. Sami, and Y. Weng. Identification of false data injection attacks with considering the impact of wind generation and topology reconfigurations. *IEEE Transactions on Sustainable Energy*, 9(3):1349–1364, 2018. doi: 10.1109/TSTE.2017.2782090.

Mostafa Mohammadpourfard, Fateme Ghanaatpishe, Marziyeh Mohammadi, Subhash Lakshminarayana, and Mykola Pechenizkiy. Generation of false data injection attacks using conditional generative adversarial networks. In *PES Innovative Smart Grid Technologies Europe*, pages 41–45. IEEE, 2020.

E. Preston and C. Barrows. Evaluation of year 2020 IEEE RTS generation reliability indices. In *IEEE International Conference on Probabilistic Methods Applied to Power Systems*, pages 1–5, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

A. Tarali and A. Abur. Bad data detection in two-stage state estimation using phasor measurements. In *IEEE PES Innovative Smart Grid Technologies Europe*, pages 1–8, 2012. doi: 10.1109/ISGTEurope.2012.6465712.

Z. Wang, H. He, Z. Wan, and Y. Sun. Detection of false data injection attacks in ac state estimation using phasor measurements. *IEEE Transactions on Smart Grid*, pages 1–1, 2020. doi: 10.1109/TSG.2020.2972781.

Allen J Wood, Bruce F Wollenberg, and Gerald B Sheblé. *Power generation, operation, and control*. John Wiley & Sons, 2013.

Zong-Han Yu and Wen-Long Chin. Blind false data injection attack using pca approximation method in smart grid. *IEEE Transactions on Smart Grid*, 6(3):1219–1226, 2015.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

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 Transactions on Power Systems*, 26(1):12–19, 2011. doi: 10.1109/TPWRS.2010.2051168.