Reconstruction of Grid Measurements in the Presence of Adversarial Attacks

Amirmohammad Naeini
Dept. of Electrical Engineering and Computer Science
York University
Toronto, ON
amirda9@yorku.ca

Samer El Kababji
Faculty of Medicine
CHEO Research Institute
Toronto, ON
skababji@gmail.com

Pirathayini Srikantha
Dept. of Electrical Engineering and Computer Science
York University
psrikan@yorku.ca

Abstract

In efforts to mitigate the adverse effects of climate change, policymakers have set ambitious goals to reduce the carbon footprint of all sectors - including the electric grid. To facilitate this, sustainable energy systems like renewable generation must be deployed at high numbers throughout the grid. As these are highly variable in nature, the grid must be closely monitored so that system operators will have elevated situational awareness and can execute timely actions to maintain stable grid operations. With the widespread deployment of sensors like phasor measurement units (PMUs), an abundance of data is available for conducting accurate state estimation. However, due to the cyber-physical nature of the power grid, measurement data can be perturbed in an adversarial manner to enforce incorrect decision-making. In this paper, we propose a novel reconstruction method that leverages on machine learning constructs like CGAN and gradient search to recover the original states when subjected to adversarial perturbations. Experimental studies conducted on the practical IEEE 118-bus benchmark power system show that the proposed method can reduce errors due to perturbation by large margins (i.e. up to 100%).

1 Introduction

Climate change concerns are instigating major transformations in the way the modern power grid operates (1). These include the widespread integration of renewable generation systems, electric vehicles and storage systems along with ubiquitous connectivity enabled by information and communication technologies (ICTs) (2) (3) (4). Although sustainable power entities are highly variable in nature (e.g. generation by renewables), with the aid of ICTs, system operators will have real-time monitoring and actuation capabilities for maintaining stable grid operations. However, the ICTs are associated with inherent communication and/or software vulnerabilities that can be leveraged by adversaries to mislead system operators in making incorrect decisions and the triggering of control systems that can lead to cascading outages in the grid (5) (6).

Adversarial attacks on real power grids have taken place and these have been reported to have inflicted extensive losses and damages to the affected parties. For instance, cyber-attacks perpetuated in Ukraine in 2015 (7) and Venezuela in 2019 (8) resulted in extended power losses that lasted for days and affected hundreds of thousands of consumers. Another example is the security breach that occurred in Iran in 2011 (9). The Stuxnet worm was utilized to infiltrate Siemen's SIMATIC winCC

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

monitoring and data acquisition systems and inflict costly damages to nuclear centrifuges of power generation systems in Iran. During this period, this worm was estimated to have infected 60% of PCs in Iran. Hence, cybersecurity is an important consideration for the smooth operation of the electric grid especially with the proliferation of low-carbon technologies. As such, one common mode of attack is false data injection (FDI) where sensor measurements generated by phasor measurement units (PMUs) are perturbed. When these measurements are utilized to estimate grid states, incorrect values will be produced (9). Although mechanisms for detecting these perturbations have been built into traditional state estimation processes (e.g. residual based techniques (10)), a seminal work published in reference (11) demonstrated that stealthy attacks leveraging on the null-space of the transformation from measurements to states can bypass these safeguards. There have been many advances in FDI attacks henceforth that aim to stealthily affect grid states so that incorrect control actions can be taken (12).

As such, one important line of work that will reduce the impact of FDI attacks is the partial or full recovery of the original grid states when measurements are subjected to adversarial perturbations (e.g. (13)). The main issue with these is that details of the grid topology is necessary for the recovery process. This is confidential information that can result in dire consequences if leaked to adversarial entities. With our proposal in this paper, our contributions are: 1) We leverage a cycle GAN (CGAN) model that captures the non-linear mappings of grid measurements to states and vice versa for detecting specific perturbed measurements; 2) We utilize the gradients computed using the trained model to iteratively recover the perturbed measurements; and 3) We demonstrate the efficacy of the proposed reconstruction method on a benchmark IEEE 118 bus system.

2 Methodology

We utilize CGAN (14) which is a generative machine learning model to learn the forward mapping G between grid measurements y and states x and reverse mapping H vice versa.

$$y = G(x),\ x = H(y)$$

The general architecture of CGAN is illustrated in Fig. 1. The CGAN is composed of two GANs - one for each direction of mapping. GANs are generally composed of a discriminator and generator. The two generators in the CGAN aims to learn G and H respectively while the discriminators D_y and D_x aim to distinguish whether its inputs are from the actual training datasets or synthesized by its generator pair. For brevity, we assume that the trained CGAN model is already available.

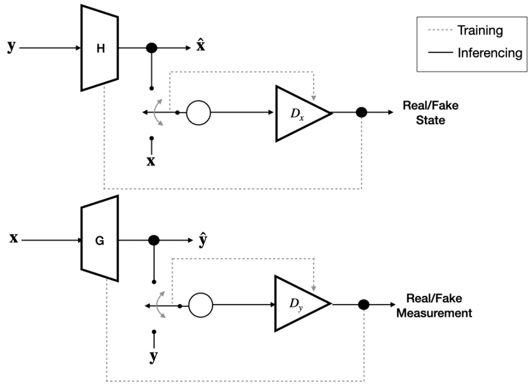

Figure 1: General architecture of CGAN.

We will utilize the discriminator D_y corresponding to the generator G to identify specific measurements that are perturbed. After training, the discriminator will output a probability lower than 0.5 if it deems its input (i.e. measurements) synthetic (i.e. perturbed). If this is the case, then, it is concluded that the measurements have been perturbed. To identify specific components of the residual that have been corrupt, the following residual vector is first calculated:

$$r = |y - G(H(y))|$$

If a component of the r vector is greater than the positive threshold α, then this component is flagged as perturbed. This is repeated for all components of the vector r. Once the perturbed components are all flagged, then just the perturbed components of y are updated in an iterative manner. To derive these updates, first the optimization problem $\mathcal{P}_{err}$ is formulated.

$$\mathcal{P}_{err}: \min_{y} ||y - G(H(y))||_2^2$$

where G and H are mappings that are already trained in the CGAN and $||.||_2$ is the second norm. $\mathcal{P}_{err}$ aims to minimize the error or the gap between the measurement and the forward and reverse transformation of y. The main idea here is that when the measurements are not perturbed, this error will be close to 0 when the mappings G and H have been trained until the training errors are low. When there are perturbations, the measurements will deviate from the original distribution of the grid actual measurements. This will lead to large gaps in the forward and reverse mappings. Hence, the goal of the reconstruction process is to minimize the objective outlined in $\mathcal{P}_{OPF}$. Directly solving this problem is not straightforward as G and H are complex neural networks (architecture is presented in the Appendix).

To solve this problem, we utilize a gradient descent based approach, where the gradient of the objective function which will be referred to as $f(y)$ is first computed using the chain rule.

$$\frac{\partial f}{\partial y} \& = -2 \cdot \Big(y - G(H(y))\Big)\Big(1 - \frac{\partial G(H(y))}{\partial H(y)} \frac{\partial H(y)}{\partial y}\Big)$$

The measurement vector is updated as follows based on the slightly modified gradient derived from empirical experiments:

$$y_{t+1} = y_t - 2\beta \operatorname{sgn}\Big(y_t - m_t\Big)\frac{\partial f(y_t)}{\partial y}$$

where t is the current iteration of update, m_t is the median value of that measurement in the training dataset and sgn is the sign function. All measurements that are not perturbed are replaced in this updated measurement vector. This process is repeated until the stopping condition is met. In this paper, the stopping condition is selected to be the upper limit T imposed on the number of iterations as information on the ground truth is not available during the reconstruction process. There is definitely room for improvement with the stopping condition and this will be discussed in detail in the next section. This algorithm is summarized in Alg. 1.

Algorithm 1 Reconstruction Algorithm

```
if $D_y < 0.5$ then
    for every attacked element of measurement (i.e. $r > \alpha$) do
        $t \leftarrow 0$
        for $t \leq T$ do
            Compute $H(y)$, $G(H(y))$ and $\frac{\partial f(y_t)}{\partial y}$
            $y_{t+1} = y_t - 2\beta \operatorname{sgn}\Big(y_t - m_t\Big)\frac{\partial f(y_t)}{\partial y}$
            Replace perturbed components of $y_0$ with the corresponding columns of $y_t$
            $y_{t+1} \leftarrow y_0$
            $t \leftarrow t + 1$
        end for
    end for
end if
```

With the proposed algorithm, it is clear that only information regarding the trained CGAN model is used to compute the reconstruction of the perturbed elements of the measurement vector. Furthermore, H and G capture the non-linear mappings between the measurements and states. Thus, the residual test will not be subjected to the same issues identified in reference (11) (i.e. attack perturbations that exist in the null-space).

3 Experiment

In this section, the performance of the proposed reconstruction algorithm of perturbed measurements is evaluated. The CGAN is trained to conduct state estimation in the benchmark IEEE 118 bus system.

Perturbations are added to randomly selected measurements. The perturbations can be as high as $\pm 10\%$ of the original values. Smaller perturbations are harder to detect and thus are utilized in these experiments to evaluate the discerning ability of the discriminator.

As such, the discriminator D_y is able to identify measurements that are perturbed with 100% accuracy. The residual limit selected is $\alpha = 0.988$. With this residual, the compromised components are also identified with 100% accuracy. Next, in Fig. 2, the evolution of the following two metrics are plotted over the update iterations $t = 0$ to $t = T$.

$$GHY = ||Y - G(H(y))||_2^2$$
$$GT = ||Y - GT||_2^2$$

The x-axis in this figure reflects the iteration number and y-axis represents the norm of error. In these results, 9 measurements were randomly selected and perturbed within $\pm 10\%$. It is clear that the gap between the ground truth and the reconstructed measurement vectors are decreasing in general throughout the iterative update process. It is clear that the general trends of GHY and GT are similar. Since GT is not known during the reconstruction process, the minimum value of GHY can be used as the stopping criteria. However, as GHY is not a non-increasing function, more analysis is required and this will be our future work. Fig. 3 illustrates the evolution of GT for every perturbed component.

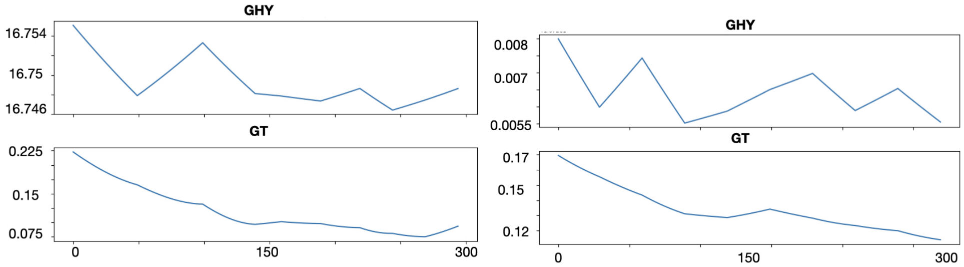

Figure 2: G(H(y)) and GT error for two different cases.

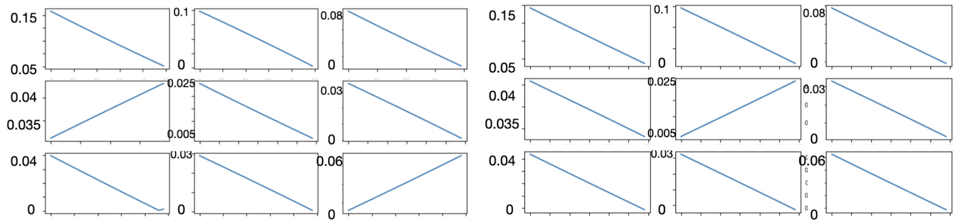

Figure 3: Component-wise ground truth error.

It is clear that the error mostly follows decreasing trends. One or two components are displaying increasing trends. Specifically, the norm of error is decreasing in both examples and the error from gt is reaching to zero for some components. Further analysis of the gradients associated with every component can reveal insights into this behaviour that can be utilized to prevent this trend. This is also future work.

4 Conclusion

In this paper, we have proposed a novel reconstruction algorithm for grid measurements that have been subjected to adversarial perturbations by leveraging on machine learning and iterative optimization constructs. The effectiveness of the proposed algorithm has been demonstrated for a practical IEEE 118 bus system. As future work, we intend to investigate how a good stopping criteria can be designed for the iterative updating algorithm so that the reconstructed grid measurement vector is as close as possible to ground truth. With detection and mitigation algorithms like this in place, the power grid can continue to accommodate highly variable green energy systems in a seamless manner even when subjected to adversarial interactions.

Appendix

4.1 Datasets

This recovery scheme is studied on the data generated for the IEEE 118-bus system (15). Gradients and chain rule has been computed in the TensorFlow (16) environment with implementation of tf.tape. The training and testing datasets have been generated using the Pandapower(17) module in Python.

4.2 CGAN Architecture

Here, the parameters of the CGAN model implemented in TensorFlow are presented for the various modules:

Grid State Generator Neural Network - H	
Input: 759 Nodes & L_1:512, L_2:1024, L_3:2048, L_4:1024, L_5:512 Output: 235	
Activation & relu, relu, relu, relu, relu, tanh	
Grid State Discriminator Neural Network- D_x	
Input: 235 Nodes & L_1:512, L_2:1024, L_3:256, L_4:64 Output: 1	
Activation & relu, relu, relu, relu, sigmoid	
Grid Measurement Generator Neural Network - G	
Input: 235 Nodes & L_1:512, L_2:1024, L_3:2048, L_4:1024, L_5:512 Output: 759	
Activation & relu, relu, relu, relu, relu, tanh	
Grid Measurement Discriminator Neural Network- D_y	
Input: 759 Nodes & L_1:512, L_2:1024, L_3:256, L_4:64 Output: 1	
Activation & relu, relu, relu, relu, sigmoid	

Table 1: Cycle GAN architecture.

4.3 Selection of α

For the selection of α, the f1-score is utilized. The following figure illustrates how the threshold and F1 are related. The threshold that results in the highest FI score is selected as α.

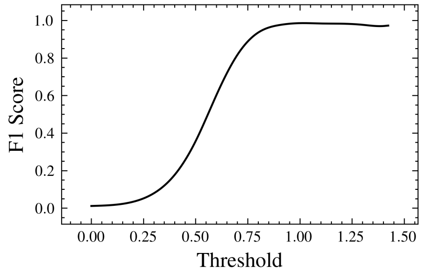

Figure 4: F1 score of tampered data detection

4.4 Selection of β

Next, we demonstrate the impact of β on the reconstruction of perturbed measurements. β is a learning parameter that can alter the convergence characteristics. If β is too small, the convergence will be too slow. On the other hand, if it is too fast, then the reconstruction process will oscillate. The following are some experimental examples of the impact of β.

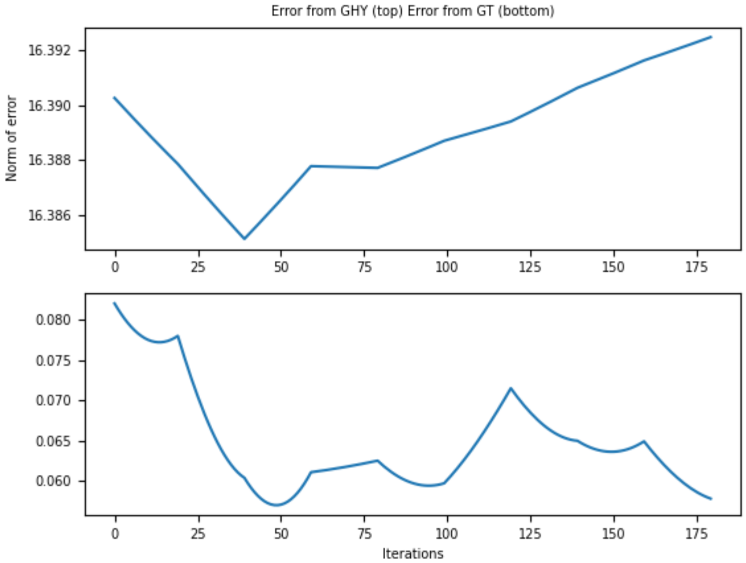

Figure 5: Beta = 0.001

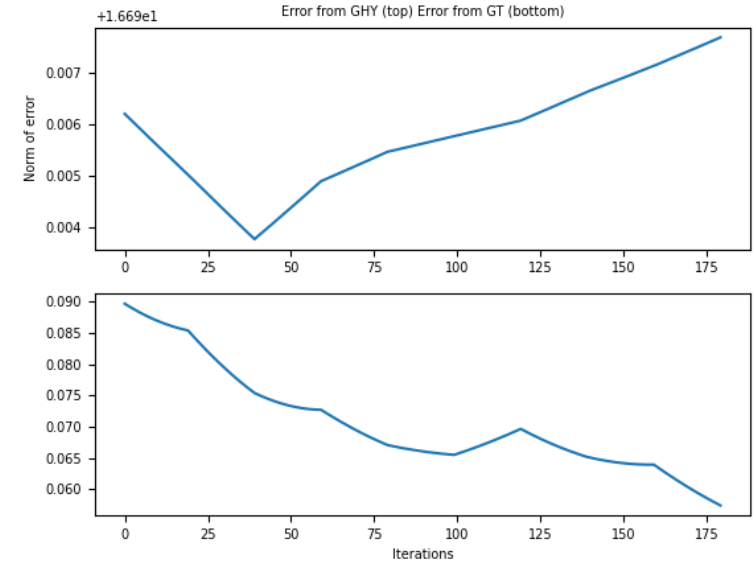

Figure 6: Beta = 0.0005

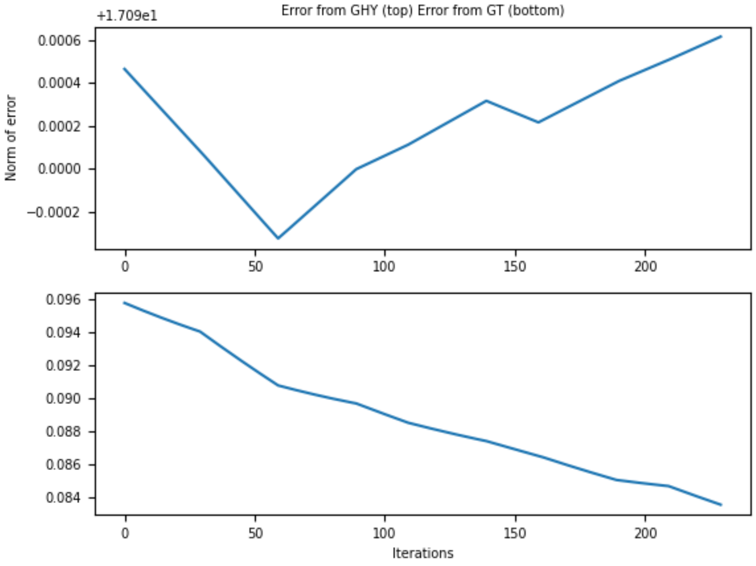

Figure 7: Beta = 0.0001

References

[1] Jerez, S., Tobin, I., Vautard, R., Montávez, J. P., López-Romero, J. M., Thais, F., ... & Wild, M. (2015). The impact of climate change on photovoltaic power generation in Europe Nat.

[2] Kurt, M. N., Yılmaz, Y., & Wang, X. (2018). Distributed quickest detection of cyber-attacks in smart grid. IEEE Transactions on Information Forensics and Security, 13(8), 2015-2030.

[3] Zhou, Y., Liu, Y., & Hu, S. (2017). Smart home cyberattack detection framework for sponsor incentive attacks. IEEE Transactions on Smart Grid, 10(2), 1916-1927.

[4] Zhou, Y., Chen, X., Zomaya, A. Y., Wang, L., & Hu, S. (2015). A dynamic programming algorithm for leveraging probabilistic detection of energy theft in smart home. IEEE Transactions on Emerging Topics in Computing, 3(4), 502-513.

[5] Deng, R., Xiao, G., Lu, R., Liang, H., & Vasilakos, A. V. (2016). False data injection on state estimation in power systems—Attacks, impacts, and defense: A survey. IEEE Transactions on Industrial Informatics, 13(2), 411-423.

[6] Li, Y., & Wang, Y. (2018). False data injection attacks with incomplete network topology information in smart grid. IEEE Access, 7, 3656-3664.

[7] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 ukraine blackout: Implications for false data injection attacks," IEEE Trans. Power Syst., vol. 32, no. 4, pp. 3317–3318, Jul. 2017.

[8] Vaz, Ricardo. "Venezuela's power grid disabled by cyber attack." Green Left Weekly 1213 (2019): 15.

[9] Mueller, P., & Yadegari, B. (2012). The stuxnet worm. Département des sciences de linformatique, Université de lArizona. Recuperado de: https://www2. cs. arizona. edu/ collberg/Teaching/466-566/2012/Resources/presentations/topic9-final/report. pdf.

[10] C. Liu, J. Wu, C. Long, and D. Kundur. "Reactance perturbation for detecting and identifying FDI attacks in power system state estimation." IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 4, pp. 763-776, 2018.

[11] Liu, C., Wu, J., Long, C., & Kundur, D. (2018). Reactance perturbation for detecting and identifying FDI attacks in power system state estimation. IEEE Journal of Selected Topics in Signal Processing, 12(4), 763-776.

[12] Tan, R., Nguyen, H. H., Foo, E. Y., Yau, D. K., Kalbarczyk, Z., Iyer, R. K., & Gooi, H. B. (2017). Modeling and mitigating impact of false data injection attacks on automatic generation control. IEEE Transactions on Information Forensics and Security, 12(7), 1609-1624.

[13] Li, Y., Wang, Y., & Hu, S. (2019). Online generative adversary network based measurement recovery in false data injection attacks: A cyber-physical approach. IEEE Transactions on Industrial Informatics, 16(3), 2031-2043.

[14] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[15] "Power Systems Test Case Archive - UWEE," Jan 2021, [Online; accessed 29. Jan. 2021]. [Online]. Available: `https://labs.ece.uw.edu/pstca/index.html`

[16] M. Abadi et al "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: `http://tensorflow.org/`

[17] L. Thurner et al, "pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems," IEEE Transactions on Power Systems, vol. 33, no. 6, pp. 6510– 6521, 11 2018.