A Hybrid Query-Efficient Black-Box Adversarial Attack on Power System Event Classifiers

Yuanbin Cheng, Koji Yamashita and Nanpeng Yu
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, California 92507 USA
ychen871@ucr.edu, kyamashi@ucr.edu, nyu@ece.ucr.edu

Yang Liu
School of Cyber Science and Engineering
Xi'an Jiaotong University
Xi'an, China
yangliu@xjtu.edu.cn

***Abstract*—The widespread deployment of advanced sensors in power transmission systems enabled machine learning (ML) models for event detection and classification. However, recent research reveals that ML models are vulnerable to adversarial attacks, which can fool the trained classifier by adding a small perturbation to the original data. Previous studies on adversarial attacks in power systems focused on the white-box configuration, which requires the attacker to have full access to the ML model's structure and parameters, which is unrealistic for real-world scenarios. This paper focuses on the black-box adversarial attack, where attackers only have access to the ML model's inputs and outputs. We propose a hybrid query-efficient black-box adversarial attack method that synergistically combines two types of attack algorithms: score-based and boundary-based attacks. Furthermore, within the score-based attack method of our hybrid method, we propose a query-efficient sign gradient estimation algorithm based on the binary representation of the coordinates through a batch of elaborate queries. Experimental results on a large-scale real-world PMU dataset and the state-of-the-art event classifier have validated the effectiveness of the proposed black-box adversarial attack method.**

***Index Terms*—Black-Box Adversarial Attack, Event Classification, Phasor Measurement Units, Power system.**

I. Introduction

The deployment of advanced sensors in the power system has empowered data-driven methods for event detection and classification [1]. A prime example is the phasor measurement units (PMUs) with high sampling frequency that can accurately measure voltage and current synchrophasors [2]. Over the past decade, PMUs have been widely deployed for power system's wide area monitoring system (WAMS) [3]. With a large amount of data from PMUs, sophisticated machine learning applications can be developed to improve the reliability of modern power systems.

In recent years, machine learning (ML)-based methods have demonstrated a high level of accuracy and computational efficiency in identifying and classifying transmission grid anomalies, including voltage, frequency, and oscillation events. To address the data quality issues of PMU raised in [4], several ML-based methods have employed feature extraction approaches, which include physics rule-based feature extraction [5], event pattern-based feature engineering [6], matrix decomposition [7], and energy similarity measurement [8]. Other methods focus on optimizing neural network structures to enhance the end-to-end power system event classification, such as convolutional neural network (CNN) [9], spatial pyramid pooling (SPP)-aided CNN [10], generative adversarial networks (GANs) [11]. The power system events are identified via an enhanced ResNet-50 model with information-loading regularization [12]. The hierarchical approaches for event classification that integrate multiple models have also been explored. For instance, [13] utilizes a hierarchical CNN model with channel filtering. [14] presents a refined two-level hierarchical CNN-based model.

Despite the advancements of ML-based methods in power systems, they often exhibit vulnerability to adversarial attacks. Such attacks inject small perturbations to the data, exploiting the inherent weaknesses of ML-based models to induce incorrect predictions [15]. Recent research by [16] demonstrates that adversarial attacks can easily sabotage ML-based event classifiers in the power system.

Adversarial attacks can be broadly divided into two categories, i.e., white-box attacks and black-box attacks. White-box attacks require knowledge of the neural classifier's architecture and trained parameters. Through back-propagation on the model, the gradient of the attacking objective function can be easily computed. However, white-box attacks are often limited in their practical application. This is because the attacker's access to the model's architecture and parameters, which, in many real-world scenarios, remain confidential and beyond the reach of potential adversaries. Conversely, black-box attacks do not require the aforementioned knowledge. Instead, they only need to access the ML models' inputs and outputs. Both types of attacks pose significant security risks to ML models, which highlights the critical need for defense strategies with sufficient robustness against such adversarial threats. This paper focuses on the black-box attacks on the ML-based power system event classifier models.

In this paper, we propose a query-efficient hybrid black-box attack algorithm that combines two categories of black-box attack methods: the score-based attack and the boundary-based attack. This hybrid algorithm partially mitigates the shortcomings of the existing methods, that is, trapped in a local minimum prematurely. Moreover, we propose a query-efficient sign gradient estimation algorithm that requires a much smaller number of the query to perform the score-based attack. The large-scale studies of black-box attacks on power system event classifiers using real-world PMU data verify that the proposed hybrid attacks can achieve faster and more effective attacks than the state-of-the-art (SOTA) methods. The contributions of this paper are listed as follows:

- We propose a two-step hybrid attack algorithm by synergistically combining score-based and boundary-based black-box adversarial attacks, which achieves better query efficiency and a higher attack success rate.
- We propose a query-efficient sign gradient estimation algorithm for score-based attack, which can efficiently estimate the sign of the classifier's gradient via a batch of designed queries.
- We demonstrate that the proposed method outperforms SOTA black-box adversarial methods on event classifiers using a large-scale real-world PMU dataset.

The rest of the paper is organized as follows. Section II provides the problem definition of adversarial attacks, and baseline black-box adversarial attack algorithms. Section III presents the proposed query-efficient sign gradient estimation algorithm and hybrid query-efficient black-box attack algorithm. Section IV compares the performance of the proposed method with multiple SOTA black-box attack algorithms. Section V concludes the paper and discusses future work.

II. Preliminaries

This section first introduces the key notations of this paper, including the samples of PMU time series, the dataset of the power system events, and the event neural classifier employed. Subsequently, it elucidates the definition of the adversarial attack on the neural classifier. Furthermore, the background of the black-box adversarial attack is introduced.

A. Key Notations

1) Notation 1: A sample from the PMU time series is formally defined as a tensor spanning a predetermined temporal window, expressed as $\mathbf{x} = [m_1, m_2, \cdots, m_W]$, where W denotes the total count of timestamps within the specified window. For each unique timestamp i, satisfying $1 \leq i \leq W$, the term m_i represents a measurement matrix that encapsulates various electrical variables collected from multiple PMUs, which include active power (P), reactive power (Q), voltage magnitude ($|V|$), and frequency (F).

2) Notation 2: Each PMU time series $\mathbf{x}$ is accompanied by an associated event label y, which signifies the specific type of event characterized by the sample. The construction of the power system event dataset thus involves aggregating pairs comprising PMU time series samples and their corresponding event labels. Formally, the dataset is represented as $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots\}$.

3) Notation 3: The ML-based power system event classifier is denoted by $f_\theta(\cdot)$, where θ represents the parameters of the classifier $f_\theta(\cdot)$. Given an input sample $\mathbf{x}$, the classifier $f_\theta(\cdot)$ computes the probability associated with each event type, yielding an output vector $\hat{y} = f_\theta(\mathbf{x})$.

B. Problem Definition

Adversarial Attack: Given a classifier $f_\theta(\cdot)$ and an original PMU sample $\mathbf{x}$, the adversarial attack aims to create an adversarial sample $\mathbf{x}' = \mathbf{x} + \delta$, where δ is a meticulously designed imperceptible perturbation vector. The goal of generating such an adversarial sample can be formally encapsulated as:

$$\arg\max_{\delta} L(f_\theta(\mathbf{x}+\delta), y), \text{ subject to } ||\delta||_2 \leq \epsilon, \tag{1}$$

where $L(f_\theta(\mathbf{x}+\delta), y)$ denotes the loss function of the classifier, specifically the cross-entropy loss between the predicted label $f_\theta(\mathbf{x}+\delta)$ and the true label y. The term $||\delta||_2$ represents the L_2 norm of the perturbation vector δ, and ϵ is a predetermined small constant that limits the perturbation's magnitude. This constraint is imposed to ensure that the adversarial sample is indistinguishable while still misleading the classifier.

The adversarial attack can be categorized into two primary types: white-box attack and black-box attack. The white-box attack assumes the attacker possesses complete knowledge and access to the target ML model, enabling direct calculation of the model's gradient. Representative white-box attack algorithms include the Fast Gradient Sign Method (FGSM) [15] and Projected Gradient Descent (PGD) [17], which leverage the sign of the gradient to generate perturbations, demonstrating a direct and potent method of attack.

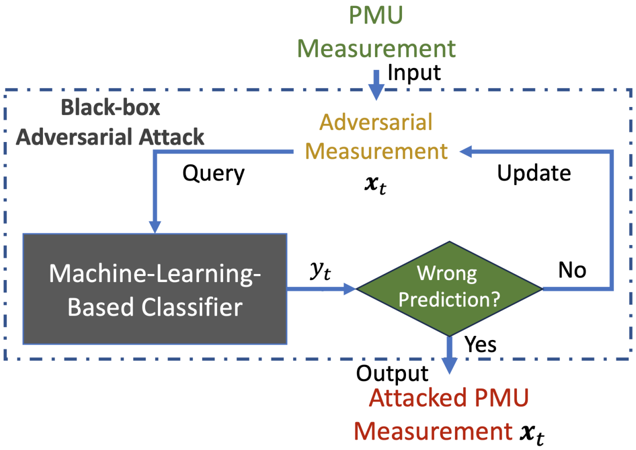

Fig. 1. The illustration of black-box attacks for PMU measurement.

However, the real-world scenarios often align more closely with the black-box attack, where the attacker is restricted to interacting with the model through its inputs and outputs, relying solely on the prediction $f(\mathbf{x})$ provided by the classifier. Under this paradigm, the attacker lacks insight into the model's internal architecture, weights, or training data.

Black-box attacks typically involve trial-and-error methods to craft adversarial examples. These methods include querying the model with varied inputs and observing the outputs to infer its behavior and formulate an effective attack strategy. This iterative process of refinement, depicted in Figure 1, involves adjusting the adversarial sample based on feedback from the model, aiming to induce misclassification by the model. Figure 2 illustrates a successful black-box attack on the neural classifier for power system events. The addition of a small perturbation fools the trained event classifier.

C. Black-box attack algorithms

The black-box attack can be further categorized into score-based attack and boundary-based attack. Here's a brief overview of algorithms in each category:

1) Score-based black-box attack: The score-based attack relies on the confidence scores predicted by the target ML model. These scores, which can be probabilities indicating how strongly the model believes its predictions are correct, serve as feedback for crafting adversarial examples. The attacker iteratively modifies the input data based on these scores to maximize the likelihood of incorrect model predictions.

Techniques like gradient estimation can be used, where the attacker estimates the gradient of the loss concerning the input by observing changes in the output score as the input is slightly varied. A common approach to estimating gradient involves numerical approximation, which can be mathematically represented as follows:

$$\nabla_{\mathbf{x}} L(\theta; \mathbf{x}, y) \approx \frac{1}{B} \sum_{b=1}^{B} [f(\mathbf{x} + \boldsymbol{\eta}^{(b)}) - f(\mathbf{x})] \cdot |\boldsymbol{\eta}^{(b)}|^{-1}, \quad (2)$$

where $\boldsymbol{\eta}^{(b)}$ are small perturbations, and the loss function calculation is repeated B times. The perturbation can be generated by different strategies. For instance, the Natural Evolutionary Strategies (NES) [18] samples $\boldsymbol{\eta}^{(b)}$ by $\beta \cdot \hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\eta}}$ is sampled from Gaussian distribution. Moreover, Liu et al. [19] extended signSGD to a zeroth-order setup with the ZOSignSGD algorithm and outperformed NES against a black-box model on the MNIST database. The ZoSignSGD algorithm approach uses the sign of the estimated gradient to achieve better convergence. These attacks, which utilize the sign of the estimated gradient, are also named sign-based attack algorithms, a subcategory of score-based attacks.

2) Boundary-based black-box attack: The core idea behind boundary-based attacks is to iteratively refine adversarial examples by exploring the target classification model's decision boundary, which refers to the hypersurface that separates different classes in the model's prediction space. By carefully perturbing input data points and observing changes in the model's output, attackers can identify points close to the decision boundary. These points are then used to generate adversarial examples—inputs, which are almost identical to legitimate examples but will be incorrectly classified by the model [20], [21]. One common technique used in boundary-based attacks is the binary search method, along with random perturbations, to efficiently search for adversarial examples close to the boundary.

III. Proposed Hybrid Query-Efficient Black-box Attack Algorithm

This section presents the proposed hybrid query-efficient black-box adversarial attack algorithm that combines the score-based and the boundary-based attack. We first present the proposed score-based attack with a more query-efficient sign gradient estimation algorithm, which we named BitSchedule. Then, we describe the two-stage hybrid black-box adversarial attack algorithm that combines the aforementioned score-based and boundary-based attacks.

A. A query-efficient score-based attack

In this subsection, we introduce a query-efficient, score-based attack algorithm, which is underpinned by a novel sign gradient estimation algorithm. Initially, we outline the overall framework of the sign-based black-box attack, followed by an in-depth exposition of the proposed query-efficient sign gradient estimation algorithm.

Algorithm 1 Sign-based black-box attack algorithm

```
Input: classifier $f_\theta$, data sample $\mathbf{x}$, learning rate $\alpha$
Output: Adversarial example $\mathbf{x}_{adv}$
Parameters: Perturbation bound $\epsilon$, Maximum iteration $M$
$\mathbf{x}_{adv} = \mathbf{x}$
$query\_cnt = 0$
while $query\_cnt < M$ do
    $\mathbf{g} = \text{SignGradientEstimate}(f_\theta, \mathbf{x}_{adv})$
    $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha \mathbf{g}$
    if $f_\theta(\mathbf{x}_{adv}) \neq f_\theta(\mathbf{x})$ then
        Success, stop the attack.
    end if
end while
return $\mathbf{x}_{adv}$
```

1) Framework of sign-based black-box attack: The overall framework of the sign-based attack algorithm is presented in Algorithm 1. This algorithm iteratively estimates the sign of the gradient of the loss function with respect to the input data and updates the sample by the sign of the estimated gradient, scaled by the learning rate until it misclassifies the sample or reaches the maximum query limit. The sign gradient estimation algorithm estimates the sign of the gradient by performing queries to the model. The sign gradient estimation algorithm is pivotal to the efficacy of the sign-based black-box attack.

B. BitSchedule: A query-efficient sign gradient estimation algorithm

As previously introduced, a critical aspect of sign-based black-box attacks involves estimating the gradient's sign by querying the model. Denote n as the dimensionality of a PMU time series sample $\mathbf{x}$. Estimating the sign gradient at point $\mathbf{x}$ typically requires performing gradient estimation across each coordinate and taking their sign, which requires $O(n)$ query operations. For example, a PMU sample comprising 360 timestamps, 40 PMUs, and 4 distinct measurements has $360 \times 40 \times 4 = 57,600$ dimensions, resulting in 57,600 queries, which is very time-consuming.

An alternative approach for sign gradient estimation involves applying a sign operator over the gradient estimation result, as elaborated in Eq. (2): $sign(\nabla_{\mathbf{x}} L(\theta; \mathbf{x}, y))$. Given that the sign-based optimization algorithm solely requires the

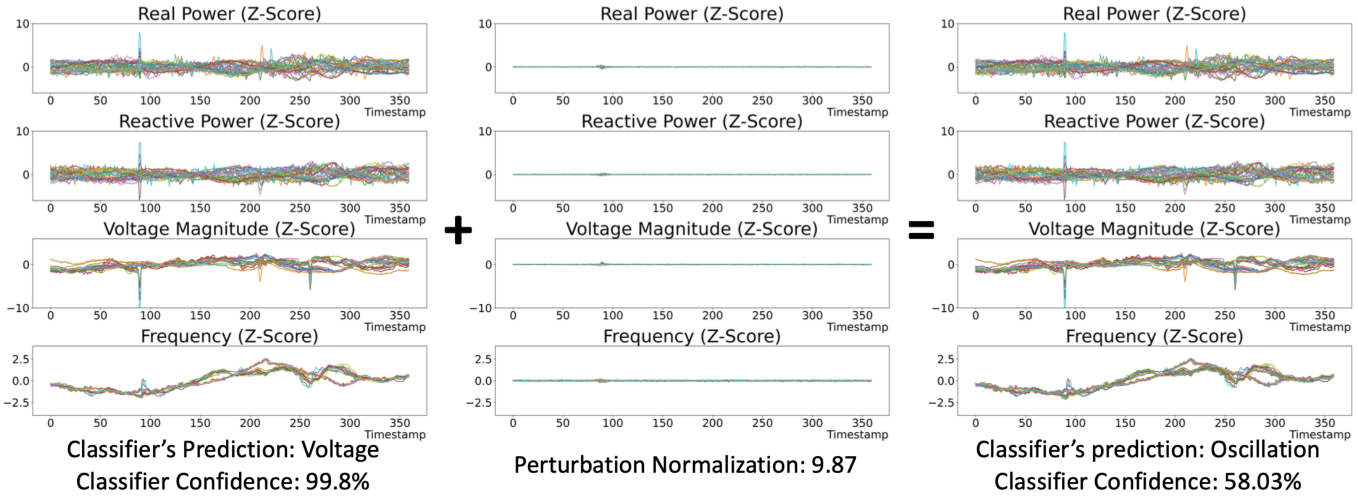

Fig. 2. Example of a successful black-box attack on PMU data: a tiny perturbation makes the model misclassify from voltage to oscillation event.

gradient's sign, maintaining magnitude variations across different coordinates in $\boldsymbol{\eta}^{(b)}$ introduces unnecessary complexity. Thus, we standardize the magnitude, focusing solely on sign information to enhance efficiency. Formally, this simplification is represented as follows:

$$\boldsymbol{\eta}^{(b)} = \delta \mathbf{v}^{(b)}, \text{ where } \mathbf{v}^{(b)} \in \{-1, +1\}^n, \tag{3}$$

with δ, a small constant set to 0.01 in this study.

Algorithm 2 BitSchedule: a sign gradient estimation algorithm

```
Input: Classifier $f_\theta$, PMU data sample $\mathbf{x}$
Output: Estimated sign($\nabla_{\mathbf{x}} L(\theta; \mathbf{x}, y)$)
$n$ = dimensionality of sample $\mathbf{x}$
$\mathbf{g} = \mathbf{0}_n$
for query = 1 to $\lceil \log_2(n) \rceil$ do
    $\mathbf{v} = \mathbf{0}_n$
    for coordinate = 1 to $n$ do
        if coordinate $\&(1 \ll query) \neq 0$ then
            $\mathbf{v}_{\text{coordinate}}$ = 1
        else
            $\mathbf{v}_{\text{coordinate}}$ = -1
        end if
    end for
    $\mathbf{g} = \mathbf{g} + (L(\theta; \mathbf{x} + \delta \cdot \mathbf{v}, y) - L(\theta; \mathbf{x}, y)) \cdot \mathbf{v}$
end for
return $sign(\mathbf{g})$
```

To reduce the number of queries required for sign gradient estimation, we introduce a method that needs only $O(\log_2(n))$ queries to estimate the sign gradient across each dimension. Assuming a total of B queries are performed, each query employs a direction vector $\mathbf{v}^{(b)} \in \{-1, +1\}^n$, indicating that each dimension's coordinate is assigned either $+1$ or -1. The strategy aims to ensure that no two coordinates, i and j, share identical direction settings across these B queries to avoid identical sign gradient estimations for these coordinates. Since there are two options for each coordinate, it is easy to derive that at least $B = \lceil O(\log_2(n)) \rceil$ queries are necessary to guarantee the above distinctiveness among coordinates.

To achieve this, our method employs binary bit encoding, preventing any two coordinates from having identical direction settings across all $\lceil \log_2(n)) \rceil$ queries. For any given coordinate i $(0 \leq i < n)$, we convert i into its binary representation $b_1 b_2 \cdots b_{\lceil \log_2 n \rceil}$, where each b_k $(1 \leq k \leq \lceil \log_2 n \rceil)$ is either 0 or 1. Consequently, for each $1 \leq k \leq \lceil \log_2 n \rceil$, we assign $\mathbf{v}_i^{(k)} = -1$ if b_k is 0, or $\mathbf{v}_i^{(k)} = +1$ if b_k is 1, in the corresponding k-th query.

This methodology is detailed in Algorithm 2, which is named "BitSchedule" due to its reliance on binary encoding for direction setting. Notably, to prevent overfitting and enhance generality during the iterative attack process, we shuffle all coordinates prior to each sign gradient estimation. Compared to the binary simultaneous perturbation stochastic approximation (BSPSA) [22], BitSchedule averages the $\log_2(n)$ designed perturbations instead of using two random symmetric Bernoulli distribution perturbations.

C. The proposed hybrid query-efficient black-box attack

In this subsection, we propose a hybrid query-efficient black-box attack, designed to combine the strengths of score-based and boundary-based attacks while mitigating their respective limitations.

Score-based attacks update the sample using the estimated gradient or its sign. However, such updates often converge to the local optima, as each iteration explores only a proximal area around the current point. Furthermore, boundary-based attacks require an initialization phase, which selects an optimal starting point from a set of randomly generated perturbations. This phase is critical to the attack's performance but requires a substantial number of queries, diminishing its success rate when there are strict query limitations.

To overcome these challenges, we introduce a hybrid attack algorithm that combines the merits of score-based and

Algorithm 3 Proposed hybrid black-box attack algorithm

```
**Input:** classifier $f_\theta$, PMU data sample $\mathbf{x}$, learning rate $\alpha$
**Output:** adversarial example $\mathbf{x}_{adv}$
**Parameters:** Perturbation bound $\epsilon$, limit for the number of query $M$, query limit in stage 1 $M_1$, query limit in stage 2 $M_2$, $(M = M_1 + M_2)$
$\mathbf{x}_{adv} = \mathbf{x}$
$\mathbf{x}_{adv} = \text{ScoreBasedAttack}(f_\theta, \mathbf{x}_{adv}, M_1)$
**if** $f_\theta(\mathbf{x}_{adv}) \neq f_\theta(\mathbf{x})$ **then**
    Successful in stage 1, return $\mathbf{x}_{adv}$
**else**
    $\mathbf{x}_{adv} = \text{BoundaryBasedAttack}(f_\theta, \mathbf{x}_{adv}, M_2)$
**end if**
**return** $\mathbf{x}_{adv}$
```

boundary-based attacks. This innovative two-step attach approach is detailed in Algorithm 3 that aims to enhance both query efficiency and the success rate of the attack.

Initially, the proposed query-efficient score-based attack is applied within a restricted query limitation. If this phase successfully fools the model, the attack is successful and finishes in this phase. If not, the process transitions to the second step, where the perturbation generated during the score-based phase serves as the initial perturbation for the boundary-based attack. This strategy effectively bypasses the boundary attack's initialization phase, leveraging the score-based attack's outcomes to reduce the total number of required queries and improve overall attack success rates.

IV. Numerical Study

In this section, we validate the proposed black-box adversarial attack algorithms with a large-scale real-world PMU dataset in the Western Interconnections of the United States. Six SOTA algorithms are used as baseline methods.

First, the PMU data used in this work is briefly described. Then, we introduce six baseline black-box adversarial attack algorithms. After that, we evaluate the performance of these six baseline methods and compare them with our proposed hybrid query-efficient black-box attack. Furthermore, the score-based method with our query-efficient sign gradient estimation method is validated and compared. Finally, we provide the success rate with different query limitations to quantify the efficiency of different attack algorithms.

A. Dataset and Target ML Event Classification Model

This study utilized two years of PMU data collected from the U.S. Western Interconnection in 2016 and 2017. The raw dataset comprises sequences of voltage phasors, current phasors, and frequency measurements. By following the methods outlined in [11, Section III-F], the raw data is cleaned and transformed into a structured tensor format, encapsulating four electrical variables: active power (P), reactive power (Q), voltage magnitude ($|V|$), and frequency (F). The preprocessing pipeline implemented several critical steps, including the removal of unreliable PMUs based on status flags or outlier thresholds and the imputation of missing data.

The power system event labels were derived from the event logs provided by the electric utility and network operators. The final dataset consists of 1,204 labeled PMU data samples, categorized into four distinct event types: 625 line events (characterized by voltage variations), 333 generator events (associated with frequency changes), 147 oscillation events, and 99 instances of normal system operation conditions.

Each data sample within the dataset corresponds to a 12-second observation window, with the PMUs recording data at a frequency of 30 Hz. Consequently, each PMU data sample is represented as a $360 \times 40 \times 4$ tensor, aligning with the number of timestamps (360), the number of PMUs (40), and the four electrical variables (P, Q, $|V|$, and F).

The ML-based classification model trained on this dataset serves as the target model of the adversarial attack, which is an enhanced ResNet-50 [23] neural network. This classifier utilized the graph signal processing-based PMU sorting algorithm and information loading-based regularization to achieve SOTA event classification performance [12].

B. Baseline Black-box Attack Algorithms

In this paper, we design a comprehensive benchmark containing six baseline algorithms: three score-based attacks and three boundary-based attacks. The following subsections provide a brief introduction to these baseline algorithms.

1) Baseline score-based attacks:

a) SimBA: The SimBA (Simple Black-Box Adversarial) [24] attack is a simple and effective score-based black-box adversarial attack on ML-based classification models. It involves perturbing the input image pixel by pixel or in small patches and observing the changes in the output. The direction of the perturbation is determined based on the model's output, aiming to maximize the error. The attack repeatedly makes small changes to the input until it finds the perturbation that can cause the model to misclassify the input.

b) ZoSignSGD: The ZoSignSGD [19] is a score-based black-box attack by using the stochastic gradient descent method. During the attack, instead of using the actual gradients, the sign of the gradients is used for the parameter updates to simplify the computation. In the ZoSignSGD attack, the attacker uses the zeroth-order optimization technique to approximate the signs of the gradients and then uses them to craft input data that can mislead the trained model.

c) SignHunter: The SignHunter [25] is a black-box adversarial attack algorithm that exploits the sign of gradients in the model's loss function. In simple terms, it focuses on the sign of the gradient (positive or negative) in the model's predictions rather than the magnitude of the change. Adjusting the sign of the gradient using a divide-and-conquer method, the SignHunter attack can effectively mislead the ML model.

2) Boundary-based attacks:

a) BoundaryAttack: The BoundaryAttack is a boundary-based black-box adversarial attack proposed in [21]. The boundary attack starts with an adversarial example already

TABLE I

SUCCESS RATE UNDER DIFFERENT QUERY NUMBER LIMITATION K, AND PERTURBATION MAGNITUDE LIMITATION ϵ

Queries Limitation K Target Distortion ϵ	**Simba**	**ZoSignSGD**	**SignHunter**	**BoundaryAttack**	**OPT**	**Sign-OPT**	**BitSchedule**	**Hybrid**
K = 10000, ϵ = 10	1.94%	5.82%	5.17%	3.66%	1.94%	3.66%	12.07%	**13.36%**
K = 20000, ϵ = 10	3.02%	9.27%	7.97%	4.31%	2.58%	7.32%	13.79%	**21.12%**
K = 30000, ϵ = 10	5.60%	10.56%	10.34%	5.81%	3.02%	12.71%	15.09%	**30.17%**
K = 40000, ϵ = 10	9.91%	13.79%	12.72%	6.68%	3.23%	16.38%	15.95%	**39.22%**
K = 10000, ϵ = 20	3.88%	20.47%	15.73%	7.11%	3.45%	25.43%	32.54%	**34.26%**
K = 20000, ϵ = 20	12.07%	28.87%	24.78%	15.51%	5.82%	49.35%	34.70%	**54.74%**
K = 30000, ϵ = 20	17.46%	32.33%	29.95%	24.14%	8.62%	59.48%	35.13%	**67.24%**
K = 40000, ϵ = 20	19.40%	34.05%	31.46%	30.39%	11.42%	62.71%	35.56%	**73.92%**

misclassified by the target model. It then iteratively refines it to reach a point close to the original data while maintaining its adversarial nature. The attack explores the classifier's decision boundary - that is, the region in the input space where the classifier's decision changes.

b) OPT Attack: The OPT Attack [26] is a boundary-based black-box adversarial attack. This method utilizes the zeroth-order optimization to minimize the distance between the original sample and the decision boundary. This method efficiently calculates the above distance using the binary search algorithm, reducing the number of require query for the attack. Instead of optimizing over the sample, this algorithm optimizes the direction of the perturbation to minimize the search space and increase the attack efficiency.

c) Sign-OPT Attack: The Sign-OPT [20] attack is another hard-label boundary-based black-box adversarial attack. This method notices that although the time complexity of the distance calculation to the boundary by binary search is $O(log(n))$, it is still very time-consuming. To increase the efficiency of the OPT attack, it proposed a Sign-OPT attack that does not need to recalculate the decision boundary distance every iteration. This method utilizes a method with $O(1)$ time complexity that can calculate the sign of the changes in the distance and use this sign information to decide the next iteration step. This method achieves greater query efficiency than the OPT attack.

C. Performance of Black-box Adversarial Attacks

Table I provides a comparative analysis of the efficacy of various black-box adversarial attacks under different constraints on the number of queries (K) and the L_2 norm of the perturbation (ϵ). The experimental setup across each row varies in terms of K and ϵ, offering insight into the impacts of the parameters on the success rates of different attack methods.

The score-based black-box attacks, including Simba, ZoSignSGD, SignHunter, and BitSchedule, are evaluated for their query efficiency and success rate under various configurations. The BitSchedule method emerges as the most effective one. It achieves higher success rates across different K and ϵ settings, which suggests its superior efficiency in utilizing queries compared to other score-based algorithms.

The boundary-based attacks, such as BoundaryAttack, OPT, and Sign-OPT, are evaluated under similar conditions. It is observed that these attacks demonstrate lower query efficiency, particularly under smaller query limitations. This reduced efficiency can be attributed to the initialization phase of these attacks, which involves generating and evaluating a batch of random perturbations to identify an optimal starting point for the attack, resulting in a higher number of queries. However, as the query number limitation increases, the Sign-OPT attack outperforms score-based methods. This advantage is likely due to the Sign-OPT approach's ability to circumvent local optima, thereby increasing the overall success rate.

The hybrid attack strategy, integrating the strengths of both score-based and boundary-based approaches, exhibits the highest success rate across all tested configurations of K and ϵ. This approach demonstrates exceptional query efficiency under limited query scenarios and maintains a high success rate even as the query and perturbation limitations are increased.

Figure 2 illustrates the effectiveness of the proposed hybrid attack on a particular PMU sample, that other baseline methods could not successfully attack under identical testing scenarios. This underscores the hybrid attack's superior query efficiency and success rate among black-box attack methods.

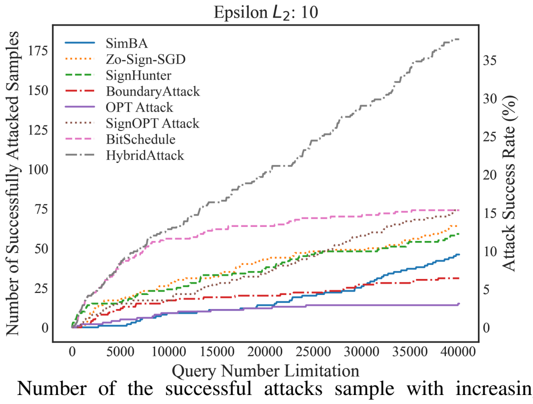

Fig. 3. Number of the successful attacks sample with increasing query limitation. ($L_2 = 10$)

D. Query-efficiency Comparison between Adversarial Attacks

Figures 3 and 4 present a detailed graphical analysis of how the success rates of various adversarial attack algorithms

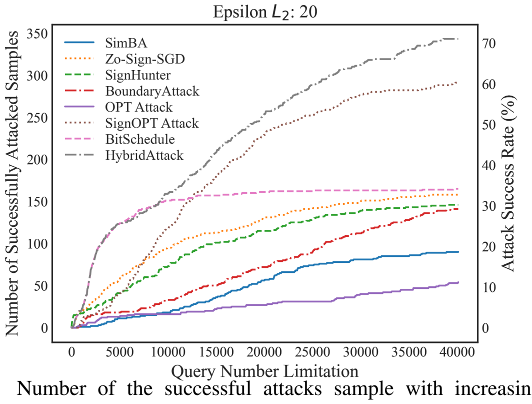

Fig. 4. Number of the successful attacks sample with increasing query limitation. ($L_2 = 20$)

evolve in response to changes in the query number limitation, ranging from 0 to 40,000, with intervals of 50 queries. These figures reveal that the proposed score-based algorithm, BitSchedule, exhibits superior query efficiency, particularly when the query number limitation is set below 10,000. This finding suggests that BitSchedule excels at achieving high success rates with a relatively low number of queries.

Furthermore, integrating score-based and boundary-based attack methods into the proposed hybrid attack algorithm demonstrates a significant advantage in both query efficiency and success rate as the limitations on query numbers and perturbation magnitudes are increased. This hybrid approach capitalizes on the strengths of both attack categories, leveraging the query efficiency of score-based algorithms in low-query scenarios and the high success rate of boundary-based algorithms when the query limitation increases.

V. Conclusion

This paper proposes a hybrid query-efficient black-box adversarial attack method to fool neural power system event classifiers. The proposed hybrid black-box adversarial attack combines the score-based attack and boundary-based attack for better query efficiency and attack success rate. Furthermore, this paper proposed a query-efficient sign gradient estimation algorithm called BitSchedule that queries with binary-encoded directions for each coordinate. This algorithm shows better query efficiency than other score-based methods. The empirical results on a large-scale real-world PMU dataset have shown that the proposed hybrid black-box adversarial attack method outperforms the other six SOTA black-box adversarial attack algorithms in query efficiency and attack success rate. In the future, our research will focus on the black-box attacks on other ML-based applications in power systems, alongside developing strategies to counteract these attacks to increase the robustness of the ML-based applications.

References

[1] J. Follum, "Real-time oscillation analysis: Technology readiness, and a vision for future needs and applications," Jun. 2020, Accessed: Sept. 20, 2023. [Online]. Available: https://www.naspi.org/sites/default/files/2020-07/20200624 NASPI Webinar - PJM ESAMS.pdf

[2] A. Monti, C. Muscas, and F. Ponci, *Phasor measurement units and wide area monitoring systems*. Academic Press, 2016.

[3] A. G. Phadke and T. Bi, "Phasor measurement units, WAMS, and their applications in protection and control of power systems," *J. of Mod. Power Syst. and Clean Energy*, vol. 6, no. 4, pp. 619–629, 2018.

[4] F. L. Grando, A. E. Lazzaretti, and M. Moreto, "The impact of PMU data precision and accuracy on event classification in distribution systems," *IEEE Trans. Smart Grid*, vol. 13, no. 2, pp. 1372–1382, 2022.

[5] A. Vosughi, S. Pannala, and A. K. Srivastava, "Event detection, classification and localization in an active distribution grid using data-driven system identification, weighted voting and graph," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 1843–1854, 2023.

[6] Y. Liu, L. Yang, A. Ghasemkhani, H. Livani, V. A. Centeno, P.-Y. Chen, and J. Zhang, "Robust event classification using imperfect real-world PMU data," *IEEE Internet Things J.*, vol. 10, no. 9, 2023.

[7] A. Vosughi, S. K. Sadanandan, and A. K. Srivastava, "Synchrophasor-based event detection, classification, and localization using koopman, transient energy matrix, best worth method, and dynamic graph," *IEEE Trans. Power Del.*, vol. 37, no. 3, pp. 1986–1996, 2022.

[8] R. Yadav, A. K. Pradhan, and I. Kamwa, "Real-time multiple event detection and classification in power system using signal energy transformations," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1521–1531, 2019.

[9] W. Wang, H. Yin, C. Chen, A. Till, W. Yao, X. Deng, and Y. Liu, "Frequency disturbance event detection based on synchrophasors and deep learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3593–3605, 2020.

[10] Y. Yuan, Y. Guo, K. Dehghanpour, Z. Wang, and Y. Wang, "Learning-based real-time event identification using rich real PMU data," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5044–5055, 2021.

[11] Y. Cheng, N. Yu, B. Foggo, and K. Yamashita, "Online power system event detection via bidirectional generative adversarial networks," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4807–4818, 2022.

[12] J. Shi, B. Foggo, and N. Yu, "Power system event identification based on deep neural network with information loading," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5622–5632, Nov. 2021.

[13] M. Pavlovski, M. Alqudah, T. Dokic, A. A. Hai, M. Kezunovic, and Z. Obradovic, "Hierarchical convolutional neural networks for event classification on PMU measurements," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[14] L. Zhu, D. J. Hill, and C. Lu, "Hierarchical deep learning machine for power system online transient stability prediction," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 2399–2411, 2020.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd Int. Conf. Learn. Represent., ICLR*, 2015.

[16] Y. Cheng, K. Yamashita, and N. Yu, "Adversarial attacks on deep neural network-based power system event classification models," in *2022 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia)*, pp. 66–70.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th Int. Conf. Learn. Represent., ICLR*, 2018.

[18] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. 35th Int. Conf. Mach. Learn., ICML*, vol. 80, 2018, pp. 2142–2151.

[19] S. Liu, P. Chen, X. Chen, and M. Hong, "signSGD via zeroth-order oracle," in *7th Int. Conf. Learn. Represent., ICLR*, 2019.

[20] M. Cheng, S. Singh, P. H. Chen, P. Chen, S. Liu, and C. Hsieh, "Sign-opt: A query-efficient hard-label adversarial attack," in *8th Int. Conf. Learn. Represent., ICLR*, 2020.

[21] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *6th Int. Conf. Learn. Represent., ICLR*, 2018.

[22] Q. Wang and J. C. Spall, "Discrete simultaneous perturbation stochastic approximation on loss function with noisy measurements," in *Proceedings of the 2011 American Control Conference*, 2011, pp. 4520–4525.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. Comput. Vis. and Pattern Recognit., CVPR*, 2016, pp. 770–778.

[24] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," in *Proc. 36th Int. Conf. Mach. Learn., ICML*, vol. 97. PMLR, 2019, pp. 2484–2493.

[25] A. Al-Dujaili and U. O'Reilly, "Sign bits are all you need for black-box attacks," in *8th Int. Conf. Learn. Represent., ICLR*, 2020.

[26] M. Cheng, T. Le, P. Chen, H. Zhang, J. Yi, and C. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in *7th Int. Conf. Learn. Represent., ICLR*, 2019.