# Vulnerability Under Adversarial Machine Learning: Bias or Variance?

Hossein Aboutalebi[1] , Mohammad Javad Shafiee[1], Michelle Karg[2] , Christian Scharfenberger[2] Alexander Wong[1]

[1] Waterloo AI Institute, University of Waterloo, [2] ADC Automotive Distance Control Systems GmbH, Continental

## Objectives

In this study, we investigate the effect of adversarial machine learning on the bias and variance of a trained deep neural network and analyze how adversarial perturbations can affect the generalization of a network. We derive the bias-variance trade-off for both classification and regression applications based on two main loss functions: (i) mean squared error (MSE), and (ii) cross-entropy. Finally, we introduce a new adversarial machine learning algorithm with lower computational complexity than well-known adversarial machine learning strategies (e.g., PGD) while providing a high success rate in fooling deep neural networks in lower perturbation magnitudes.

## Introduction

Despite of the impressive achievements of deep learning over the past decade in different fields such as computer vision, machine translation, and medicine, their vulnerability against adversarial machine learning brings different concerns regarding their robustness. A perturbation $\epsilon$ in a specific direction to the input causes the model to incorrectly classify the input sample which can be preformed in both classification or regression problems. Szegedy *et al.* introduced this drawback for deep neural networks in their seminal paper [1]. They observed that the state-of-the-art deep neural networks act poorly with high confidence when an imperceptible non-random perturbation is added to the input image. Although a rich literature developed in the field of adversarial machine learning, there has not been enough theoretical studies on why neural networks are vulnerable in facing inputs perturbed with adversarial perturbations.

## Contributions

- The bias and variance of a deep neural network facing adversarial perturbations is decomposed for both MSE and cross-entropy loss functions.
- The new derivations illustrate what should be the behavior of the adversarial attack to enforce the maximimum changes in the loss.
- Extensive experimental results validate the new theoretical findings in the network's bias and variance theorem for both MSE and cross-entropy loss functions.
- A new adversarial machine learning method (so-called BV adversarial attack) is proposed which is capable of fooling deep neural networks with comparable results with the state-of-the-art algorithms but with higher efficiency and less computational complexity.

## Theoretical Analysis

**Theorem 1**: The bias-variance trade off of a prediction model $\hat{f}(x)$ with training data $D$ for a target function $f(x)$ with noise $\epsilon$ in the presence of adversarial attacker that can inject noise of $\beta(x)$ to the system for MSE loss function is as follows:

$$\mathbb{E}_{x,D,\epsilon}\left[(y-\hat{f}(x+\beta(x)))^2\right] \approx \mathbb{E}_{x,D}[(f(x)-\bar{f}(x)-c_x)^2] \text{ -} \quad (1)$$

where $c_x = \nabla\bar{f}(x)^T\beta(x)$ and $c'_x = 2(\hat{f}(x) - \bar{f}(x))(\nabla\hat{f}(x) - \nabla\bar{f}(x))^T\beta(x)$.

**Theorem 2**: The bias-variance trade off of a prediction models $\hat{f}_0(x), ..., \hat{f}_m(x)$ with training data $D$ for a target function $f(x)$ with $m$ classes in the presence of adversarial attacker noise of $\beta(x)$ for classification with cross-entropy loss function is as follows:

$$\mathbb{E}_{x,D}\left[\sum_{i=0}^{c}\left(-\log(\frac{\hat{f}_i(x+\beta(x))}{\hat{f}_0(x\beta(x))+...+\hat{f}_m(x\beta(x))})\right)\mathbf{1}_{f(x)=i}\right] \approx$$
$$\sum_{i=0}^{c}\mathbb{E}_{x\in X_i,D}\left[-\log(\hat{f}_0(x))+\log(\hat{f}_0(x)+...+\hat{f}_m(x))+c_i(x \quad (2)$$

Where $c_i(x) = -\nabla log\left(\frac{\hat{f}_i(x)}{\hat{f}_0(x)+...+\hat{f}_m(x)}\right)^T\beta(x)$ and $X_i = \{x|x \in X \wedge f(x) = i\}$.

## BV Algorithm

```
Algorithm 1: BV Attack
Data: $\left\{(x, y(x))|x \in D\right\}$ with $c$ distinct classes
Result: $\hat{x}$ Perturbed image $x$.
Input:
$\hat{f}_i \quad i \in \{1, ..., c\}$, The prediction model scores
for all classes.
            $\epsilon$, The magnitude of perturbation.
            $x$, The input image.
            $y$, The ground truth label.
Begin
    $S = \left[-\nabla_x \log\left(\frac{\hat{f}_1(x)}{\hat{f}_1(x)+...+\hat{f}_c(x)}\right), ..., -\nabla_x \log\left(\frac{\hat{f}_c(x)}{\hat{f}_1(x)+...+\hat{f}_c(x)}\right)\right]$
    $V = \left[l_i|l_i = 1_{y=i}\ ,\ i = 1, \ldots, c\right]$
    $\hat{x} = x + \epsilon\ S^TV$
    Return $\hat{x}$
End
```

## Results

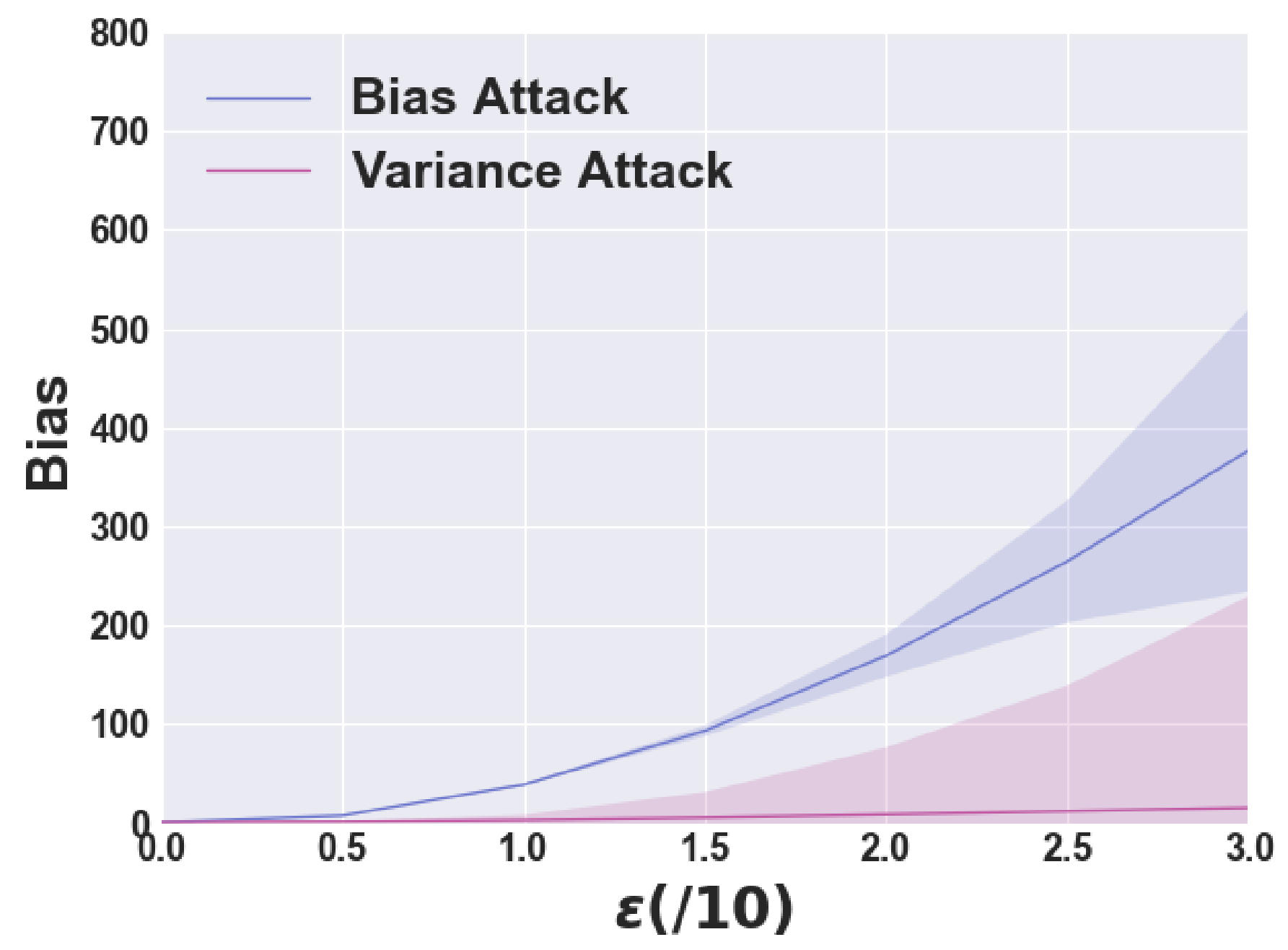

Figure 1:Effect of Bias and Variance attack

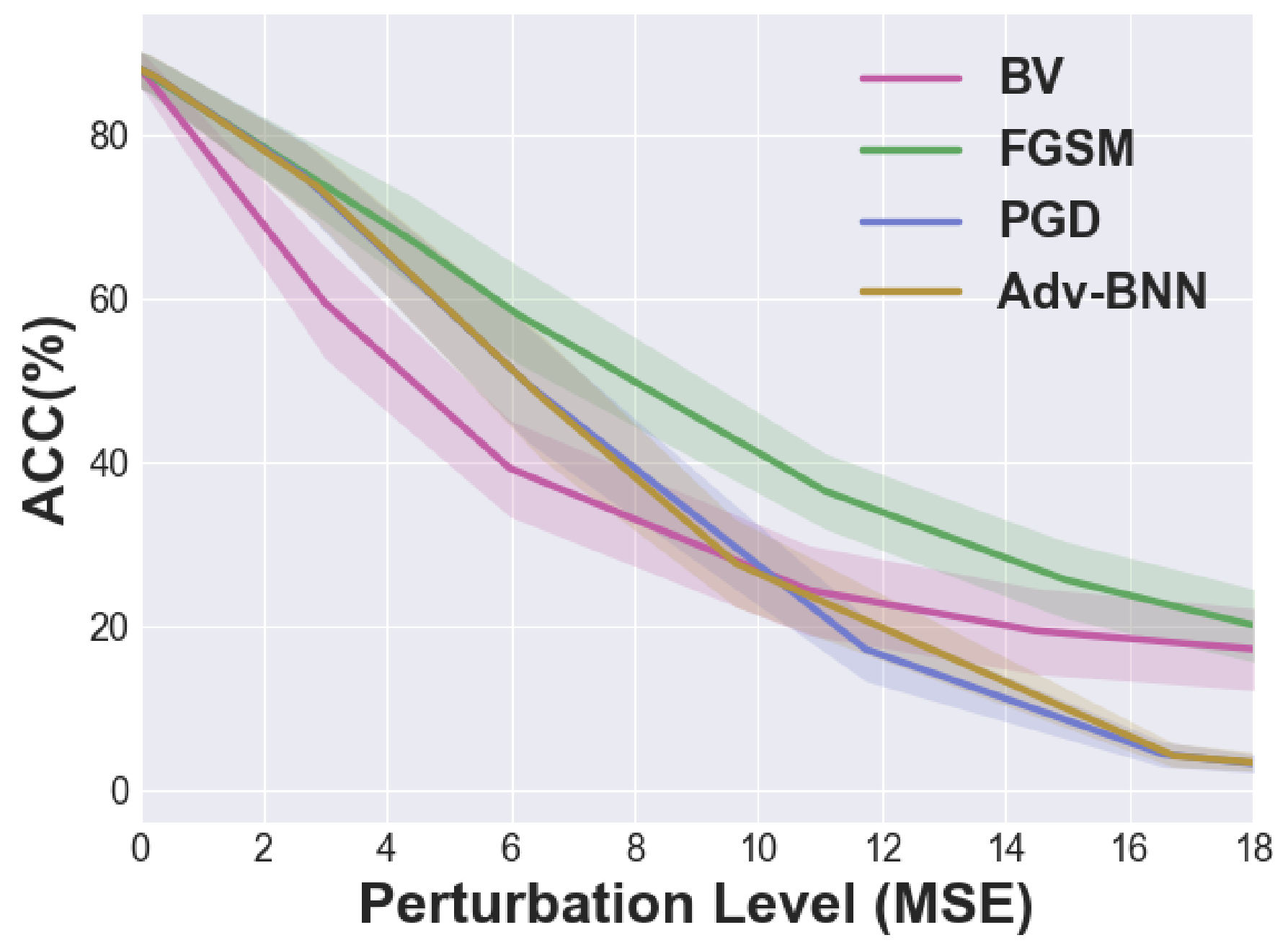

Figure 2:BV attack on ResNet50 compared with PGD, and

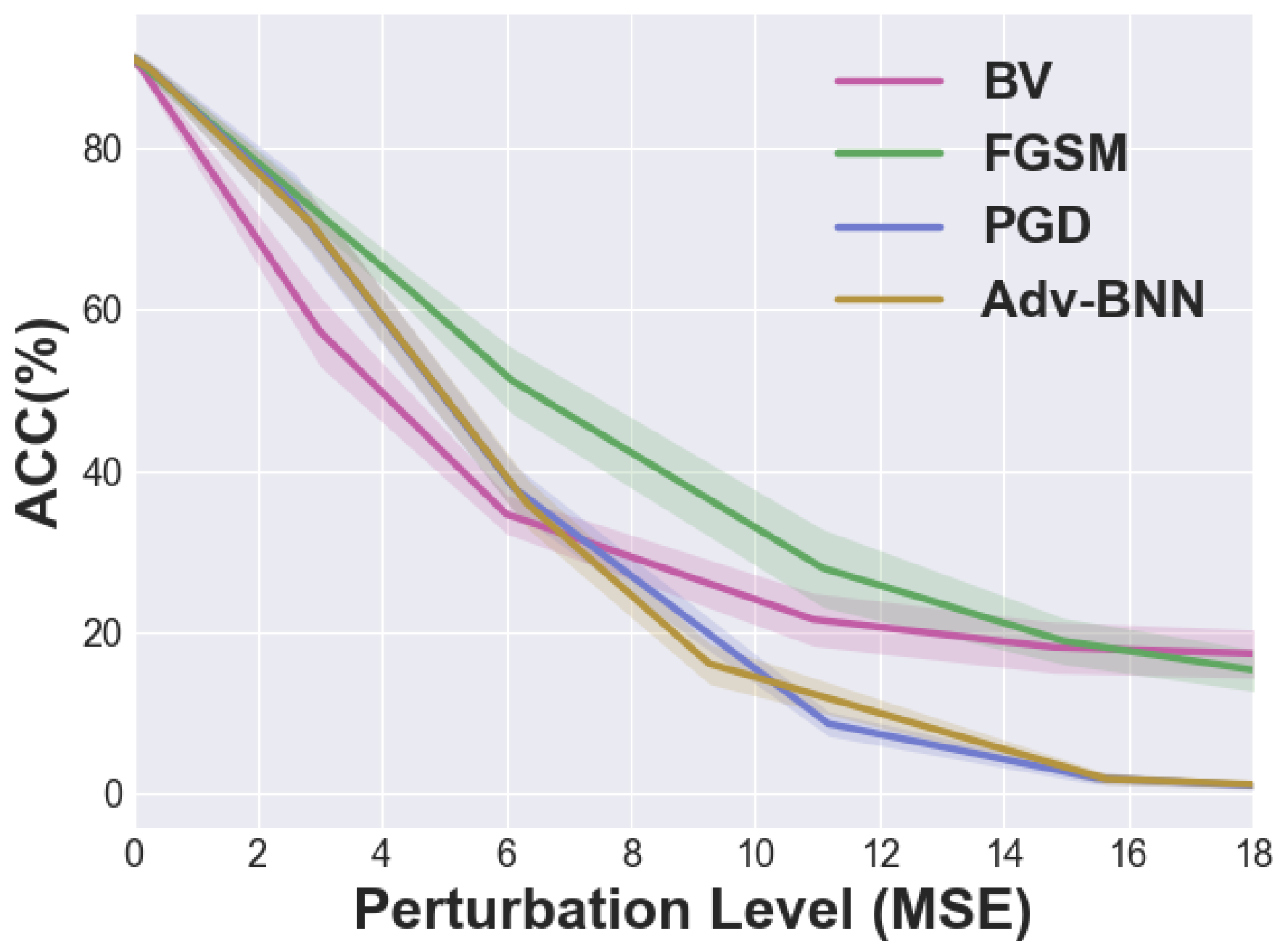

Figure 3:BV attack on ResNet34 compared with PGD, and FGSM

## Conclusion

In this work, we present a theory to help in understanding the impact of adversarial machine learning on both the variance and bias of the system, and for the first time illustrate how adversarial perturbations can manipulate the variance of the system besides its bias. We believe that these types of theoretical insights will give us a deeper understanding how these mechanisms lead deep neural networks to become vulnerable facing adversarial perturbation. Knowing how deep neural networks fail under adversarial machine learning will allow the community to discover new ways to defend against them and improve their robustness to such perturbations in order to build more reliable deep neural networks to use in real-world scenarios that impact society at large.

## References

[1] 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.

## Contact Information

- Web: https://uwaterloo.ca/vision-image-processing-lab/
- Email: haboutal@uwaterloo.ca