An Information Geometric Perspective to Adversarial Attacks and Defenses

1st Kyle Naddeo
Department of Electrical and Computer Engineering
Rowan University
Glassboro New Jersey, USA
naddeok5@rowan.edu

2nd Nidhal Bouaynaya
Department of Electrical and Computer Engineering
Rowan University
Glassboro New Jersey, USA
bouaynaya@rowan.edu

3rd Roman Shterenberg
Department of Mathematics
University of Alabama at Birmingham
Birmingham Alabama, USA
shterenb@uab.edu

***Abstract*—Deep learning models have achieved state-of-the-art accuracy in complex tasks, sometimes outperforming human-level accuracy. Yet, they suffer from vulnerabilities known as adversarial attacks, which are imperceptible input perturbations that fool the models on inputs that were originally classified correctly. The adversarial problem remains poorly understood and commonly thought to be an inherent weakness of deep learning models. We argue that understanding and alleviating the adversarial phenomenon may require us to go beyond the Euclidean view and consider the relationship between the input and output spaces as a statistical manifold with the Fisher Information as its Riemannian metric. Under this information geometric view, the optimal attack is constructed as the direction corresponding to the highest eigenvalue of the Fisher Information Matrix - called the Fisher spectral attack. We show that an orthogonal transformation of the data cleverly alters its manifold by keeping the highest eigenvalue but changing the optimal direction of attack; thus deceiving the attacker into adopting the wrong direction. We demonstrate the defensive capabilities of the proposed orthogonal scheme - against the Fisher spectral attack and the popular fast gradient sign method - on standard networks, e.g., LeNet and MobileNetV2 for benchmark data sets, MNIST and CIFAR-10.**

***Index Terms*—deep neural network, adversarial defense, information geometry**

I. Introduction

MODELS utilizing deep neural networks (DNNs) have become ubiquitous in the machine learning research community; yet, the same trend does not appear for critical tasks in industry. Although DNN models have achieved human-level accuracy in many tasks, including object recognition in computer vision and natural language processing, they are vulnerable to adversarial examples, i.e., imperceptible malicious input perturbations that cause well-trained state-of-the-art models to fail "with high confidence". Adversarial attacks were first discovered in computer vision [20] and later induced in natural language processing [7] and reinforcement learning [3] domains. More severely, it was found that adversarial examples have cross-model generalization ability, i.e., that adversarial examples generated from one model can fool another different model with a high probability [21]. Given such evidence, we can justly assume that all DNN models are vulnerable, and therefore the deployment of these models in real-world applications is hinged on the understanding and diminishing of the adversarial threat [4].

Currently, there is no consensus on why adversarial examples exist or how their transferability mechanism operates, leading to further difficulty in addressing the problem. There are, however, several speculative explanations to be considered. Starting with the seminal paper on the subject, it was suspected that the high non-linearity of neural networks could explain the existence of adversarial examples [20]. This was later refuted by Goodfellow *et al*. when it was shown that linear behavior in high-dimensional spaces is sufficient to cause adversarial examples [4]. They argued that adversarial examples are a result of models being too linear, rather than too non-linear. This view enabled the design of fast methods of generating adversarial examples. While the linear explanation may seem possible, it only analyses the decision boundary under the first-order approximation of the model non-linearity. A further investigation by Moosavi-Dezfooli *et al*. derived a more generic analysis in terms of the geometric properties – notably curvature - of the boundary [13]. Specifically, they showed that classifiers with curved decision boundaries are vulnerable to malicious perturbations, and provided the existence of a shared subspace along which the decision boundary is positively curved (for most directions).

The explanations thus far are limited to the Euclidean metric and only analyze the decision boundary. An alternative view analyzes the relationship between the input and output spaces, which map vectors of real numbers to probability distributions over classes. From an information geometric perspective, this

This work was supported by the National Science Foundation Award NSF ECCS-1903466. We are also grateful to UK EPSRC support through EP/T013265/1 project NSF-EPSRC: "ShiRAS: Towards Safe and Reliable Autonomy in Sensor Driven Systems". Kyle Naddeo is supported by the U.S. Department of Education Graduate Assistance in Areas of National Need (GAANN) Program.

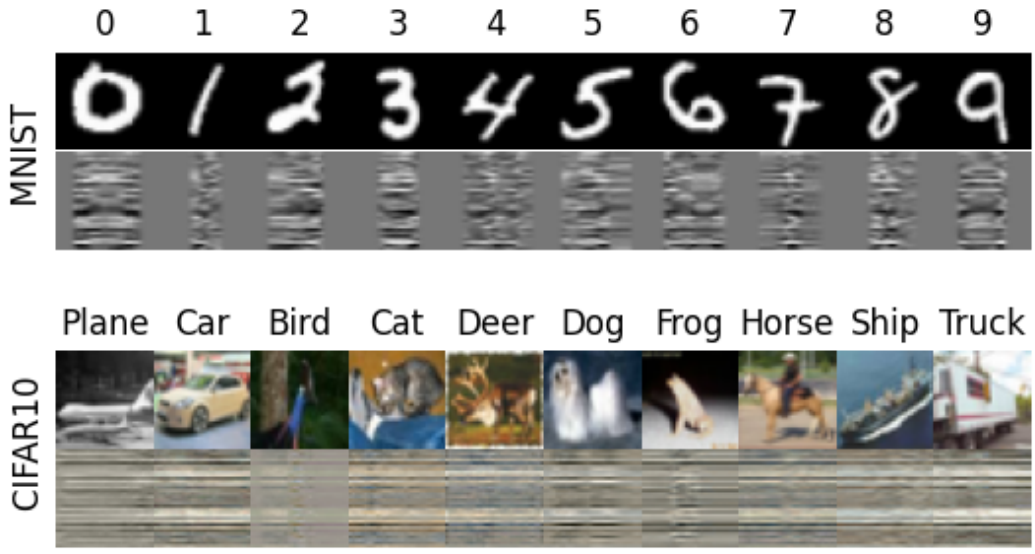

Fig. 1. Illustration of an orthogonal transformation on images from MNIST and CIFAR-10. The resulting images are unrecognizable to the human eye but are congruent to the original image.

relationship can be defined as a statistical manifold. The inputs represent coordinate points on the manifold while the outputs determine the Fisher Information Metric (FIM), used as the Riemannian structure [1]. The eigensystem of the FIM, therefore, represents the magnitudes and directions of curvature on the manifold. An input constructed in the direction of the eigenvector associated with the highest eigenvalue results in the highest divergence of the output distribution and thus can be considered an optimal attack under this formulation [22]. Further analysis indicated that these attacks were highly transferable implying a similarity in manifolds for separate and unique networks.

To defend against the FIM-based attack, Shen *et al*. developed an approach to suppress the eigenvalues, which was shown to be equivalent to the well-known label smoothing method [17]. Label smoothing was empirically shown to prevent the network from becoming over-confident [14]. This paper advances a very different and clever approach that aims at *deceiving the attacker* - who is utilizing the eigensystem of the FIM - by obfuscating the eigenvectors of the FIM. While smoothing the data manifold does reduce adversarial risk, the real danger of attack is from the directions of the eigenvectors. By transforming the data with an orthogonal transformation, we show that the eigenvectors are rotated - causing the data manifold to be distinct (ideally orthogonal) from the attacker's data manifold - while the eigenvalues are kept unchanged. Figure 1 shows an orthogonal transformation on images from MNIST and CIFAR-10. The resulting images are unrecognizable to the human eye but are congruent to the original image.

The rest of the paper is organized as follows. Section II describes the necessary background and common language used throughout this work. Section III describes the optimal adversarial problem from an information geometric perspective. Section IV formally introduces our defense approach of preprocessing the data with an orthogonal transformation and describes the effects on the model's manifold. In Section V, we present empirical evidence to validate our method. Lastly, Section VI summarizes the main findings and proposes directions for future work.

II. Background

A. Deep Neural Network Image Classifiers

In image classification, a data-set, $\mathcal{D} = \{(X, Y)\}$, is a set of tuples containing an image $x \in \mathbb{R}^{c \times h \times w}$, where c, h and w represent channels, pixel height and width, respectively, and a one-hot encoded label $y \in \mathbb{R}^k$ of k classes. A DNN image classifier $f(x; \theta)$ is a parameterized statistical model that maps an image x to a distribution over the k classes $f : X \mapsto P(Y|X)$. The distribution is created by normalizing the unbounded output of the DNN, known as scores or logits, and denoted as $Z(x)$. The logits are normalized to a pseudo probability distribution over the k classes $p(y_i|x)$ for score z_i via the softmax function σ, and are given by $\sigma(z_i) = e^{z_i} / \sum_{j=1}^{k} e^{z_j}$. The resulting softmax layer has the properties: $p(y|x; \theta) > 0, \ \forall \ y$, and $\sum p(y|x; \theta) = 1$, and therefore can be considered a probability distribution. The predicted class is determined as the class with the highest density, $\hat{y} = \arg\max_i \ p(y_i|x; \theta)$.

The loss function calculates the error between the model's prediction and the ground truth. For instance, the cross-entropy loss, or log loss, measures the difference between the true

distribution of the data and the model's learned distribution as follows:

$$\mathcal{L}_{CE} = \quad -\sum_i y_i \log p(y_i|x), \tag{1}$$

where y_i is the i^{th} label from the dataset distribution and $p(y_i|x)$ is the model's learned distribution. Equation (1) simplifies to $-\log p(y_{\text{true}}|x)$ for datasets with one hot encoded labels, where the probability of the true class label y_{true} is 1, while all others are 0; thus, removing the summation.

B. Adversarial Attacks

An adversarial image x' is the summation of the original image x and the attack perturbation $\delta \in \mathbb{R}^{c\times h\times w}$. The goal of the attack is to solve the following constrained optimization problem:

$$\begin{aligned} &\arg\max_i f(x) \neq \arg\max_i f(x+\delta) \\ &\text{such that} \quad ||\delta||_p < \epsilon, \end{aligned} \tag{2}$$

where $||\cdot||_p$ is the p norm, p can be 0, 1, 2 or ∞ and $\epsilon \in \mathbb{R}^+$. It is important - from the attacker's perspective - that p and ϵ are chosen such that the resulting image (x') is indistinguishable from the original (x) to the naked human eye.

Adversarial attacks can be classified by the amount of information known about the model being attacked. In a white box setting, the adversary has full knowledge of the target model, that is: architecture, parameters, and loss gradients. In contrast, a black box attack is generated when the adversary is completely ignorant. White box attacks are more powerful but are less likely to occur in real-world scenarios because any legitimate entity utilizing DNNs would not make their models' information publicly available. It is then more probable that the adversary will train their own model, generate a white box attack on it, then transfer the attack. In good favor of the adversary, adversarial attacks are transferable at a high rate making black-box attacks a real threat to neural network applications [4].

C. An Information Geometric Perspective of Deep Learning

Information geometry studies the intrinsic differential geometric structure over families of probability distributions. It can be applied to statistical models, such as DNNs, by considering the parameter space as points on a statistical manifold and using the model's estimation to generate the Riemannian metric. It was shown that the Fischer Information is a valid Riemannian metric, and it was further proven by Chenstov's Theorem that it is the only invariant measure for statistical models [1].

Now, we will formally introduce the Fischer Information Metric (FIM) for a n-dimensional statistical model $p(y|x;\xi)$ where $(x,y) \in \mathcal{D}$ and $\xi \in \mathbb{R}^n$ are the inputs, outputs and n model parameters, respectively. Given a point ξ, the FIM of the model is an $n \times n$ symmetric positive semi-definite matrix, defined by

$$G_\xi = \mathbb{E}_\xi \left[\nabla_\xi \log\, p(y|x;\xi) \nabla_\xi^\intercal \log p(y|x;\xi)\right] \tag{3}$$

$$= -\mathbb{E}_\xi \left[\nabla_\xi^2 \log\, p(y|x;\xi)\right]. \tag{4}$$

The distribution of the expectation in Eqs. (3)-(4) is $p(y|x;\xi)$, the random variable is the Hessian of the log likelihood with respect to the parameters. The term $-\log p(y|x;\xi)$ is interpreted as the amount of information at x. The FIM can therefore be interpreted as the "interest level" of the data point compared to the data used to fit the model [12]. When a low probability event occurs, the interest is high because it does not normally occur under the current distribution.

It should be noted that there are two forms of the FIM. The first considers the weights and biases of a network as parameters [15, 12]. The second view considers only the input as parameters while keeping the weights and biases fixed [18, 22, 17]. The former describes a family of distributions given a model's architecture, while the latter describes a family of distributions from one fixed weight model given a dataset ($\xi = x$). We denote the latter as *the data manifold of a trained classifier*, and is the form adopted in this work.

D. Orthogonal Transformations

An orthogonal (or orthonormal) matrix Q is a real square matrix with orthonormal bases for its row and column spaces, i.e., $Q^\intercal Q = QQ^\intercal = I$ and therefore $Q^\intercal = Q^{-1}$, making Q always invertible. The matrix can be used as an operator for a linear transformation $Q : T \mapsto T$ on a real inner product space T. Such a transformation preserves the inner product space, i.e., $\langle v, w\rangle = \langle Qv, Qw\rangle \ , \forall\ v, w \in T$, thus, an isometry of the Euclidean space.

The eigenvalues of an orthogonal matrix have modulus one, and therefore orthogonal transformations only rotate or reflect and do not scale. Suppose two matrices A and B are related by an orthogonal transformation, i.e., $A = QBQ^T$, then the eigenvalues of A and B will be equal but the eigenvectors are represented in different bases. An orthogonal matrix generalized to complex numbers is called a unitary matrix.

III. Attacks and Defenses

A. Attacks

1) Fast Gradient Sign Method (FGSM): was developed to make adversarial training, at any scale, fast and practical [4]. In general, methods for producing adversarial examples are computationally taxing, and result in unreasonably long training times. Although other attack methods may surpass the attack abilities of FGSM, few have the speed to enable rapid adversarial training. To comprehend the attack, let us first consider this linear single layer model below:

$$f(x;\theta) = \theta^\top x. \tag{5}$$

The model simply takes the dot product of the image and the model parameters. If a perturbation is added to the input image, the output will grow by $\theta^\top \delta$ as seen below:

$$f(x+\delta;\theta) = \theta^\top (x+\delta) = \theta^\top x + \theta^\top \delta. \tag{6}$$

Under the max norm (largest element) constraint of the perturbation ($||\delta||_\infty < \epsilon$), the greatest allowable increase is:

$$\delta = \epsilon \cdot \text{sign}(\theta). \tag{7}$$

The max norm is the most accurate norm to use; the reason being if the largest singular pixel perturbation is below human perception, then none of the perturbations are detected.

Under the assumption that the gradient with respect to the input yields the fastest changing direction in the output, the formulation in Eq. (7) is extended to DNNs as follows

$$\delta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(y, f(x))). \tag{8}$$

The effectiveness of this attack validates the assumption that training methods induce a linear structure to DNNs. As a consequence, adversarial examples would then exist along *directions* rather than *pockets* on the decision boundary. Indeed, experiments of scaling attacks corroborated the theory but furthermore these insights lead to the concept of the *transferability property* of adversarial examples. Since unique models are fitted to data in similar ways, their decision boundaries may be co-linear and therefore adversarial examples generated on one model will fool other models as well.

2) One Step Spectral Attack (OSSA): exploits the knowledge of the data manifold given by the FIM [22]. Let us first build an intuition by analyzing the effect of the change in input in the context of the output via the Kullback–Leibler (KL) divergence,

$$D_{KL}\left(p(y|x) \,||\, p(y|x+\delta)\right) \;=\; \mathbb{E}_x\left[\log \frac{p(y|x)}{p(y|x+\delta)}\right]. \tag{9}$$

Taking the 2nd order Taylor series expansion of $\log p(y|x+\delta)$ about point x, we obtain

$$\begin{aligned} D_{KL} \;&=\; \mathbb{E}_x\Bigg[\underbrace{\log p(y|x) - \log p(y|x)}_{=\ 0} - \\ &\qquad \delta\nabla_x \log p(y|x) - \frac{1}{2}\delta\nabla_x^2 \log p(y|x)\delta\Bigg] \qquad (10) \\ &=\; \mathbb{E}_x\Big[-\delta\nabla_x \log p(y|x)\Big] - \mathbb{E}_x\left[\frac{1}{2}\delta\nabla_x^2 \log p(y|x)\delta\right] \qquad (11) \\ &=\; -\delta\underbrace{\int p(y|x)\frac{\nabla_x p(y|x)}{p(y|x)}dx}_{=\ 0} - \\ &\qquad \frac{1}{2}\delta\underbrace{\mathbb{E}_x\left[\nabla_x^2 \log p(y|x)\right]}_{-G_x}\delta \qquad (12) \\ &\approx\; \frac{1}{2}\delta G_x \delta + \mathcal{O}\left(\delta^3\right). \qquad (13) \end{aligned}$$

The first term in Eq. (12) is equal to zero because $\int p(y|x)\left[\nabla_x p(y|x)/p(y|x)\right]dx \;=\; \int \nabla_x p(y|x)dx \;=\; \nabla_x \int p(y|x) = \nabla_x 1 = 0$. Note that the final approximate KL-divergence between the original and adversarial input - shown in Eq. (13) - is a quadratic form of the FIM.

The objective of the attacker is to maximize Eq. (13) and obtain the optimal perturbation δ. Maximizing Eq. (13) has a well-known solution in mathematics given by the eigenvector corresponding to the highest eigenvalue of G_x; thus, the name 'Fisher spectral attack' or 'One Step Spectral Attack' [22]. This direction can be viewed as the steepest direction in the data manifold, i.e., a movement in this direction results in the highest KL-divergence, and therefore the highest likelihood to confuse the DNN.

B. Defenses

1) Suppressing the Eigenvalues of the Fisher Information Matrix (EVS): The suppression of the FIM eigenvalues was inspired by the derivation of the one-step spectral attack [17]. The intuition being if the highest eigenvalue of G_x indicates the level of adversarial threat, then suppressing this value will produce a more robust network. The max eigenvalue problem is mathematically intractable. A reasonable approximation is to consider instead the trace of G_x since it is equivalent to the summation of the eigenvalues, $\text{tr}\ G_x = \sum \lambda_i$. The new training loss then becomes

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \mu\ \text{tr}\, G_x. \tag{14}$$

Calculating all eigenvalues directly is also intractable for large input layers. The problem can be further simplified by considering the softmax layer $s(x) = [p_1, ..., p_k]$ rather than the input layer since the number of classes is much smaller than the input size. The two representations are related by the Jacobian of the softmax layer with respect to the input, the $cmn \times k$ matrix $J = \partial s/\partial x$.

$$G_x = J^\intercal G_s J. \tag{15}$$

Due to the linear mapping between the two FIMs, the suppression of the eigenvalues of one will result in a suppression of the eigenvalues of the other. Lastly, to avoid direct computation of the eigenvalues of G_s, the following derivation is performed,

$$\begin{aligned} \text{tr}\, G_s &= \text{tr}\, \mathbb{E}_{y|s}\left[\nabla_s \log p(y|s) \cdot \nabla_s^\intercal \log p(y|s)\right] \\ &= \int_{y|s} p(y|s)\ \text{tr}\ \left[\nabla_s \log p(y|s) \cdot \nabla_s^\intercal \log p(y|s)\right] \\ &= \int_{y|s} p(y|s) \cdot \|\nabla_s \log p(y|s)\|_2^2 \\ &= \sum_{i=1}^{k} p_i \sum_{j=1}^{k} \left[\nabla_{p_j} \log p(y_i|s)\right]^2 \\ &= \sum_{i=1}^{k} \frac{1}{p(y_i|x)}. \end{aligned} \tag{16}$$

By substituting Eq. (16) into (14), we have,

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \mu \cdot \sum_{i=1}^{k} \frac{1}{p_i}. \tag{17}$$

Intuitively, the new loss term in Eq. (17) will drive the output towards a uniform distribution, $s(x) = [1/k, ..., 1/k]$ while the cross-entropy term will force the correct class to have more density. In practice, this accomplishes label smoothing regularization, a technique that is already understood to improve robustness [14] but now has a strict mathematical deduction as its foundation.

IV. Proposed Orthogonal Defense Approach

We now describe our approach for hardening a deep neural network to the optimal OSSA adversarial attack. To begin, we rely on the initial formulation of the quadratic form of the Fisher Information matrix given in Eq. (13) and formally define the objective function of the adversary as:

$$\begin{aligned} \max_{\delta} \quad & \frac{1}{2}\delta^{\intercal} G_x \delta \\ \text{s.t.} \quad & ||\delta||_2^2 = \epsilon \\ & \mathcal{L}\left(y, x+\delta\right) > \mathcal{L}\left(y, x\right) \end{aligned} \tag{18}$$

The expression $v^{\intercal}Av$ is bounded by the eigenvalues of the symmetric matrix A, such that $\lambda_{\min} \leq v^{\intercal}Av \leq \lambda_{\max}$, and the vectors v that produce these bounds are the respectively associated eigenvectors. The direction of the eigenvector associated with the highest eigenvalue will, therefore, produce the greatest KL-divergence between the normal and attacked output with a magnitude equal to the eigenvalue.

White-box attacks are the worst-case scenario for DNN models; however, since G_x is derived with fixed model parameters θ, it is only valid as the Riemannian metric for the training data manifold of this particular model while other models will have their unique data representations. The existence of the transferability property for adversarial attacks indicates that - although networks have different architectures and parameters - they represent the data in a similar fashion. Under the Fisher Information matrix view, it is conjectured that the eigenvectors associated with the largest eigenvalue of two such networks are co-linear.

Given this knowledge, an obvious approach to defend against black-box attacks is to suppress the eigenvalues of the FIM. This was solved by Shen *et al*. in [17] and the solution was shown to be equivalent to label smoothing. In this paper, we propose a new approach that aims at *deceiving the attacker* into adopting an erroneous direction of attack. To achieve this goal, we transform the input (prior to training) such that the highest eigenvalue of the FIM is preserved but the corresponding direction (of attack) is changed.

Mapping the training dataset with an orthogonal transformation (Orthogonal PreProcess, OPP) will change the eigenvectors of the resulting FIM while maintaining the same eigenvalues; thus fooling the attacker by changing the optimal direction of attack.

V. Experiments and Discussion

A. Data sets and Models

We evaluate the proposed OPP defense on two standard benchmark data sets in image classification: MNIST [2] and CIFAR-10 [9]. The former is relatively small, consisting of 60,000 single channel gray scale 28x28 images of hand-written digits from zero to nine. The latter is the same size but the images are three channel colored 32x32 natural images encompassing 10 classes, including airplanes, cars and birds. For standard preprocessing, both data sets are normalized channel-wise and augmented in training via rotation, horizontal flip, color jitter and affine transformation. We implement LeNet-5 on MNIST, a small architecture of three convolutional layers followed by two fully connected layers. It is one of the earliest convolutional neural network (CNN) architectures to be successfully implemented [10]. We trained using a Stochastic Gradient Decent (SGD) optimizer with a batch size of 512, a momentum of 0.9 and a weight decay of 0.0001. Our learning rate was scheduled through Cosine Annealing [11] starting at 0.5 and atrophying to 0.

On CIFAR-10, a light weight network called MobileNetV2 [16] is fitted. MobileNet uses depth-wise separable convolutions, which reduces the model complexity leading to less over fitting [6]. It is trained with the Adam optimizer [8] using a batch size of 124 and the same momentum and weight decay values of 0.9 and 0.0001, respectively. In this case, the learning rate was scheduled using One Cycle LR [19] starting at 0.00004 escalating to 0.001 then deescalating to 0.

B. Model Analysis

It is hypothesized that attacks live along directions and not pockets and therefore can span directions with scalar multiplication [4]. In our experiments, the scaling factor is based on the ratio of the Euclidean norms of the perturbation and the original image, i.e.,

$$\epsilon = \frac{\|\delta\|_2}{\|x\|_2}. \tag{19}$$

The Euclidean norm measures the energy of the signal. If the energy of the perturbation is too low, the image will overpower the signal resulting in an unsuccessful attack. If the perturbation is too high then the resulting attack will no longer be considered adversarial, as a human could easily identify the perturbation. In our experiments, we exhaustively analyzed different ϵ values to check if they could be detected by the human eye. It was determined heuristically, from Figure 2, that $\epsilon = 0.3$ is the threshold for both MNIST and CIFAR-10 under FGSM attacks.

The effectiveness of an attack is measured by its fooling ratio or the percentage of images misclassified that were originally correctly classified. Images already misclassified before the attack are not considered because it is already known that the model is confused by such inputs. Thus, the fooling ratio is defined as

$$\text{Fooling Ratio} = \frac{|A_{test} - A_{adversary}|}{A_{test}}, \tag{20}$$

where A represents the accuracy in terms of the ratio of images correctly classified. A fooling ratio of 0 would be the result of the adversarial accuracy equaling the original accuracy indicating no change occurred to the network's performance. A fooling ratio of 1 occurs when the adversarial accuracy is 0% and therefore the network is completely fooled.

C. Classification Accuracy

In this study, the orthogonal matrices are generated from a QR decomposition of a randomly generated matrix. Figure 1

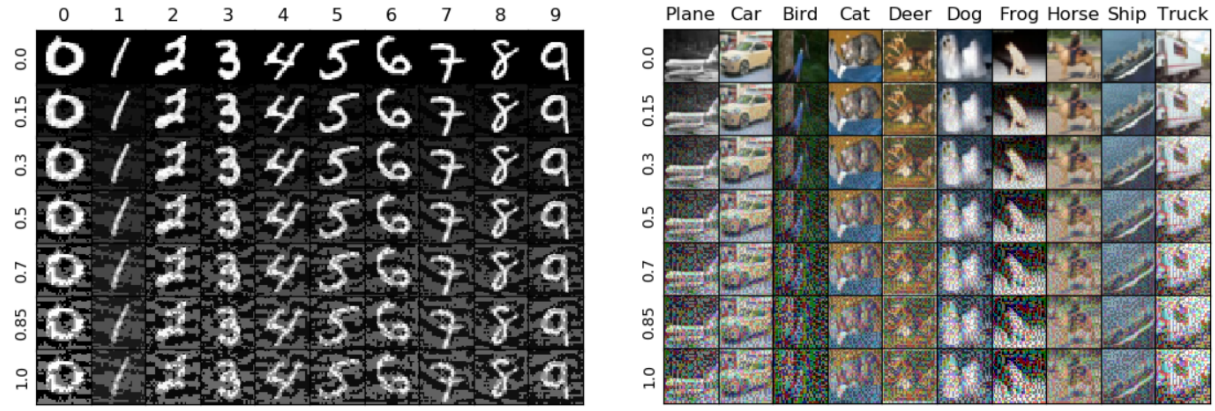

Fig. 2. **Progression of FGSM perturbation visibility** with respect to the noise-to-signal ratio, ϵ, for each class of MNIST and CIFAR-10. It can be observed that the threshold at which the perturbations remains imperceptible is at $\epsilon = 0.3$ for both MNIST and CIFAR-10.

TABLE I
A COMPARISON OF CLASSIFICATION ACCURACY AND TRAINING TIME FOR THE STANDARD AND OPP MODELS ON MNIST AND CIFAR-10.

Dataset	Metric	Standard Model	OPP Model
MNIST	Training Time	5m 49s	5m 45s
	Classification Accuracy	0.9824	0.9448
CIFAR-10	Training Time	4h 33m 44s	5h 24m 2s
	Classification Accuracy	0.9327	0.7682

shows sample orthogonally-transformed inputs from MNIST and CIFAR-10 datasets. Although the resulting images are indiscernible, a deep neural network is still *theoretically* guaranteed to learn by the universal approximation theorem [5]. For training outcomes to be comparable between a standard network and an OPP network, identical training regimens and computer resources were used, as described earlier in Section V-A.

Table I shows the evaluation accuracy and training time of the standard model and the proposed OPP model for the MNIST and CIFAR-10 datasets. Note that both models perform similarly on the MNIST data. For the CIFAR-10 data, it seems that the OPP model underperforms with a drop in accuracy of $\sim 16\%$ and an additional 1h in training time. However, this is due to the fact that the adopted model (architecture and hyperparameters) was optimized for the standard dataset and not its OPP counterpart, i.e., the orthogonally-transformed dataset. The OPP accuracy could match the standard training model by using a network with higher model capacity and more advanced training precautions. The time requirement issue could also be mitigated if the orthogonal transformation is done once prior to testing. In this experiment, the preprocessing is done online, one image at a time.

D. Adversarial Robustness Evaluation

Figure 3 plots the fooling ratios of an undefended model (blue curve) and the OPP defense approach (orange curve) against FGSM and OSSA attacks, for MNIST and CIFAR-10 datasets. There is a noticeable increase in robustness with the proposed OPP - orthogonal transformation - approach for both MNIST and CIFAR-10. Although the fooling tends to increase monotonically with the ratio of noise to signal ϵ, the real adversarial attack stops much earlier at the point of being perceived by the naked human eye. In Figure 2, we heuristically determined the maximum value of ϵ to be 0.3 for MNIST and CIFAR-10 attacks to be perceptible.

E. OPP vs. EVS

To the best of our knowledge, our method and the method of suppressing eigenvalues of the FIM (EVS) are the only defense methods that intentionally manipulate the data manifold as formulated in Eq. (13). In this section, we show that although eigenvalues indicate the severity of the adversarial threat, it is the directions, i.e., the eigenvectors, of the associated eigenvalues that are the real menace.

Table II displays the fooling ratios of the proposed OPP defense and the EVS technique at the perception threshold of $\epsilon = 0.3$. The OPP model achieves nearly half the fooling ratio of the EVS model. Both methods are based on theoretical foundations and manipulate the data manifold in such a way to diminish the transferability property of black box attacks.

VI. CONCLUSION

In this study, the adversarial attack and defense were viewed through an information geometric lens. From this perspective, we introduced a method of orthogonally transforming the data to directly manipulate the data manifold of a deep learning model. More specifically, our defense changes the basis of

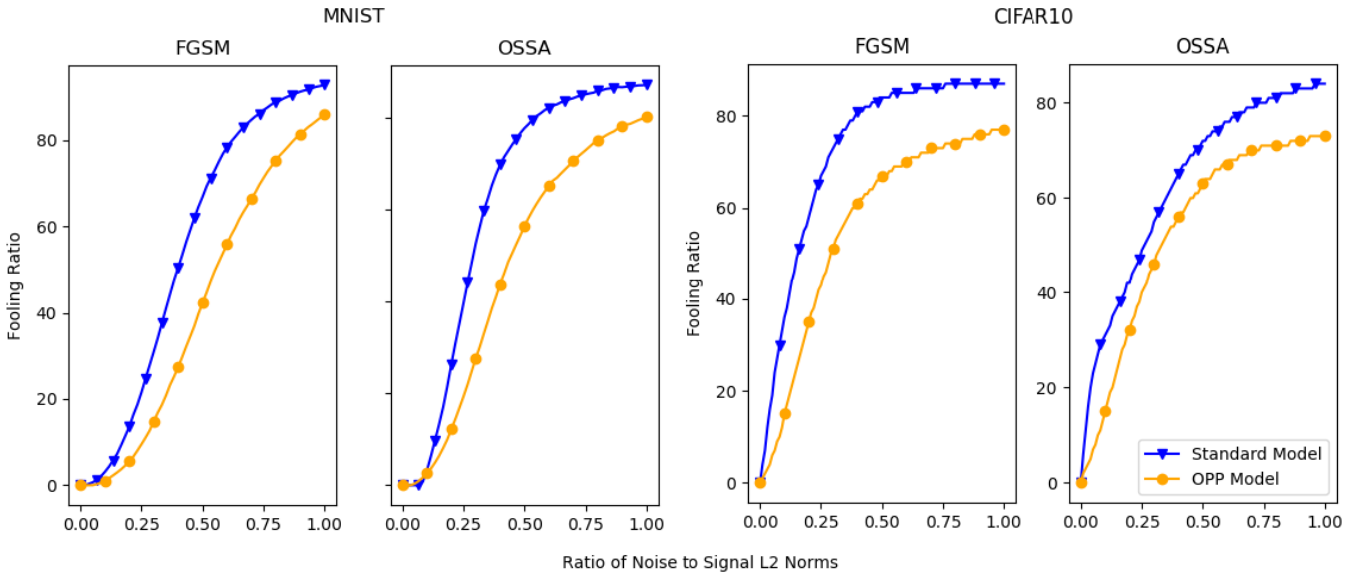

Fig. 3. **Fooling Ratio of the proposed OPP defense**. The y-axis displays the fooling ratio and the x-axis displays the noise to signal ratio. The blue curves denote a standard (undefended) model and the orange curves denote the proposed OPP defense model - trained with orthogonal transformation of the data. The attacks were generated using the FGSM and OSSA models. Observe that OPP consistently mitigates the effects of the attack as compared to an undefended model.

TABLE II
COMPARISON OF THE FOOLING RATIOS BETWEEN THE PROPOSED OPP MODEL AND THE EVS MODEL AT THE PERCEPTION THRESHOLD OF NOISE-TO-SIGNAL RATIO $\epsilon = 0.30$.

Dataset	Attack Type $\epsilon = 0.30$	Fooling Ratio	
		OPP Model	EVS Model
MNIST	FGSM	**15%**	34%
	OSSA	**28%**	52%
CIFAR-10	FGSM	**25%**	49%
	OSSA	**24%**	38%

the eigenvectors of the Fisher Information metric at each point of the data manifold. The manipulation is unique to the choice of the orthogonal matrix; thus, attackers without knowledge of the specific orthogonal matrix are unable to form black box attacks. We validated our method on two poplar benchmark datasets, MNIST and CIFAR-10, against two attack methods, FGSM and OSSA. In comparison with the eigenvalue suppression (EVS) defense - the only other method to our knowledge to leverage an information geometric formulation - we consistently scored half fooling ratio.

The experiments were limited to image classification; yet, the theory generalizes to many other machine learning tasks that are vulnerable to adversarial attacks. An important observation is that the orthogonal matrix acts as a lock and key for the data representation - like an encryption.

This paper considered a fixed random orthogonal transformation. Future work will investigate the effect of different orthogonal transformations, or even a distribution of these transformations, which could lead to theoretical properties and bounds of the proposed OPP defense.

REFERENCES

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2007.

[2] L. Deng. "The MNIST Database of Handwritten Digit Images for Machine Learning Research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[3] A. Gleave et al. "Adversarial Policies: Attacking Deep Reinforcement Learning". In: *CoRR* abs/1905.10615 (2019). arXiv: 1905.10615.

[4] I. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and Harnessing Adversarial Examples". In: *arXiv 1412.6572* (2014).

[5] K. Hornik, M. Stinchcombe, and H. White. "Multilayer Feedforward Networks are Universal Approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[6] A. Howard et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *CoRR* abs/1704.04861 (2017).

[7] R. Jia and P. Liang. “Adversarial Examples for Evaluating Reading Comprehension Systems”. In: *CoRR* abs/1707.07328 (2017).

[8] D. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015.

[9] A. Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: 2009.

[10] Y. Lecun et al. “Gradient-Based Learning Applied to Document Recognition”. In: *Proceedings of the IEEE* 86 (Dec. 1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[11] I. Loshchilov and F. Hutter. “SGDR: Stochastic Gradient Descent with Restarts”. In: *CoRR* abs/1608.03983 (2016).

[12] J. Martin and C. Elster. “Inspecting adversarial examples using the fisher information”. In: *Neurocomputing* 382 (2020), pp. 80–86. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.11.052.

[13] S. Moosavi-Dezfooli et al. “Analysis of universal adversarial perturbations”. In: *CoRR* abs/1705.09554 (2017).

[14] R. Müller, S. Kornblith, and G. Hinton. “When Does Label Smoothing Help?” In: *CoRR* abs/1906.02629 (2019).

[15] A. Nayebi and S. Ganguli. “Biologically inspired protection of deep networks from adversarial attacks”. In: (Mar. 2017). arXiv: 1703.09202 [stat.ML].

[16] M. Sandler et al. “Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation”. In: *CoRR* abs/1801.04381 (2018).

[17] C. Shen et al. “Defending Against Adversarial Attacks by Suppressing the Largest Eigenvalue of Fisher Information Matrix”. In: *CoRR* abs/1909.06137 (2019).

[18] Y. Shi et al. *Understanding Adversarial Behavior of DNNs by Disentangling Non-Robust and Robust Components in Performance Metric*. 2019. arXiv: 1906.02494 [stat.ML].

[19] L. Smith and N. Topin. “Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates”. In: *CoRR* abs/1708.07120 (2017). arXiv: 1708.07120.

[20] C. Szegedy et al. “Intriguing properties of neural networks”. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2014.

[21] L. Wu and Z. Zhu. “Towards Understanding and Improving the Transferability of Adversarial Examples in Deep Neural Networks”. In: *Proceedings of The 12th Asian Conference on Machine Learning*. Ed. by Sinno Jialin Pan and Masashi Sugiyama. Vol. 129. Proceedings of Machine Learning Research. PMLR, 18–20 Nov 2020, pp. 837–850.

[22] C. Zhao et al. “The Adversarial Attack and Detection under the Fisher Information Metric”. In: *CoRR* abs/1810.03806 (2018).