# ADVERSARIAL ROBUSTNESS WITH PARTIAL ISOMETRY

*Loïc Shi-Garrier*[⋆] *Nidhal Carla Bouaynaya*[†] *Daniel Delahaye*[⋆]

[⋆] Ecole Nationale de l'Aviation Civile, Université de Toulouse, France
[†] Dept. of Electrical and Computer Engineering, Rowan University, New Jersey, USA

## ABSTRACT

Despite their outstanding performance, deep learning models still lack robustness guarantees, notably in the face of adversarial examples. This major weakness prevents the introduction of these learning systems in critical domains where a certain level of robustness must be certified. In this paper, we present an information geometric framework to derive a precise robustness criteria for $l_2$ white-box attacks in a multi-class classification setting. We endow the output space with the Fisher information metric and derive a criteria on the input-output Jacobian to ensure robustness. We show that model robustness can be achieved by constraining the model to be partially isometric around the training points. The approach is tested on MNIST and CIFAR-10 against adversarial attacks. It is shown to be significantly more robust than defensive distillation and Jacobian regularization for mediums-sized perturbations and more robust than adversarial training for large perturbations, while still maintaining desired accuracy.

***Index Terms*—** Adversarial robustness, Information geometry, Fisher information metric, Multi-class classification.

## 1. INTRODUCTION

The machine learning community has started to study the robustness problems of machine learning models, neural networks in particular. This study was motivated by the high sensitivity of neural networks to adversarial attacks, i.e., small perturbations in the input data that are able to fool a network. Adversarial attacks have been shown to be both ubiquitous and transferable [1], [2], [3]. Beyond the obvious security threat, adversarial attacks are the evidence of the dramatic lack of robustness in machine learning models [4], [5].

In this paper, we shed an information geometric perspective to adversarial robustness in machine learning models. We show that robustness can be achieved by encouraging the model to be isometric in the orthogonal space of the kernel of the pullback Fisher metric. We subsequently formulate a regularization defense method for adversarial robustness. We focus on $l_2$ white-box attacks against multi-class classification tasks; but the approach could be extended to more general settings, e.g., unrestricted attacks and black-box attacks, as well as to other supervised learning tasks. The regularized model is evaluated on MNIST and CIFAR-10 against PGD $l_\infty$ attacks and AutoAttack [6] with $l_\infty$ and $l_2$ norms. Comparisons with unregularized model, defensive distillation [7], Jacobian regularization [8], and Fisher information regularization [9] show significant improvement in robustness. Moreover, the regularized model is able to ensure robustness for larger perturbations compared to adversarial training. We pay special attention to the computational efficiency of the method since we hope that it could be used in real-world applications.

The remaining of this paper is divided into four parts. Section 2 introduces the notations and presents a sufficient condition for adversarial robustness at a sample point. The partial isometry regularization defense is then derived by approximating this condition for mathematical tractability. Section 3 presents several experiments to evaluate the proposed method. Section 4 discusses the results in the lights of related work on adversarial defense. Finally, section 5 concludes the paper and suggests potential extensions of this work.

## 2. PARTIAL ISOMETRY REGULARIZATION

We will denote the $i^{\text{th}}$ component of a vector $v$ as $v^i$. Smooth means $C^\infty$.

### 2.1. Family of Categorical Distributions

Consider a multi-class classification task. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the *input domain*, and let $\mathcal{Y} = \{1, \dots, m+1\} \subset \mathbb{N}$ be the set of labels for the classification task. We assume than $n > m$. For example, in MNIST, we have $\mathcal{X} = [0,1]^n$ (with $n = 784$) and $m = 9$. We assume that $\mathcal{X}$ is an $n$-dimensional embedded smooth connected submanifold of $\mathbb{R}^n$. A machine learning model (e.g., a neural network) is often seen as assigning a label $y \in \mathcal{Y}$ to a given input $x \in \mathcal{X}$. Instead, in this work, we see a model as assigning the *parameters* of a random variable $Y$ to a given input $x \in \mathcal{X}$. The random variable $Y$ has a probability density function $p_\theta$ belonging to the *family of* $(m+1)$*-dimensional categorical distributions* $\mathcal{S} = \{p_\theta : \theta \in \Delta^m\}$. The set $\Delta^m$ is defined as $\Delta^m = \{\theta \in \mathbb{R}^{m+1} : \sum_{i=1}^{m+1} \theta^i = 1, 0 < \theta^i < 1\}$. $\Delta^m$ is called the *probability* $m$*-simplex*.

$\mathcal{S}$ can be endowed with a differentiable structure by using $p_\theta \in \mathcal{S} \mapsto (\theta^1, \dots, \theta^m) \in \mathbb{R}^m$ as a global coordinate system. Hence, $\mathcal{S}$ becomes a smooth manifold of dimension $m$ (more details on this construction can be found in Amari 1985 [10] Chapter 2). We can identify $p_\theta$ with $(\theta^1, \dots, \theta^m)$.

We call *model* any smooth map $F : \mathcal{X} \to \Delta^m$, that assigns to an input $x \in \mathcal{X}$, the parameters $\theta = F(x) \in \Delta^m$ of a $(m+1)$-dimensional categorical distribution $p_\theta \in \mathcal{S}$. In practice, a neural network produces a vector of *logits* $s(x)$. Then, these logits are transformed into the parameters $\theta$ with the softmax function: $\theta = \text{softmax}(s(x))$.

### 2.2. Riemannian Metrics

Let $F$ be a model. In order to study the sensitivity of the predicted $F(x) \in \Delta^m$ with respect to the input $x \in \mathcal{X}$, we need to be able to measure distances both in $\mathcal{X}$ and in $\Delta^m$. In order to measure distances on smooth manifolds, we need to equip each manifold with a Riemannian metric.

Let us begin with $\mathcal{X}$. Since we are studying adversarial robustness, we need a metric that formalizes the idea that two close data points must be "indistinguishable" from a human perspective (or any other relevant perspective). A natural choice is the *Euclidean metric*. Using the standard coordinate of $\mathbb{R}^n$ as a global coordinate system for $\mathcal{X}$, the Euclidean metric is defined as

$$\overline{g}_x = \sum_{i,j=1}^{n} \delta_{ij} dx^i dx^j,$$

where $\delta_{ij} = 1$, if $i = j$, and 0 otherwise. The Euclidean metric corresponds to the $l_2$ norm.

Now, we consider $\Delta^m$. As described above, we see $\Delta^m$ as the family of categorical distributions. A natural Riemannian metric for $\Delta^m$ (i.e., a metric that reflects the statistical properties of $\Delta^m$) is the *Fisher information metric* (FIM) defined by:

$$g_\theta^{FIM} = \sum_{i,j=1}^{m+1} \frac{1}{\theta^i} \delta_{ij} d\theta^i d\theta^j.$$

The FIM has two remarkable property. First, it is the "infinitesimal distance" of the the *relative entropy* (Theorem 4.4.5 in [11]), which is the loss function used to train a multiclass classification model. The other remarkable property of the FIM is Chentsov's theorem [12] claiming that the FIM is the *unique* Riemannian metric on $\Delta^m$ that is invariant under sufficient statistics (up to a multiplicative constant).

In order to see $\Delta^m$ as a manifold in its own right, we must define the FIM with respect to a coordinate system. Let $g$ be the FIM expressed in the coordinate system $(\theta^1, \dots, \theta^m)$. In these coordinates, we have:

$$g_\theta = \sum_{i,j=1}^{m} \left( \frac{1}{\theta^i} \delta_{ij} + \frac{1}{\theta^{m+1}} \right) d\theta^i d\theta^j, \tag{1}$$

where $\theta^{m+1} = 1 - \sum_{k=1}^{m} \theta^k$.

We also define the *pullback metric* $\tilde{g}$ of the FIM by $F$, i.e. $\tilde{g} = F^* g$. Using the standard coordinates on $\mathcal{X}$ and the coordinates $(\theta^1, \dots, \theta^m)$ on $\Delta^m$, we have for any $x \in \mathcal{X}$:

$$\tilde{g}_x = J_x^T g_{F(x)} J_x, \tag{2}$$

where $J_x$ is the Jacobian matrix of $F$ at $x$.

### 2.3. Adversarial Robustness - a sufficient condition

Now, consider the decision boundary in $\Delta^m$ defined as the set $\mathcal{B}$ such that $\theta \in \mathcal{B}$ if the maximum of $\theta$ is achieved at two different components, i.e., $\max_{1 \le k \le m+1} \theta^k = \theta^i = \theta^j$ for some $i \neq j$. We are interested in the distance $d(F(x), \mathcal{B})$ between $F(x)$ and the decision boundary $\mathcal{B}$. Instead of deriving an exact formula, we use the following upper bound:

$$d(F(x), \mathcal{B}) \le d(F(x), c),$$

where $c = \frac{1}{m+1}(1, \dots, 1)$ is the center of the simplex $\Delta^m$. Let us denote $\delta(x) = d(F(x), c)$. It can be shown that:

$$\delta(x) = 2 \arccos \sum_{i=1}^{m+1} \sqrt{\frac{F(x)^i}{m+1}}. \tag{3}$$

Let $x \in \mathcal{X}$ be a training point. Let $\epsilon > 0$ be a chosen budget. Consider the Euclidean open ball $\mathcal{B}(x, \epsilon) = \{z \in \mathcal{X} : \|z - x\| < \epsilon\} \subset \mathcal{X}$. Let $\widetilde{B}(0, \delta) = \{X \in T_x\mathcal{X} : \sqrt{\tilde{g}_x(X, X)} < \delta\}$ where $T_x\mathcal{X}$ is the tangent space of $\mathcal{X}$ at $x$. The *geodesic ball* $\widetilde{\mathcal{B}}(x, \delta)$ is defined as $\widetilde{\mathcal{B}}(x, \delta) = \{\exp_x(X) \in \mathcal{X} : X \in \widetilde{B}(0, \delta)\}$.

We say that the model $F$ is *adversarially robust* at $x$ if every point in $\mathcal{B}(x, \epsilon)$ has the same class, i.e., if $\mathcal{B}(x, \epsilon)$ does not intersect any decision boundary. Hence, *for $F$ to be adversarially robust at $x$, it is sufficient that:*

$$\mathcal{B}(x, \epsilon) \subseteq \widetilde{\mathcal{B}}(x, d(F(x), \mathcal{B})), \tag{4}$$

### 2.4. Adversarial Robustness - a condition on the Jacobian matrix

In order to obtain an implementable condition, we make the following simplifications. First, we use $\delta(x)$ (Eq. (3)) instead of $d(F(x), \mathcal{B})$. Second, we neglect the curvature of $\tilde{g}$ in $\widetilde{\mathcal{B}}(x, \delta)$. Third, we impose that the Euclidean ball be equal to the geodesic ball instead of being only included in it. However, since $n > m$, the pullback metric $\tilde{g}$ cannot be positive definite, it is only positive semi-definite. In particular, it has a nontrivial kernel and thus the two balls cannot be equal. To solve this problem, let $D = (\ker \tilde{g}_x)^\perp$ be the orthogonal space of the kernel of $\tilde{g}_x$, where the orthogonality is defined using the inner product induced by the Euclidean metric $\overline{g}$. Our third simplification consists in imposing that the Euclidean ball *in* $D$ is equal to the geodesic ball *in* $D$.

Given these simplifications, we replace Eq. (4) by the following stronger condition:

$$\tilde{g}_x|_D = \frac{\delta(x)^2}{\epsilon^2}\overline{g}_x|_D. \tag{5}$$

Then, using Eq. (2) and the coordinate change $t = T(\theta) = 2\sqrt{\theta}/\left(1-\sqrt{\theta^{m+1}}\right)$, it can be shown that Eq. (5) is equivalent to:

$$\tilde{J}_x\tilde{J}_x^T = \left(\frac{\delta(x)\kappa(x)}{\epsilon}\right)^2 I_m, \tag{6}$$

where $\tilde{J}_x$ is the Jacobian matrix of $T \circ F$ at $x$, $I_m$ is the identity matrix, and $\kappa(x)$ is defined as:

$$\kappa(x) = \frac{\sqrt{F^{m+1}(x)}}{2\sqrt{F^{m+1}(x)} - \|F(x)\|_1}$$

with $\|F(x)\|_1 = \sum_{i=1}^m F^i(x)$, and $F^i$ is the $i$-th component of $F$ in the coordinates $(\theta^1, \ldots, \theta^{m+1})$. Equation (6) constrains the Jacobian matrix $\tilde{J}_x$ to be a *semi-orthogonal matrix* multiplied by a homothety matrix. With this condition, $F$ becomes a *partial isometry*, at least in the neighborhood of the training points.

Now, we can define a regularization term:

$$\alpha(x, F) = \epsilon^2 \left\| \tilde{J}_x\tilde{J}_x^T - \frac{\delta(x)^2}{\rho(x)^2\epsilon^2} I_m \right\|_F, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm. We multiplied the regularization term by $\epsilon^2$ in order to compensate the $1/\epsilon^2$. This will facilitate the hyperparameter tuning when using different values of $\epsilon$. To compute $\alpha(x, F)$, we only need to compute the Jacobian matrix $\tilde{J}_x$ which can be efficiently achieved with backpropagation. The regularized loss function is given as:

$$\mathcal{L}(\hat{\theta}, x, F) = (1-\eta)H(\hat{\theta}||F(x)) + \eta\alpha(x, F), \tag{8}$$

where $H$ is the relative entropy (or the cross entropy in practice), and $\eta$ is a hyperparameter to control the trade-off between the relative entropy and the regularization term.

## 3. EXPERIMENTS

### 3.1. Experimental Setup

The regularization method introduced in Section 2 is evaluated on MNIST dataset. We implemented a LeNet model with two convolutional layers of 32 and 64 channels respectively, followed by one hidden layer with 128 neurons. We train three models: one regularized model, one baseline unregularized model, and one model trained with adversarial training. All three models are trained with Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 30 epochs, with a batch size of 64, and a learning rate of $10^{-3}$. For the regularization term, we use a budget of $\epsilon = 5.6$, which is chosen to contain the $l_\infty$ ball of radius 0.2. The adversarial training is conducted with 10 iterations of PGD with a budget $\epsilon_{adv} = 0.2$ using $l_\infty$ norm. We found that $\eta = 10^{-6}$ yields the best performance in terms of robustness-accuracy tradeoff; this value is small because we did not attempt to normalize the regularization term.

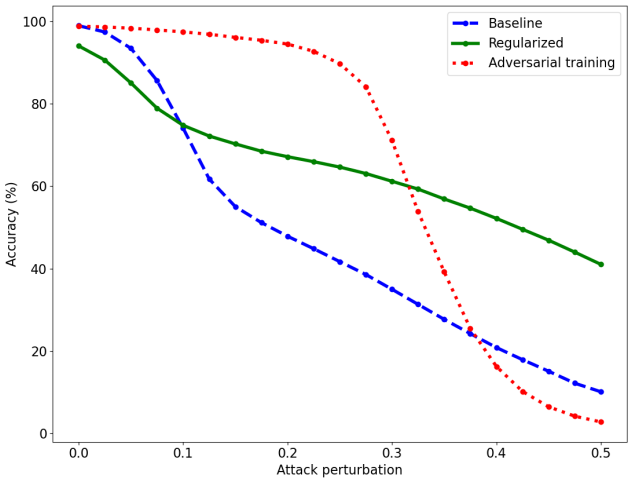


**Fig. 1**. Accuracy of the baseline (dashed, blue), regularized (solid, green), and adversarially trained (dotted, red) models for various attack perturbations on the MNIST dataset. The perturbations are obtained with PGD using $l_\infty$ norm.

The models are trained on the 60,000 images of MNIST's training set, then tested on the 10,000 images of the test set. The baseline model achieves an accuracy of 98.9% (9893/10000), the regularized model achieves an accuracy of 94.0% (9403/10000), and the adversarially trained model achieves an accuracy of 98.8% (9883/10000). Although the current implementation of the regularized model is almost 6 times slower to train than the baseline model, it may be possible to accelerate the training using, for example, the technique proposed by Shafahi *et al.* [13], or using another method to approximate the spectral norm of $\tilde{J}_x$. Even without relying on these acceleration techniques, the regularized model is still faster to train than the adversarially trained model.

### 3.2. Robustness to Adversarial Attacks

To measure the adversarial robustness of the models, we used the PGD attack with the $l_\infty$ norm, 40 iterations, and a step size of 0.01. The $l_\infty$ norm yields the hardest possible attack for our method, and corresponds more to the human notion of "indistinguishable images" than the $l_2$ norm. The attacks are performed on the test set, and only on images that were correctly classified by each model. The results are reported in Fig. 1. The regularized model has a slightly lower accuracy than the baseline model for small perturbations, but

the baseline model suffers a drop in accuracy above attack level $\epsilon = 0.1$. Adversarial training achieves high accuracy for small to medium-sized perturbations but the accuracy decreases sharply above $\epsilon = 0.3$. The regularized model remains robust even for large perturbations. The baseline model reaches 50% accuracy at $\epsilon = 0.2$ and the adversarially trained model at $\epsilon = 0.325$, while the regularized model reaches 50% accuracy at $\epsilon = 0.4$.

## 4. DISCUSSION AND RELATED WORK

In 2019, Zhao *et al.* [14] proposed to use the Fisher information metric in the setting of adversarial attacks. They used the eigenvector associated to the largest eigenvalue of the pullback of the FIM as an attack direction. Following their work, Shen *et al.* [9] suggested a defense mechanism by suppressing the largest eigenvalue of the FIM. They upper-bounded the largest eigenvalue by the trace of the FIM. As in our work, they added a regularization term to encourage the model to have smaller eigenvalues. Moreover, they showed that their approach is equivalent to label smoothing [15]. In our framework, their method consists in expanding the geodesic ball $\widetilde{\mathcal{B}}(x, \delta)$ as much as possible. However, their approach does not guarantee that the constraint imposed on the model will not harm the accuracy more than necessary. In our framework, the quantity $\delta(x)\kappa(x)/\epsilon$ informs the model on the precise restriction that must be imposed to achieve adversarial robustness in the $l_2$ ball of radius $\epsilon$.

Cisse *et al.* [16] introduced an other adversarial defense called *Parseval networks*. To achieve adversarial robustness, the authors aim at controlling the Lipschitz constant of each layer of the model to be close to unity. This is achieved by constraining the weight matrix of each layer to be a *Parseval tight frame*, which is synonymous with semi-orthogonal matrix. Since the Jacobian matrix of the entire model with respect to the input is almost the product of the weight matrices, the Parseval network defense is similar to our proposed defense, albeit with completely different rationales. This suggests that geometric reasoning could successfully supplement the line of work on Lipschitz constants of neural networks.

Following another line of work, Hoffman *et al.* [8] advanced a Jacobian regularization to improve adversarial robustness. Their regularization consists in using the Frobenius norm of the input-output Jacobian matrix. To avoid computing the true Frobenius norm, they relied on random projections, which are shown to be both efficient and accurate. This method is similar to the method of Shen *et al.* [9] in the sense that it will also increase the radius of the geodesic ball. However, the Jacobian regularization does not take into account the geometry of the output space (i.e., the Fisher information metric) and assumes that the probability simplex $\Delta^m$ is Euclidean.

Although this study focuses on $l_2$ norm robustness, it must be pointed out that there are other "distinguishability" measures that can be used to study adversarial robustness, including all other $l_p$ norms. In particular, the $l_\infty$ norm is often considered to be the most natural choice when working with images. However, the $l_\infty$ norm is not induced by any inner product, and hence, there is no Riemannian metric that induces the $l_\infty$ norm. However, given an $l_\infty$ budget $\epsilon_\infty$, we can choose an $l_2$ budget $\epsilon_2 = \sqrt{n}\epsilon_\infty$ such that any attack in the $\epsilon_\infty$ budget will also respect the $\epsilon_2$ budget. When working on images, other dissimilarity measures are: rotations, deformations, or color changes of the original image. Contrary to the $l_2$ or $l_\infty$, these measures are not based on a pixel-based coordinate system. However, it is possible to define *unrestricted attacks* based on these spatial dissimilarities, for example in [17].

In this work, we derived the partial isometry regularization for a classification task. The method can be extended to regression tasks by considering the family of multivariate normal distributions as the output space. On the probability simplex $\Delta^m$, the FIM is a metric with constant positive curvature, while it has constant negative curvature on the manifold of multivariate normal distributions [18].

Finally, the precise quantification of the robustness condition presented in Eqs. (4) and (6) paves the way to the development of a certified defense [19] in this framework. By strongly enforcing Eq. (6) on a chosen proportion of the training set, it may be possible to maximize the accuracy under the constraint of a chosen robustness level, which offers another solution to the robustness-accuracy trade-off [20], [21]. Certifiable defenses are a require step for the deployment of deep learning models in critical domains and missions, such as civil aviation, security, defense and healthcare, where a certification may be required.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we introduced an information geometric approach to the problem of adversarial robustness in machine learning models. The proposed defense consists of enforcing a partial isometry between the input space endowed with the Euclidean metric and the probability simplex endowed with the Fisher information metric. We subsequently derived a regularization term to achieve robustness during training. The proposed strategy is tested on the MNIST dataset, and shows considerable increase in robustness without harming the accuracy. A journal paper that extends the MNIST results to other Benchmark and real-world datasets is under preparation. Several attack methods will also be considered in addition to PGD. Although this work focuses on $l_2$ norm robustness, future work would consider other "distinguishability" measures.

## 6. REFERENCES


[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations*, 2018.

[4] Nicholas Carlini and David Wagner, "Towards Evaluating the Robustness of Neural Networks," in *IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[5] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow, "Adversarial spheres," in *6th International Conference on Learning Representations*, 2018.

[6] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *37th International Conference on Machine Learning (ICML)*, 2020.

[7] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a defense to adversarial perturbations against deep neural networks.," *CoRR*, 2015.

[8] Judy Hoffman, Daniel A. Roberts, and Sho Yaida, "Robust learning with jacobian regularization," *ArXiv*, 2018.

[9] Chaomin Shen, Yaxin Peng, Guixu Zhang, and Jinsong Fan, "Defending against adversarial attacks by suppressing the largest eigenvalue of fisher information matrix," *ArXiv*, 2019.

[10] Shun-ichi Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*, Springer New York, 1985.

[11] Ovidiu Calin and Constantin Udrişte, *Geometric Modeling in Probability and Statistics*, Springer International Publishing, 2014.

[12] N.N. Čencov, "Algebraic foundation of mathematical statistics," *Series Statistics*, vol. 9, no. 2, pp. 267–276, 1978.

[13] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein, "Adversarial training for free!," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.

[14] Chenxiao Zhao, P. Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen, "The Adversarial Attack and Detection under the Fisher Information Metric," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[15] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[16] Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier, "Parseval Networks: Improving Robustness to Adversarial Examples," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 854–863.

[17] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song, "Spatially transformed adversarial examples," in *International Conference on Learning Representations*, 2018.

[18] Lene Theil Skovgaard, "A Riemannian Geometry of the Multivariate Normal Model," *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 211–223, 1984.

[19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter, "Certified Adversarial Robustness via Randomized Smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, May 2019, pp. 1310–1320.

[20] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 7472–7482.

[21] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2019.