Robustness Bound of Classifiers on Riemannian Manifolds

1 Introduction

1.1 Motivation

- "Manifold hypothesis" motivates us to model the input space as a Riemannian (sub)manifold, say (M, g).
- Many datasets can be naturally assumed to be collections of samples of some Riemannian manifolds. E.g.:
  - in computer vision, the so-called covariance descriptors are elements of a manifold of symmetric-positive-definite (SPD) matrices,
  - subspace tracking (e.g. by incremental PCA) is naturally modeled as moving around on a Grassmann manifold. I.e. the solution of PCA is a point on a Grassmannian. One can also naturally think the orthogonal matrix given by PCA is a point on a Stiefel manifold.
- Generalizes the notion of robust classifier on arbitrary Riemannian manifolds.
- We are dealing with *general* perturbations on inputs, not necessarily have to be adversarial perturbations. I.e. our approach should work on any data type (not necessarily images) modeled on a Riemannian manifold.
- Adversarial examples found on this manifold are hypothetically stronger as they are very similar to the real data and thus harder to detect.
  - Gives rise to "on-manifold perturbation/attack".

1.2 Questions

- Especially for adversarial perturbation: What is the qualitative and quantitative differences between adversarial examples found on an image manifold and the usual adversarial examples?
- Can we derive a robustness bound under the above assumption?
  - E.g. following [Hein and Andriushchenko, 2017], let $p \in M$, bound $\|v\|_g$ for any $v \in T_pM$ and make a guarantee than under this condition, $\exp_p(v) =: q$ has the same class as p.
- How do we in practice exploit that bound to make classifiers more robust to perturbations (adversarial or not)? How do the resulting classifiers perform under input perturbations?

2 Robustness bound on Riemannian manifolds

Let (M, g) be a Riemannian n-manifold and $p \in M$. Let $f : M \to \mathbb{R}^k$ in $C^\infty(M)$ be a multi-class, smooth classifier. Denote f_i to be the i-th component of f. Let $c = \arg\max_{i=1,\dots,k} f_i(p)$ be the class which f predicts for p.

Definition 1 (Input perturbation). An ***input perturbation*** is a tangent vector $v \in T_pM$ such that $f_c(\exp_p(v)) \leq f_i(\exp_p(v))$ for some $i \neq c$. The point $\exp_p(v) =: q \in M$ is called the ***perturbed input***. If, in addition, M is an image manifold, and q is visually similar to p, then we call q an ***adversarial example***.

That is, the resulting perturbed inputs are constrained on the manifold. This definition also gives rise to a simple algorithm to find perturbed inputs or adversarial examples. Simply do the Riemannian gradient ascent on the manifold. Furthermore, Definition 1 also gives us the notion of the robustness of a classifier f.

Definition 2 (Robust classifier). f is ***ϵ-robust*** to input perturbations if for every $v \in T_pM$ with $\|v\|_g \leq \epsilon$ and for every $i \neq c$, $f_c(\exp_p(v)) > f_i(\exp_p(v))$.

Now, we give a lower bound of the norm of any tangent vector $v \in T_pM$ at p, such that the perturbed point $\exp_p(v)$ is an adversarial example.

Theorem 1. *Let (M, g) be a Riemannian manifold and $p \in M$. Let $f : M \to \mathbb{R}^k$ in $C^\infty(M)$ be a multi-class, smooth classifier. Denote f_i to be the i-th component of f. Let $c = \arg\max_{i=1,\dots,k} f_i(p)$ be the class which f predicts for p. Then, for any $i \neq c$, we have a lower bound on the norm of any $v \in T_pM$ such that $\exp_p(v)$ is an adversarial example, given by*

$$\|v\|_g \geq \frac{f_c(p) - f_i(p)}{\|\operatorname{grad} f_i|_p - \operatorname{grad} f_c|_p\|_g}. \tag{1}$$

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Proof. Let $v \in T_pM$ be arbitrary. By Taylor's theorem, for any $i = 1, \dots, k$,

$$f_i(\exp_p(v)) \approx f_i(p) + \langle \operatorname{grad} f_i|_p, v \rangle_g \,.$$

To obtain an adversarial example, we want $f_i(\exp_p(v)) \geq f_c(\exp_p(v)$. Thus,

$$\begin{aligned} f_c(p) - f_i(p) &\leq \langle \operatorname{grad} f_i|_p, v \rangle_g - \langle \operatorname{grad} f_c|_p, v \rangle_g \\ &= \langle \operatorname{grad} f_i|_p - \operatorname{grad} f_c|_p, v \rangle_g \\ &\leq \|\operatorname{grad} f_i|_p - \operatorname{grad} f_c|_p\|_g \|v\|_g \,, \end{aligned}$$

where we have used the bilinearity of and the Cauchy-Schwarz inequality on the inner product $\langle \cdot, \cdot \rangle_g$. Finally, by rearranging, we obtain the desired result. □

Note that the bound is probably loose. Indeed, we could set the bound to be arbitrarily large and the condition still holds true. However, at least, up to the first-order approximation of f, Elsayed et al. [2018] showed (on the Euclidean space) that the bound in Theorem 1 gives us the *minimum* perturbation strength such that $\exp_p(v)$ lies on the decision boundary. We state the adaptation of their result on Riemannian manifolds, in the following proposition.

Proposition 1 (Adaptation of Elsayed et al. [2018]). *Assumption as in Theorem 1. For any $i \neq c$,*

$$\frac{f_c(p) - f_i(p)}{\|\operatorname{grad} f_i|_p - \operatorname{grad} f_c|_p\|_g}$$

is the solution of

$$\min_v \|v\|_g \;\; \textit{s.t.} \;\; f_c(p) + \langle \operatorname{grad} f_c|_p, v \rangle = f_i(x) + \langle \operatorname{grad} f_i|_p, v \rangle \,.$$

□

Corollary 1 (Computation of the bound). *Assumption as in Theorem 1. Denote G to be the matrix representation of g and d_i to be the column vector representation of $df_i|_p$ for all $i = 1, \dots, k$. Then, for any $i \neq c$ and $v \in T_pM$ satisfying*

$$\|v\|_g \geq \frac{f_c(p) - f_i(p)}{[(d_i - d_c)^T G^{-1} (d_i - d_c)]^{\frac{1}{2}}} \,, \tag{2}$$

$\exp_p(v) \in M$ is an adversarial example.

Proof. We focus on the denominator of eq. (1). By definition of Riemannian gradient and norm:

$$\begin{aligned} \|\operatorname{grad} f_i|_p - \operatorname{grad} f_c|_p\|_g &= \|G^{-1} d_i - G^{-1} d_c\|_g \\ &= \|G^{-1}(d_i - d_c)\|_g \\ &= [(G^{-1}(d_i - d_c))^T G (G^{-1}(d_i - d_c))]^{\frac{1}{2}} \\ &= [(d_i - d_c)^T G^{-1} G G^{-1} (d_i - d_c)]^{\frac{1}{2}} \\ &= [(d_i - d_c)^T G^{-1} (d_i - d_c)]^{\frac{1}{2}} \,. \end{aligned}$$

We end the proof by substituting the above equation to eq. (1). □

Let ϵ be the value of the bound in Theorem 1. The theorem says that a classifier under conditions stated in Theorem 1 will be guaranteed to be ϵ-robust. It is, therefore, in our best interest to maximize ϵ, when training the classifier. Two ways of doing that are

1. by following Hein and Andriushchenko [2017], i.e. by adding regularization term to minimize the denominator, while maximizing the numerator of eq. (1),
2. or by following Elsayed et al. [2018], i.e. by using the bound as the loss function directly, as a large-margin loss.

3 Related work

Hein and Andriushchenko [2017] presented one of the earliest results on robustness bound of classifiers. They incorporated the bound as a regularization term during training. Meanwhile, similarly, Elsayed et al. [2018] derived a robustness bound similar to ours and Hein and Andriushchenko [2017] for deep neural networks. The main point in their proof is the linearization argument: They approximated the classifier function up to the first-order and showed that indeed their bound is optimal under such approximation. In contrast to Hein and Andriushchenko [2017], they incorporate their bound as a max-margin loss during training.

Geometric analyses of adversarial examples have been conducted before [3, 4, 5]. However, they did not consider the input or latent spaces of a classifier to be Riemannian manifolds and thus their analyses are limited to the Euclidean geometry or its variants using various L^p norms. Fawzi et al. [2018a] and Moosavi-Dezfooli et al. [2018b], meanwhile presented an

empirical study and the robustness bound based on the geometry of the decision boundary as a hypersurface in the input space, and not the input space itself.

Computation of adversarial examples on the latent space of a generative model has also been studied by [8, 9, 10, 11, 12, 13] and [14]. They showed that there exist adversarial examples that lie on the same manifold as the clean inputs. These examples are thus more believable and thus harder to detect. However, they assume that the latent space is an L^p metric space and only loosely define the term "manifold". In contrast, our work provides a principled and rigorous analysis of adversarial examples on the input and latent spaces through the lens of Riemannian geometry.

Zhao et al. [2018b] meanwhile, proposed adversarial attack and detection methods by assuming the input space to be equipped with the Fisher information metric, thus assumed to be a Riemannian manifold. In contrast, we do not assume a specific manifold and metric to be used, thus our work can be seen as the generalization of their work. Moreover, while Zhao et al. [2018b] focused on the attack and detection methods, we study of the analysis of on-manifold adversarial examples and the robustness bound of classifiers under such examples. Kanbak et al. [2018] also studied the Riemannian geometry of adversarial examples although they focus on the Lie group of transformations of inputs. That is, they focused on finding a geometric transformations such the transformed images are adversarial examples. In contrast, we focus on finding tangent vectors such that their image under the exponential map are adversarial examples.

4 Experiment sketches

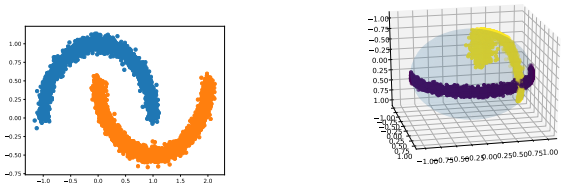
Given a dataset $\mathcal{D} = \{(x, y) : x \in M, y = 1, \dots, K\}$ where the inputs are sampled from a Riemannian manifold (M, g), we would like to find perturbed inputs or adversarial examples on M, and verify that they are indeed qualitatively better than vanilla perturbations. We then want to try to defend against this attack by using the result from Theorem 1.

In this section, we will present experiments on:

1. a simple, toy dataset sampled from a model Riemannian manifold other than $\mathbb{R}^n$, e.g. spheres.
2. on a given *real world* dataset sampled from a predefined manifold.
   - E.g. inputs are coming from a Grassmannian [Huang et al., 2018] or the manifold of positive-definite matrices [Huang and Van Gool, 2017].
3. Experiment on a standard dataset, e.g. MNIST, with a learned metric tensor [Hauberg et al., 2012].
4. Experiment on adversarial examples on latent space of a generative model, with metric tensor given by the pullback metric by the generator [Tosi et al., 2014, Arvanitidis et al., 2017, Chen et al., 2018, Shao et al., 2018].

4.1 Classification on a sphere

We construct the dataset as follow: We create a standard moon-shaped dataset in $\mathbb{R}^2$ (Figure 1) and project it onto $\mathbb{S}^3$ by applying the inverse stereographic projection.

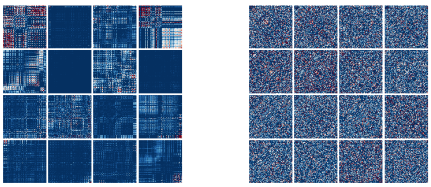


(a) On $\mathbb{R}^2$ (b) On $\mathbb{S}^2$

Figure 1: Moons dataset on a plane and its projection onto $\mathbb{S}^2$.

We train a two-layer NN with 100 hidden units and `tanh` nonlinearity. We achieved 89% test accuracy in this experiment. We attack the model using the on-manifold perturbation and the standard Euclidean perturbation on 2000 test points and noticed that the perturbation were successful in 1237 and 1373 cases respectively. However, we noticed that in general the Euclidean attack yields perturbed inputs *outside* the manifold, as indicated by the mean of their norms 0.93 ± 0.131. Meanwhile, as expected the on-manifold perturbation successfully found all the adversarial examples *on* the manifold (i.e. with norm 1 ± 0).

By incorporating the bound as a regularization, following Hein and Andriushchenko [2017], we found that the classifier is able to defend itself better to both on-manifold and Euclidean perturbations. Now, only 470 and 532 perturbations are successful out of the 2000 test points, for the on-manifold and the Euclidean perturbation, respectively. Meanwhile, the accuracy of the classifier is now 86%.



(a) On-mfld, diff. (b) Euclidean, diff.

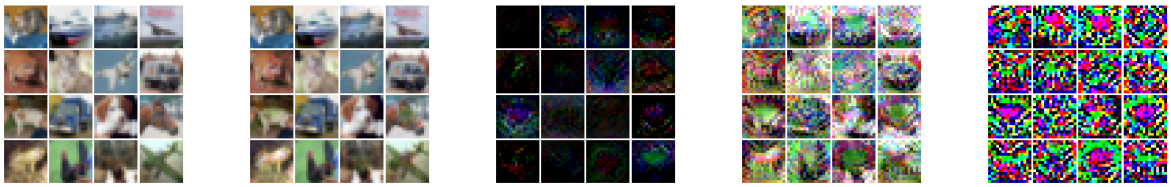
Figure 2: Difference between the clean inputs and their perturbed versions obtained via on-manifold perturbation and Euclidean perturbation (under Frobenius norm), magnified 1000 times.

4.2 Classification of SPD matrices

In computer vision, covariance matrices, which are (SPD) matrices, have been widely used as image descriptors [24, 18]. The space of SPD matrices is a Riemannian manifold $(Sym^{++}(n, \mathbb{R}), g)$, where $Sym^{++}(n, \mathbb{R})$ denotes the space of $n \times n$ real SPD matrices and g is the choice of the Riemannian metric for $Sym^{++}(n, \mathbb{R})$, e.g. the log-Euclidean [Arsigny et al., 2006] metric.

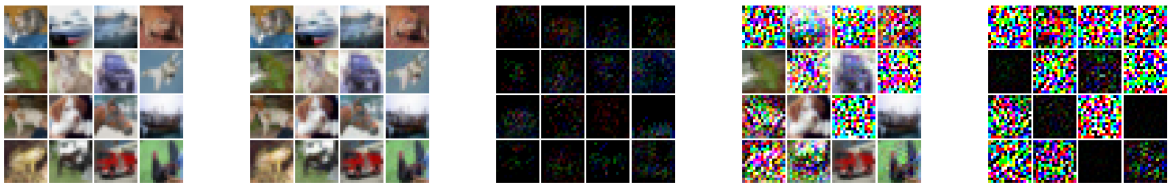
We consider the emotion recognition problem on AFEW dataset Dhall et al. [2011], where the covariance descriptor of an image is classified using a specialized deep network for processing SPD matrix as an input, dubbed SPDNet Huang and Van Gool [2017]. We compare the difference between the clean inputs and their perturbed versions obtained via on-manifold perturbation and Euclidean perturbation (under Frobenius norm), in Figure 2. We noticed that often, the perturbations found on the manifold are very small and structured, compared to the Euclidean perturbations which in general are more apparent and random.

4.3 Adversarial examples on a learned Riemannian manifold



(a) Real (b) On-mfld, adv. (c) On-mfld, diff. (d) L^∞, adv. (e) L^∞, diff.

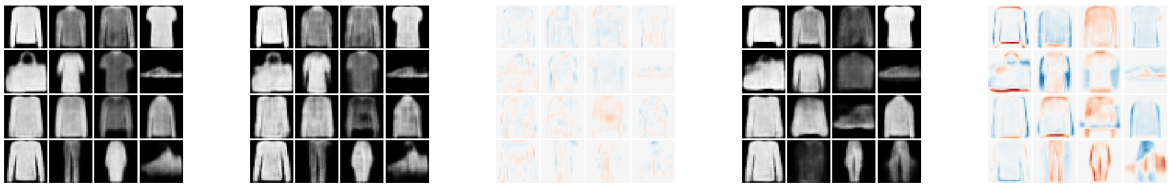
Figure 3: Adversarial examples found on the input space of CIFAR-10 equipped with a learned metric tensor and the standard L^∞ norm.



(a) Real (b) On-mfld, adv. (c) On-mfld, diff. (d) L^2, adv. (e) L^2, diff.

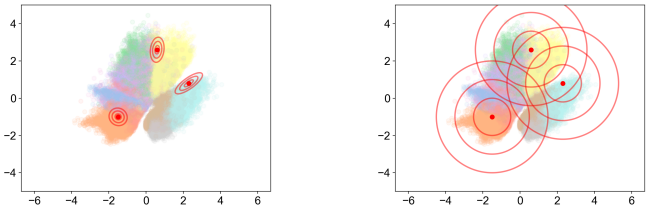
Figure 4: Adversarial examples found on the input space of CIFAR-10 equipped with a learned metric tensor and the standard L^2 norm, where the bounds for the norms are picked such that the resulting model accuracies are comparable ($\sim 18\%$).

Assume that our dataset consists of images in $\mathbb{R}^n$ (assume flattened-vector representation). We can endow $\mathbb{R}^n$ with a suitable Riemannian metric, learned from data. That is, we would like to learn a smooth SPD matrix field on $\mathbb{R}^n$ [19]. One of the popular approach to do metric learning is by partitioning the input space onto k regions and learn a metric tensor for each region. One can then define a smooth weighting scheme at each point in $\mathbb{R}^n$ and average the metrics w.r.t. these weights. However, the exponential and logarithm map can still be very expensive as we have to solve high-dimensional initial value problem and boundary value problem, respectively. Fortunately, as the base manifold in this case is $\mathbb{R}^n$, one can use a particularly simple first-order approximation of the exponential map, called retraction, which essentially moving following a straight line.



(a) Real (b) Pullback, adv. (c) Pullback, diff. (d) Euclidean, adv. (e) Euclidean, diff.

Figure 5: Adversarial examples found on the latent space of a VAE equipped with the pullback and Euclidean metric. The perturbation bound for $\|v\|$ is picked such that the test accuracy wrt. those two metrics are about the same ($\sim 36\%$).



(a) Pullback (b) Euclidean

Figure 6: Spheres in tangent spaces of the latent space (using the canonical identification of $T_p\mathbb{R}^2$ with $\mathbb{R}^2$) with radius 1, 2, and 3, under the norm induced by the (a) pullback metric and (b) Euclidean metric.

We use a resized-version of CIFAR-10 dataset ($16 \times 16 \times 3$ images) in this experiment. To learn the metric tensors, we partition the inputs w.r.t. their labels, and compute the mean and the covariance matrix for each of the partition. The metric tensor for each partition is then its inverse covariance matrix. The weighting scheme follows Hauberg et al. [2012], where an RBF is computed for each partition, with the anchor point of the RBF is the mean of the partition.

We compare the on-manifold adversarial attack with the usual L^∞-norm attack ($\epsilon = 0.3$). The results are presented in Figure 3. We noticed that by respecting the geometry of the input space, we get more believable adversarial examples, while still reducing drastically the model's accuracy. Note that using L^2-norm attack failed in this case, as we found that the resulting perturbations are very small and the model accuracy did not change.

Finally, we also show the effectiveness of the on-manifold attack compared to the L^2 attack in Figure 4, where we picked the bound for the perturbation norms separately, such that the resulting accuracies match (at around 18%). Evidently, given the same accuracy target, the adversarial examples produced by the on-manifold attack are more subtle and imperceptible to those produced by the standard L^2 attack. Thus we conclude that following the geometry of the input manifold, even if only by a crude approximation, yields a better quality adversarial examples.

4.4 Adversarial examples on latent spaces

Several studies [21, 23, 22] have been done to analyze the geometry of the latent space of deep generative models (such as VAEs [27]). Specifically, they treat the latent space as a Riemannian submanifold $(\mathbb{R}^d, g)$ of the Euclidean space $(\mathbb{R}^n, \bar{g})$, where $\bar{g}$ is the Euclidean metric. Assuming the conditions hold[1], we can see the generator function $F : \mathbb{R}^d \to \mathbb{R}^n$ of a VAE to be a smooth immersion. The metric tensor g is then simply the pullback of $\bar{g}$ by F. That is, at any $p \in \mathbb{R}^d$ and any $v, w \in T_p\mathbb{R}^d$

$$g(v, w) = F^* \bar{g}(v, w) = \bar{g}(dF(v), dF(w)) \tag{3}$$

where dF is the pushforward of F. In coordinates, $g = \bar{g}_{ij}\, dF^i \otimes dF^j$ and dF can be represented by the Jacobian matrix J of F, thus in matrix notation, $g(v, w)$ can be written as

$$v^T G w = (Jv)^T I (Jw) = v^T J^T J w \implies G = J^T J \in \mathbb{R}^{d \times d}. \tag{4}$$

Assume that F governs the underlying generative process of input x of a neural network. We can therefore apply on-manifold perturbation on the underlying latent space to obtain adversarial examples, instead of on the input space $\mathbb{R}^n$ as it is usually done. By following the curvature of the latent space, we can hypothesize that the resulting adversarial examples will be more "believable", i.e. very close to the true examples, and thus harder to detect.

We train a VAE and a DenseNet-121 classifier NN on Fashion-MNIST (FMNIST). The classifier achieved 0.921 accuracy. We then carry out on-manifold perturbation on the latent representations of samples in the test set. The resulting "adversarial

[1] The map must be smooth and has constant rank injective Jacobian.

Table 1: Vanilla and regularized models' accuracy under clean, on-manifold, and Euclidean adversarial test sets.

Dataset	Vanilla			Regularized		
	Clean	**Adv.-Mfld.**	**Adv.-Eucl.**	**Clean**	**Adv.-Mfld.**	**Adv.-Eucl.**
Moons-sphere	0.890	0.382	0.314	0.860	0.765	0.734
CIFAR-10	0.627	0.199	0.131	0.628	0.153	0.083
FMNIST	0.906	0.589	0.503	0.903	0.597	0.542

latent representations" are then mapped to $\mathbb{R}^n$ by F. The resulting adversarial examples, constructed using the constrains for the perturbation norms picked to match the resulting accuracies under both metrics ($\sim 36\%$), are presented in Figure 6. As we can see, by taking into account the geometry of the latent space, the quality of the adversarial examples are much higher compared to when assuming that the latent space is a Euclidean space ($\mathbb{R}^d$ with the Euclidean metric).

In this case, we have another corollary of Theorem 1, which gives us the robustness bound in term of the perturbation norm on the latent space.

Corollary 2 (Robustness bound on latent spaces). *Assumption as in Theorem 1. Let N be a smooth manifold and let $F : N \to M$ be a smooth immersion. We equip N with the pullback metric $\hat{g} := F^* g$. Then, for any $i \neq c$ and any $q \in N$, we have a lower bound on any $w \in T_q N$ such that $F \circ \exp_q(w)$ is an adversarial example, given by*

$$\|w\|_{\hat{g}} \geq \frac{(f_c \circ F)(q) - (f_i \circ F)(q)}{\|\text{grad}\,(f_i \circ F)|_q - \text{grad}\,(f_c \circ F)|_q\|_{\hat{g}}}\,. \tag{5}$$

In coordinates, if $\hat{G}$ is the matrix representation of $\hat{g}$, d_i is the column vector representation of df_i, and J is the Jacobian matrix of F, then

$$\|w\|_{\hat{g}} \geq \frac{(f_c \circ F)(q) - (f_i \circ F)(q)}{[(d_i - d_c)^T J^T \hat{G}^{-1} J (d_i - d_c)]^{\frac{1}{2}}}\,. \tag{6}$$

□

Remark 1. The term $J(d_i - d_c)$ can be obtained "for free" by automatic differentiation.

4.5 Effects of regularization derived from the bound

As noted by Hein and Andriushchenko [2017], the bound above can be used to regularize the classifier during training. Specifically, we want to minimize the denominator of the bound, which can easily be computed thanks to Corollary 1 and Corollary 2. We present the results in Table 1.

References

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276, 2017.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems*, pages 842–852, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. *arXiv preprint arXiv:1811.09716*, 2018a.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Marc Khoury and Dylan Hadfield-Menell. On the geometry of adversarial examples. *arXiv preprint arXiv:1811.00525*, 2018.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018a.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Robustness of classifiers to universal perturbations: A geometric perspective. 2018b.

Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 36–42. IEEE, 2018.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, pages 1178–1187, 2018b.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. *arXiv preprint arXiv:1812.00740*, 2018.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018.

Bing Yu, Jingfeng Wu, Jinwen Ma, and Zhanxing Zhu. Tangent-normal adversarial regularization for semi-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Andrew Ilyas, Ajil Jalal, Eirini Asteri, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=H1BLjgZCb.

Chenxiao Zhao, P Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. The adversarial attack and detection under the fisher information metric. *arXiv preprint arXiv:1810.03806*, 2018b.

Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4441–4449, 2018.

Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on grassmann manifolds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Søren Hauberg, Oren Freifeld, and Michael J Black. A geometric take on metric learning. In *Advances in Neural Information Processing Systems*, pages 2024–2032, 2012.

Alessandra Tosi, Søren Hauberg, Alfredo Vellido, and Neil D Lawrence. Metrics for probabilistic geometries. *arXiv preprint arXiv:1411.7432*, 2014.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick Smagt. Metrics for deep generative models. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1540–1550, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, and Xilin Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *International conference on machine learning*, pages 720–729, 2015.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56 (2):411–421, 2006.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, 2:1, 2011.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, April 2014.