frontiers

Supplementary Material

1 PROOF OF THEOREM 1

THEOREM 1 (Risk Decomposition). *Let $\mathcal{M}$ be a smooth compact manifold in $\mathbb{R}^D$, and let data be drawn from $\mathcal{M} \times \{-1, 1\}$ according to some distribution p. There exists a $\Delta > 0$ depending only on $\mathcal{M}$ such that the following statements hold for any $\epsilon < \Delta$. For any score function f satisfying assumption A,*

(i)

$$R_{adv}(f, \epsilon) \leq R_{std}(f) + R^{nor}_{adv}(f, \epsilon) + R^{in}_{adv}(f, 2\epsilon) + \textit{NNR}(f, \epsilon).$$

(ii)If $R^{nor}_{adv}(f, \epsilon) = 0$, then

$$R_{adv}(f, \epsilon) \leq R_{std}(f) + R^{in}_{adv}(f, 2\epsilon)$$

Proof of i): We first address the existence of the constant Δ that only depends on $\mathcal{M}$ in the theorem statement.

DEFINITION 1 (Tubular Neighborhood). *A tubular neighborhood of a manifold $\mathcal{M}$ is a set $\mathcal{N} \subset \mathbb{R}^D$ containing $\mathcal{M}$ such that any point $z \in \mathcal{N}$ has a unique projection $\pi(z)$ onto $\mathcal{M}$ such that $z - \pi(z) \in N_{\pi(z)}\mathcal{M}$.*

By Theorem 11.4 in Bredon (2013), we know that there exists $\Delta > 0$ such that $N := \{y \in \mathbb{R}^D : dist(y, \mathcal{M}) < \Delta\}$ is a tubular neighborhood of $\mathcal{M}$. This also implies that for any $0 < \epsilon < \Delta$, the normal line segments of length ϵ at any two points $x, x' \in \mathcal{M}$ are disjoint, a fact that will be used later.

The Δ guaranteed by Theorem 11.4 is the Δ referred to in our theorem, and the budget $\epsilon > 0$ is constrained to be at most Δ.

Next we consider the left hand side, the adversarial risk:

$$R_{adv}(f, \epsilon) := \mathop{\mathbb{E}}_{(x,y)\sim p} \mathbb{1}(\exists x' \in B_\epsilon(x) : f(x')y \leq 0)$$

Denote by $E(x, y)$ the event that $\exists x' \in B_\epsilon(x) : f(x')y \leq 0$.

We will write the indicator function above as the sum of indicator functions of four events. Specifically, define by $E_1(x, y), E_2(x, y), E_3(x, y), E_4(x, y)$ the following four events:

- $E_1(x, y)$: $f(x)y \leq 0$.
- $E_2(x, y)$: $f(x)y > 0$ and $\exists x' \neq x \in B_\epsilon(x)$ such that $x' - x \in N_x\mathcal{M}$ and $f(x')y \leq 0$.

For the next two cases, let $x' \neq x \in B_\epsilon(x)$ be such that $x' - x \notin N_x\mathcal{M}$ and $f(x')y \leq 0$ (if such an x' exists). Let $x'' = \pi(x')$ be the unique projection of x' onto $\mathcal{M}$. Note that $x'' \neq x$. Define:

- $E_3(x, y)$: $f(x'')y \leq 0$.
- $E_4(x, y)$: $f(x'')y > 0 \iff f(x'')f(x') \leq 0$.

LEMMA 1.

$$\mathbb{1}(E(x, y)) = \mathbb{1}(E_1(x, y)) + \mathbb{1}(E_2(x, y)) + \mathbb{1}(E_3(x, y)) + \mathbb{1}(E_4(x, y))$$

PROOF. Assume $E(x, y)$ occurs, i.e, $\exists x' \in B_\epsilon(x) : f(x')y \leq 0$. Either $x' = x$ satisfies the condition (which is event E_1) or some $x' \neq x$ satisfies the condition.

Now we further divide into the case when $f(x)y > 0$ and $x' - x \in N_x\mathcal{M}$ (which is event E_2), or $f(x)y > 0$ and $x' - x \notin N_x\mathcal{M}$. In the latter case, note that $x'' = \pi(x')$ cannot equal x as otherwise $x' - x$ would be in the normal space at x, since the projection map is unique inside the tubular neighborhood. Thus x'' is well-defined, and it is easy to see that the last two cases are disjoint and cover this remaining case. Thus we have shown that if $E(x, y)$ occurs, then one of the four disjoint events E_i must occur, proving the lemma.

Finally we have the following lemma, which completes the proof of the theorem after combining with Lemma 1.

LEMMA 2. *The following relation holds between the risk and the expectation of the indicator functions in Lemma 1*

1. $\mathop{\mathbb{E}}_{(x,y)\sim p} \mathbb{1}(E_1(x, y)) = R_{std}(f)$
2. $\mathop{\mathbb{E}}_{(x,y)\sim p} \mathbb{1}(E_2(x, y)) \leq R^{nor}_{adv}(f, \epsilon)$
3. $\mathop{\mathbb{E}}_{(x,y)\sim p} \mathbb{1}(E_3(x, y)) \leq R^{in}_{adv}(f, 2\epsilon)$
4. $\mathop{\mathbb{E}}_{(x,y)\sim p} \mathbb{1}(E_4(x, y)) \leq NNR(f, \epsilon)$

PROOF. 1) and 2) follow by definitions of standard adversarial risk and normal adversarial risk, respectively. Consider the setting of $E_3(x, y)$: i.e., $f(x)y > 0$, the adversarial perturbation x' is not in the normal direction (so $f(x')y \leq 0$), and $f(x'')y \leq 0$. Observe that by the triangle inequality, $d(x, x'') \leq d(x, x') + d(x', x'') \leq \epsilon + \epsilon = 2\epsilon$, simply because a) x' is within the ϵ-ball of x, and b) x'' is closer to x' than x.

This means that there is a point $x'' \in B^{in}_{2\epsilon}(x)$ such that $f(x'')y \leq 0$. The expectation over a random $(x, y) \sim p$ of this event is clearly at most $R^{in}_{adv}(f, 2\epsilon)$ (the inequality need not be tight because x may have adversarial perturbation within 2ϵ and also satisfy some other events like E_1).

Lastly, by the definition of the NNR, we see that $A(x, y)$ occurs when $E_1(x, y)$ or $E_2(x, y)$ do not. Also $C(x, y)$ implies that the event $E_3(x, y)$ does not occur. We are now in the situation where x'' is within 2ϵ of x, $f(x')y \leq 0$, and $f(x'')y > 0$. But this implies that $f(x'')f(x') \leq 0$, and since $x' \in B^{nor}_{\epsilon}(x'')$, it implies that $B(x, y)$ occurs. Thus all of $A(x, y)$, $B(x, y)$ and $C(x, y)$ occur, which is the definition of NNR.

Proof of ii)

If $R^{nor}_{adv}(f, \epsilon) = 0$, we claim that $NNR(f, \epsilon) = 0$. Setting these two terms to zero in i) proves ii).

Note that although $R^{nor}_{adv}(f, \epsilon) = 0$, it does not imply that there are no normal adversarial perturbations for any x— it just means that the measure of such x with normal adversarial perturbation is zero.

Also note that $R^{nor}_{adv}(f, \epsilon) = 0$ does not exclude $A(x, y)$ or $C(x, y)$ from occurring (in fact A occurs for almost all x). Thus the proof will focus on the measure of points where $B(x, y)$ can occur. We will prove the following lemma, which will complete the proof of the theorem.

LEMMA 3. *Let (x, y) be such that $B(x, y)$ occurs, i.e., there exist $x' \in B_\epsilon(x)$ and $x'' = \pi(x)$ such that $f(x')y \leq 0$, $f(x'')y > 0$ and $d(x, x'') \leq 2\epsilon$. Then $C(x, y)$ cannot occur, i.e., there exists a point $w \in B^{in}_{2\epsilon}(x)$ such that $f(w)y \leq 0$. Consequently, $NNR(f, \epsilon) = 0$.*

PROOF. We first claim that if $B(x, y)$ occurs, it must be the case that $f(x'') = 0$. Assuming this, if $f(x'') = 0$, then by Assumption A we know there exists an $s \in B_\epsilon(x'') \cap B_{2\epsilon}(x)$ such that $f(s)y \leq 0$, which imply that $C(x, y)$ cannot occur. This will complete the proof of the lemma.

To prove that $f(x'') = 0$, consider what happens if $f(x'') \neq 0$. Assume first that $f(x') \neq 0$, and note that $f(x')f(x'') \leq 0$. By continuity of f, there exist open neighborhoods $U \ni x''$ and $V \ni x'$ such that f has the same sign on all of U and the same sign on all of V, i.e., $sign(f|U) = sign(f(x''))$ and $sign(f|V) = sign(f(x'))$.

Consider the normal bundle on U defined as the set $U' = \{y \in \mathcal{M}_\Delta : \pi(y) \in U\}$. In other words, U' is the union of the normal line segments passing through points in U (here $\mathcal{M}_\Delta$ denotes the tubular neighborhood of $\mathcal{M}$). Note that U' is an open set.

Define $W' = U' \cap V$, and $W = \pi(W')$. $W \subset \mathcal{M}$ is an open set, but for every $w \in W$, there exists a point $w' \in W' \cap B^{nor}_{\epsilon}(w)$ such that $f(w')f(w) \leq 0$. Therefore there exists anormal adversarial perturbation for every point in W. Since the measure of W is not zero, this contradicts the fact that $R^{nor}_{adv}(f, \epsilon) = 0$.

The proof is completed by observing that in the remaining case when $f(x'') \neq 0$ but $f(x') = 0$, there must exist (by assumption A) a point w near x' such that $f(w) \neq 0$ and $f(w)y < 0$. This lands us in the previous case, which we showed contradicts the hypothesis that $R^{nor}_{adv}(f, \epsilon) = 0$.

Remark: In Corollary 1, $\mu(\overline{Z^{nor}(f,\epsilon)} \cap B_{2\epsilon}(Z^{nor}(f,\epsilon))$ is the NNR under deterministic case. Therefore, Corollary 1 follows directly from the proof of Theorem 1.

2 ADDITIONAL EXPERIMENTS

In the main paper, we leave some experimental results to discuss in this supplementary materials. In the following section, we compare different ways of generating in-manifold attack data.

2.1 In-Manifold Attack Algorithm

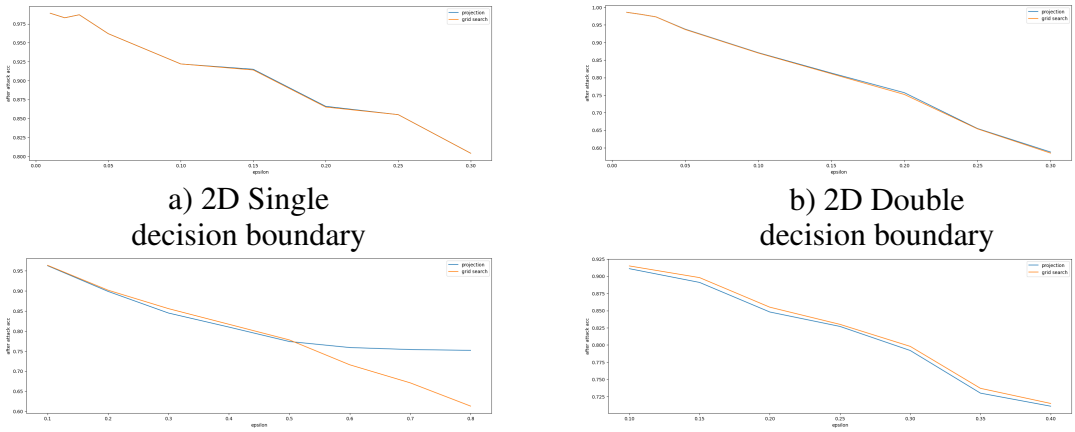
To estimate the *in-manifold adversarial risk*, we have tested two potential algorithms for generating in-manifold adversarial examples. We present our observations on these two methods. Our empirical study in the paper leverage one of the two methods presented below, which generates a more powerful in-manifold adversarial example.

One way to generate the adversarial samples is by brutal force. We use the grid search method to search the $B_\epsilon^{in}(x)$ region and find the maximum loss point in that region. We treat the maximum loss point as the in-manifold adversarial data. We call this approach the grid search method. Another approach we name as the projected method. We set the step size of the grid search proportional to the perturbation budget ϵ. In general, we search 100 points in 1D cases and 400 points in the 2D manifold. In the projected method, we first use a general adversarial attack algorithm to generate adversarial data in ambient space. Then we project the generated adversarial example back to the manifold and return the results as our in-manifold adversarial data. In the following experiment, we use PGD as our generator of adversarial data in ambient space. Both methods will find in-manifold data that is adversarial to the given model. The rest of the experiment settings follow Section 3 in the main paper.

In Figure S1 we plot the after-attack accuracy of these two in-manifold attack methods. The experiments follow the same setting as the one we described in the main paper. We could observe that the grid search is slightly stronger in the 3D single boundary case and equivalent to the projection method in the rest of the cases. In the graph, the after-attack accuracy of the grid search method matches with the projection methods in the 2D case. And in the 3D case, when the ϵ is larger than 0.5, then the grid search method achieves smaller after attack accuracy. This is due to the projection method searching the adversarial example in a smaller in-manifold ball. In other words, it hasn't fully explored the ϵ ball around the original data point. Therefore we could observe this small gap between these two methods. In the paper, we rely on the grid search method for generating in-manifold adversarial examples.

2.2 Real-World Data under L_∞ norm

In this section, we additionally outline various adversarial risks within the context of L_∞ norm attacks in Table S1. Our theoretical findings remain consistent when applied to the FASHIONMNIST dataset. However, these findings did not hold up for the other two datasets.



a) 2D Single decision boundary

b) 2D Double decision boundary

c) 3D Single decision boundary

d) 3D Double decision boundary

Figure S1: We compare the grid search method and projection method to generate in-manifold attack data. The first row is after attack accuracy on the 2D data set. The blue line is the accuracy of the projection approach. Orange is for the grid search method. The ϵ range is smaller than the range we choose in the discussion of the main paper. This is because ϵ-budget is larger than 0.05. The after-attack accuracy remains zero. The lower row is after attack accuracy on two different 3D data sets.

Table S1. Listing standard risk and different adversarial attack risks under L_∞ norm. Each risk is plotted in separate columns, and in the last column, we sum up the standard risk, in-manifold adversarial risk, and normal adversarial risk for comparison with the general adversarial risk.

Dataset	L_∞ Attack Risk	Standard Risk	In-Manifold Adversarial Risk	Normal Adversarial Risk	Sum of RHS
MNIST	0.9997	0.0076	0.0262	0.5654	0.5992
FASHIONMNIST	0.9862	0.0522	0.083	0.8633	0.9985
SVHN	0.98	0.0326	0.0988	0.1843	0.3157

This leads us to propose that the decision boundary within the FASHIONMNIST dataset closely aligns with the data manifold, making it susceptible to attacks from both directions.

2.3 Manifold Reconstruction

To verify the accuracy of our approximation, we plotted the difference between the reconstructed images and the original images under L_2 and L_∞ norms in Figure S2. The horizontal axes represent the distance between input and reconstructed images, while the vertical axes represent the number of images falling within a specific distance from the original images.

From the images in Figure S2, we observe that for gray-scale images the majority of the reconstructed examples are very close to the original images. Specifically, under L_2 norm with a perturbation budget of 1.5, nearly 95% of the reconstructed images for MNIST and FASHIONMNIST datasets fall within the 1.5 distance away from the original images, demonstrating a accurate approximation of the underlying data manifold.

Similarly, under L_∞ norm, a significant percentage of the reconstructed images also remain within the specified perturbation budget. However, for the SVHN dataset, most of the images exceed the perturbation budget, resulting in visible differences between the reconstructed and original images.

Furthermore, upon comparing the reconstructed images to the original ones, we noticed that the reconstructed SVHN images appear blurrier than their corresponding original images. This is a common issue with Autoencoders, where certain fine details may not be accurately captured during the reconstruction process.

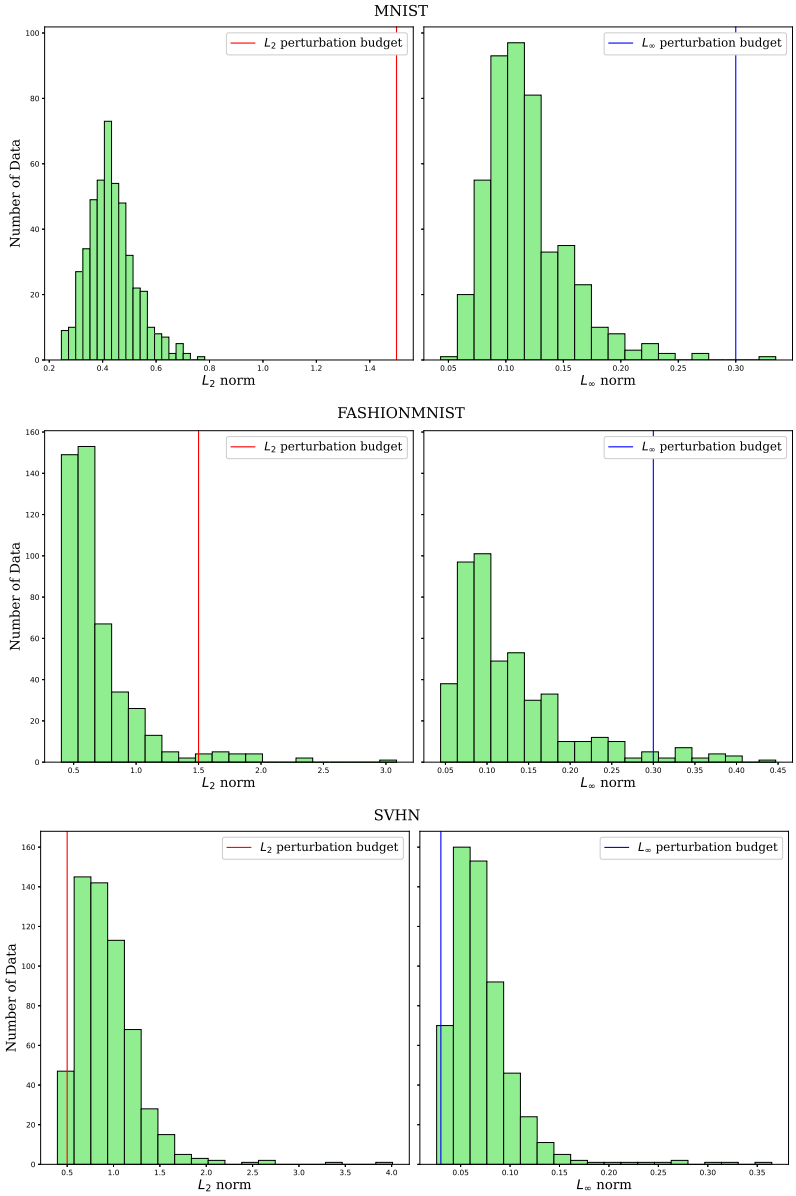


Figure S2: In this plot, we present histograms for the differences between input images and their reconstructed counterparts, using both the L_2 and L_∞ norms. The figure reveals that the invisibility of differences in MNIST and FASHIONMNIST datasets is due to the majority of the differences falling within the L_2 or L_∞ budget for gray-scale images. Conversely, in the case of SVHN, only 1% of the differences remain under the perturbation budget, making them easily noticeable.