A law of adversarial risk, interpolation, and label noise

Daniel Paleka [*1] **Amartya Sanyal** [*1 2]

Abstract

In supervised learning, it is known that label noise in the data can be interpolated without penalties on test accuracy. We show that interpolating label noise induces adversarial vulnerability, and prove the first theorem showing the dependence of label noise and adversarial risk in terms of the data distribution. Our results are almost sharp without accounting for the inductive bias of the learning algorithm. We also show that inductive bias makes the effect of label noise much stronger.

1. Introduction

Label noise is ubiquitous in data collected from the real world. Such noise can be a result of both malicious intent as well as human error. A relatively benign form of such noise is one that is distributed uniformly randomly on the data distribution. The well-known work of Zhang et al. [29] observes that training overparameterised neural networks with gradient descent can memorize large amounts of label noise without increased test error. Recently, Bartlett et al. [2] investigated this phenomenon and termed it *benign overfitting*: perfect interpolation of the noisy training dataset still leads to satisfactory generalization for overparameterized models. A long series of works [7, 11, 17] have focused on providing generalisation guarantees for models that interpolate data under uniform label noise. This gives some hope that noisy training data does not hurt the test error of overparameterized models, and therefore such models can be deployed in the real world.

Adversarial vulnerability is a practical security threat [14, 24, 8] for deploying machine learning algorithms in critical environments. An adversarially vulnerable classifier, that is accurate on the test distribution, can be forced to err on carefully perturbed inputs even when the perturbations are *small*. This has motivated a large body of work towards improving the *adversarial robustness* of neural networks [10, 19, 26, 20, 6]. Despite the empirical advances, the theoretical guarantees on robust defenses are still poorly understood.

Consider the setting of uniformly random label noise. Under certain distributional assumptions, Sanyal et al. [21] claim that with moderate amount of label noise, when training classifiers to zero training error, the adversarial risk is always large, even when the test error is low. However, it is not clear whether their distributional assumptions are realistic and if their result is tight. To responsibly deploy machine learning models in the real world, it is important to understand the extent to which a common phenomenon like label noise can adversely impact adversarial robustness. In this work, we improve upon previous theoretical results [21], proving that label noise *guarantees* adversarial risk for large enough sample size. We provide a lower bound on the required sample size and show that, without further assumptions on the data distribution or the machine learning model, our result cannot be improved.

On the contrary, previous experimental results from Sanyal et al. [21] show that neural networks suffer from large adversarial risk even in the small sample size regime. Our results suggests that explaining such a phenomenon necessarily requires further assumptions on the data distributions, learning algorithm, or the machine learning model. While specific biases of machine learning models and algorithms (referred to as inductive bias) have usually played a "positive" role in machine learning literature [28, 27, 16, 1], we show how some biases may make the model more vulnerable to adversarial risks under noisy interpolation.

2. Main theoretical results

Our setting Choose a norm $\|\cdot\|$ on $\mathbb{R}^d$, for example $\|\cdot\|_2$ or $\|\cdot\|_\infty$. For $\boldsymbol{x} \in \mathbb{R}^d$, let $\mathcal{B}_r(\boldsymbol{x})$ denote the $\|\cdot\|$-ball of radius r around $\boldsymbol{x}$. Let μ be a distribution on $\mathbb{R}^d$ and let $f^* : \mathcal{C} \to \{0, 1\}$ be a measurable ground truth classifier. Then we can define the adversarial risk of any classifier f with respect to f^*, μ, given an adversary with perturbation budget $\rho > 0$ under the norm $\|\cdot\|$, as

[*]Equal contribution [1]ETH Zurich [2]ETH AI Center. Correspondence to: Amartya Sanyal <amartya.sanyal@ai.ethz.ch>, Daniel Paleka <daniel.paleka@inf.ethz.ch>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) = \mathbb{P}_{\boldsymbol{x}\sim\mu}\left[\exists \boldsymbol{z} \in B_\rho(\boldsymbol{x}),\ f^*(\boldsymbol{x}) \neq f(\boldsymbol{z}))\right]. \tag{1}$$

Next, consider a training set $((\boldsymbol{z}_1, y_1), \ldots, (\boldsymbol{z}_m, y_m))$ in $\mathbb{R}^d \times \{0,1\}$, where the $\boldsymbol{z}_i$ are independently sampled from μ, and each y_i equals $f^*(\boldsymbol{z}_i)$ with probability $1-\eta$, where $\eta > 0$ is the label noise rate. Let f be any classifier which correctly interpolates the training set. We can now state the main theoretical result of Sanyal et al. [21]:

Theorem 1 (Sanyal et al. [21])**.** *Suppose that there exist $c_1 \geq c_2 > 0$, $\rho > 0$, and a finite set $\zeta \subset \mathbb{R}^d$ satisfying*

$$\mu\left(\bigcup_{\boldsymbol{s}\in\zeta} \mathcal{B}_{\rho/2}(\boldsymbol{s})\right) \geq c_1 \quad \textit{and} \quad \forall \boldsymbol{s} \in \zeta,\ \mu\left(\mathcal{B}_{\rho/2}(\boldsymbol{s})\right) \geq \frac{c_2}{|\zeta|} \tag{2}$$

Further, suppose that each of these balls contains points from a single class. Then for $\delta > 0$, when the number of samples $m \geq \frac{|\zeta|}{\eta c_2}\log\left(\frac{|\zeta|}{\delta}\right)$, with probability $1-\delta$

$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) \geq c_1. \tag{3}$$

This is the first guarantee for adversarial risk caused by label noise in the literature. However, Theorem 1 has two extremely strong assumptions:

- The input distribution has mass c_1 in a union of balls, each of which has probability mass at least c_2;
- Each ball only contains points from a single class.

The assumptions are unrealistic: it is not clear why such balls would exist for real-world datasets, or even MNIST or CIFAR-10. In Theorem 2, we remove these assumptions and show that our guarantees hold for all compactly supported input distributions, with comparable guarantees on adversarial risk.

Denote a compact subset of $\mathbb{R}^d$ by $\mathcal{C}$. An important quantity in our theorem will be the *covering number* $N = N(\rho/2; \mathcal{C}, \|\cdot\|)$ of $\mathcal{C}$ in the metric $\|\cdot\|$. The covering number N is the minimum number of $\|\cdot\|$-balls of radius $\rho/2$ such that their union contains $\mathcal{C}$.

Theorem 2. *Let $\mathcal{C} \subset \mathbb{R}^d$ satisfy $\mu(\mathcal{C}) > 0$, and let $N = N(\rho/2; \mathcal{C}, \|\cdot\|)$ be its covering number. For $\delta > 0$, when the number of samples satisfies $m \geq \frac{8N}{\mu(\mathcal{C})\eta}\log\frac{2N}{\delta}$. with probability $1-\delta$ we have that*

$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) \geq \frac{1}{4}\mu(\mathcal{C}). \tag{4}$$

Note that the compact $\mathcal{C}$ can be chosen freely to make tradeoffs between the required number of samples m and the lower bound on the adversarial risk. As the covering number of the chosen $\mathcal{C}$ increases, the lower bound on the adversarial risk increases, but we also increase the required number of samples for the theorem to kick in. The tradeoff curve depends on the distribution μ; we discuss this in Section 3.

For compactly supported μ, we can take $\mathcal{C}$ to be the support of μ to prove a general statement.

Corollary 3. *Let N be the covering number of* $\mathrm{supp}(\mu)$ *with balls of radius $\rho/2$. For $\delta > 0$, when the number of samples satisfies $m \geq \frac{8N}{\eta}\log\frac{2N}{\delta}$. with probability $1-\delta$ we have that*

$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) \geq \frac{1}{4}. \tag{5}$$

This is easier to interpret than Theorem 2: if interpolating a dataset with label noise, the number of samples required to guarantee constant adversarial risk scales with the covering number of the support of the distribution.

Our Theorem 2 avoids the unwieldy assumptions, and in fact gives a slightly stronger guarantee than Theorem 1. When Equation (2) holds, our theorem requires the number of samples $m = \widetilde{\Omega}\left(\frac{|\zeta|}{\eta c_1}\right)$ instead of $m = \widetilde{\Omega}\left(\frac{|\zeta|}{\eta c_2}\right)$ in Theorem 1. We leave the proof of Theorem 2 to Appendix A, but we provide a brief sketch of the ideas behind it.

Proof sketch We want to prove that a large portion of points from μ have a label noised point nearby when m is large enough. With the label noise probability $\eta > 0$, the expected number of label noise training points is ηm; however a priori those could be anywhere in the support of μ.

The key idea is that we can always find a set of $\|\cdot\|$-balls covering a lot of measure, with each of the balls having a large enough density of μ. We prove this in Lemma 6.Then, if we take a lot of $\|\cdot\|$-balls with large density of a single class, we can prove that label noise induces an opposite-labeled point in each of the chosen balls given m large enough.

Concretely, the probability for a single chosen ball to not be adversarially vulnerable is on the order of $\left(1-\frac{\eta}{2N}\right)^{\mu(\mathcal{C})m}$, and summing this up over the chosen balls goes to zero in the assumed regime. Each of these balls is then adversarially vulnerable, summing up to a constant adversarial risk.

3. Practical implications

In his section, we discuss the limitations of results such as Theorem 2 in practical settings. When we allow arbitrary classifiers, we show that Theorem 2 paints an accurate picture of the interaction of label noise, interpolation and adversarial risk. However, we also show that this par-

ticular theoretical framework does not offer much hope in explaining the strong effect of label noise previously shown in Figure 3, as discussed in Section 4. We argue that this requires a better understanding of the inductive biases of the hypothesis class and the optimisation algorithm.

Large sample size is sometimes needed The number of required samples m in Theorem 2 can be very large, depending on the density and the covering number of the chosen compact $\mathcal{C}$. Consider $\|\cdot\|$ to be the maximum-norm $\|\cdot\|_\infty$, as is customary in adversarial robustness research [10]. Then the balls B_ρ are small hypercubes in $\mathbb{R}^d$. If we choose $\mathcal{C}$ to be the hypercube $[0,1]^d$, the covering number scales exponentially:

$$N = N(\rho; [0,1]^d, \|\cdot\|_\infty) \simeq \left(\frac{1}{\rho}\right)^d. \tag{6}$$

This can scale badly even for standard datasets such as MNIST ($d = 784$) or CIFAR-10 ($d = 3072$), since in Theorem 2 we need $m \gtrsim \frac{N}{\mu(\mathcal{C})\eta}$. This amounts an impossibly large sample size ($m \gtrsim 10^{784}$) to explain the effect present in $m = 50000$ MNIST training samples in Figure 3.

Hence our result often does not guarantee any adversarial risk if the number of samples m is small. In general, the covering number of a dataset is not polynomial in the dimension, except if the data has special properties in the given metric. For example, if the data distribution is supported on a subspace of $\mathbb{R}^d$ of dimension $k < d$, we can pick a $\mathcal{C}$ for which the covering number in $\|\cdot\|_2$ will depend only on k and not on d.

The large required sample size is not just a limitation of Theorem 2; in fact, we can show that if arbitrary classifiers are allowed, it is not always possible to lower bound the adversarial risk for $m = \mathrm{poly}(d)$.

Our result is tight It is a priori possible that the true dependence of adversarial risk on label noise kicks in for much lower sample size regimes than in Theorem 2. This might suggest that the lower bound on sample complexity can be improved. We can show this is not the case and in fact our bound is sharp. In particular, we exhibit a simple distribution on $\mathbb{R}^d$ such that there exist classifiers which correctly and robustly interpolate datasets of $m = \mathrm{poly}(d)$ samples from the distribution[1].

Proposition 4. *Let μ be the uniform distribution on $\mathbb{S}^{d-1} = \{x_1, \ldots, x_d \in \mathbb{R}^d : x_1^2 + \ldots + x_d^2 = 1\}$, and let the ground truth classifier f^* be a threshold function on x_1: $f^*(x) = \mathbb{1}_{x_1 > \frac{1}{2}}$. Consider any adversarial radius $\rho < \frac{1}{4}$ in the Euclidean metric. Then, for any label noise $\eta < 1$: with high probability, there exists a classifier f that interpolates $m = \lfloor 1.01^d \rfloor$ samples from the label noise distribution, such that $\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) = o_d(1)$.*

Proof sketch The main ingredient of the proof is the concentration of measure on $\mathbb{S}^{d-1}$, which makes the training samples far apart in the Euclidean or $\|\cdot\|_\infty$ metrics.We leave the full proof to Appendix B. Similar statements in the clean data setting have appeared before, most recently in [4].

Note that Proposition 4 shows that Theorem 2 cannot give an adversarial risk lower bound with sample size polynomial in m. Hence the covering number of any substantial portion of $\mathbb{S}^{d-1}$ is super-polynomial in the dimension d. This unintentionally proves that the covering number of the sphere $\mathbb{S}^{d-1}$ in the Euclidean metric is exponential, which is well-known.[2]

Optimizing $\mathcal{C}$ can avoid large sample size While Proposition 4 shows that our result in Theorem 2 is sharp in the worst case, it is possible a smaller sample size requirement under certain conditions. In particular, if we can pick a compact $\mathcal{C}$ with small covering number, such that the measure $\mu(\mathcal{C})$ is not too small, then Theorem 2 allows for a small sample size while guaranteeing a large adversarial risk.

Example Take an adversarial radius $\rho > 0$ in the $\|\cdot\|_\infty$ metric, Let $\mu = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2$ be the average of two measures, μ_1 and μ_2, with μ_1 the uniform distribution on $[0,1]^d$, and μ_2 the uniform distribution on a smaller hypercube $[0,\rho]^d$.

The first choice $\mathcal{C} = [0,1]^d$ as in Corollary 3 has covering number on the order of ρ^{-d}. Theorem 2 is then vacuous until $m \gtrsim \rho^{-d}/\eta$, which is very large in high dimensions. However, if we instead use $\mathcal{C} = [0,\rho]^d$, the covering number is 1 and we can use Theorem 2 for $m = O(\frac{1}{\eta})$.

Formally, to get the "best possible" m in Theorem 2 for a certain adversarial risk lower bound r, we should solve the following optimization problem over subsets of $\mathrm{supp}(\mu)$:

$$\min_{\mu(\mathcal{C}) \geq 4r} \frac{N(\rho/2, \mathcal{C}, \|\cdot\|_\infty) \log N(\rho/2, \mathcal{C}, \|\cdot\|_\infty)}{\mu(\mathcal{C})}. \tag{7}$$

It is not known whether this problem is tractable in general. However, the concept of having to solve an optimisation problem in order to get a tight lower bound is common in the literature. Some examples are the *representation dimension* [3] in differential privacy and the *SQ dimension* [9] in learning theory.

[1] We believe we can improve this to be exponential in m here.

[2] See Proposition 4.16 in https://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/20/material/lecture04.pdf

4. The role of inductive bias

We have seen, in the previous section, that without further assumptions the theoretical guarantees in Theorem 2 only hold for very large training sets. Proposition 4 shows that the result of Theorem 2 is sharp in that these results cannot be improved. In this section, we discuss how the inductive bias of the hypothesis class or the learning algorithm can lower the sample size requirement. This is practically relevant as Figure 3 shows that state of the art neural networks, in common vision datasets, show a much stronger dependance between label noise and adversarial robustness than what Theorem 2 prescribes.

Inductive bias can hurt robustness even further There is already ample empirical evidence [18, 13, 23] in existing works that neural networks exhibit an inductive bias that is different from what is required for robustness. For example, this is evident from the experiments shown in Figure 8 and Figure 9(a) in Sanyal et al. [21]. Shah et al. [22] also provides empirical evidence that neural networks exhibit a certain inductive bias, that they call simplicity bias, that hurts adversarial robustness.

Here, we show a simple example to illustrate the role of inductive bias. Consider a binary classification problem on a data distribution μ and an m-sized dataset $S_{m,\eta}$ sampled i.i.d. from μ such that the label of each example is flipped with probability η. We use $\mathcal{H}, \mathcal{F}$ to denote two hypothesis classes.

Theorem 5. *For any $\rho > 0$, there exists a distribution μ and two hypothesis classes $\mathcal{H}$ and $\mathcal{F}$, such that for any label noise rate $\eta \in (0, 1/2)$ and dataset size $m = \Theta\left(\frac{1}{\eta}\right)$, we have that: for all $h \in \mathcal{H}$ that interpolate $S_{m,\eta}$,*

$$\mathcal{R}_{\mathrm{Adv},\rho}\left(h; \mu\right) \geq \Omega\left(1\right); \tag{8}$$

whereas there exists an $f \in \mathcal{F}$ that interpolates $S_{m,\eta}$ and

$$\mathcal{R}_{\mathrm{Adv},\rho}\left(f; \mu\right) = \mathcal{O}\left(\rho\right). \tag{9}$$

We state the detailed proof in Appendix C but provide a short proof sketch here. The data distribution μ, of our construction, is defined on the domain $\mathbb{R}^2$ and distributed uniformly on the set $[0, W] \times \{0\}$ *i.e.,* the data is just supported on the first coordinate where $W \gg \rho$. The ground truth classifier is a threshold function on the first coordinate. The hypothesis $f \in \mathcal{F}$ simply labels everything according to the ground truth classifier except the mislabelled data points; where it constructs infinitesimally small intervals around the point on the first coordinate. Note that this construction is similar to the one in Proposition 4. By construction, it interpolates the training set and its expected adversarial risk is upper bounded by $2m\eta\rho$.

Each hypothesis in $\mathcal{H}$ can be thought of as a union of T-shaped decision regions as illustrated in Figure 1. The region inside the T-shaped regions are classified as 1 and the rest 0. Note that the head of the Ts make the region on the data manifold (first coordinate) directly below it adversarially vulnerable. Thus, for a significantly large width of the head, the total measure of the adversarially vulnerable set is large for any interpolating classifier. The width of the T can be interpreted as the inductive bias of the learning algorithm. The decision boundaries of neural networks usually lie on the manifold of the data [25]; and the network behaves more smoothly off the data manifold. A natural consequence of this is that the head of the Ts will be large. We don't propose this to be the exact inductive bias but rather an illustrative example for what might be happening in practice.

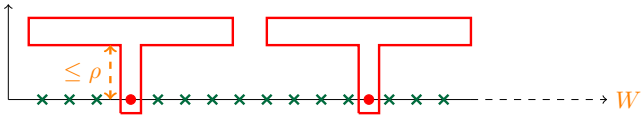


Figure 1: Visualization of a portion of the distribution μ and the hypothesis class $\mathcal{H}$ used in Theorem 5. The crosses are the mislabelled examples and the circles are correctly labelled examples. All the circles are adversarially vulnerable if perturbed upwards with magnitude less than ρ.

There are two important properties of this relatively simple example that make them relevant for understanding adversarial vulnerability of neural networks. First, the adversarial examples constructed here are off-manifold *i.e.,* they do not lie on the manifold of the data. This has been observed in prior works [12, 20]. Secondly, implicitly our examples also exhibits the *dimpled manifold* phenomenon described in Shamir et al. [23].

Is Theorem 2 about the wrong function class? When fitting deep neural networks to real datasets, the results of Theorem 2 still hold even when the number of samples m is smaller than required. We think that proving guarantees on adversarial risk in the presence of label noise is within reach for simple neural networks. Towards this goal, we propose a conjecture in a similar vein to Bubeck et al. [5]:

Conjecture 1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a neural network with a single layer with k neurons. Under the same conditions as in Theorem 2, for $m = \widetilde{\Omega}(\frac{1}{\eta}\mathrm{poly}(k, d))$,

$$\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) \geq \text{const.} \tag{10}$$

for a distribution μ supported on $[0, 1]^d$.

In short, we conjecture that neural networks exhibit properties (inductive biases) which hurt robustness when interpolating label noise. Understanding these properties is important for deploying neural networks in real world environments, where uniform label noise is not a possibility but rather a norm.

5. Acknowledgements

Amartya Sanyal acknowledges the ETH AI Center for the postdoctoral fellowship. We thank Mislav Balunović for an useful discussion regarding Theorem 2 and Florian Tramèr for general feedback.

References

[1] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning (ICML)*, 2018.

[2] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[3] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of pure private learners. *Journal of Machine Learning Research*, 2019.

[4] Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Sébastien Bubeck, Yuanzhi Li, and Dheeraj Nagaraj. A law of robustness for two-layers neural networks. *arXiv:2009.14444*, 2020.

[6] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017.

[7] Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effects of inductive bias. In *International Conference on Machine Learning (ICML)*, 2022.

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[9] Vitaly Feldman. A general characterization of the statistical query complexity. In *Conference on Learning Theory*, 2017.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[11] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 2022.

[12] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[13] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[14] A Kurakin, I Goodfellow, and S Bengio. Adversarial examples in the physical world. *arXiv:1607.02533*, 2016.

[15] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 2000.

[16] Chris Mingard, Joar Skalse, Guillermo Valle-Pérez, David Martínez-Rubio, Vladimir Mikulik, and Ard A. Louis. Neural networks are a priori biased towards boolean functions with low entropy. In *International Conference on Learning Representations (ICLR)*, 2020.

[17] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv:2005.08054*, 2020.

[18] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, and Pascal Frossard. Neural anisotropy directions. *Advances in Neural Information Processing Systems*, 33:17896–17906, 2020.

[19] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE symposium on security and privacy (SP)*, 2016.

[20] Amartya Sanyal, Varun Kanade, Philip HS Torr, and Puneet K Dokania. Robustness via deep low-rank representations. *arXiv:1804.07090*, 2018.

[21] Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. How benign is benign overfitting? In *International Conference on Learning Representations (ICLR)*, 2021.

[22] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv:2106.10151*, 2021.

[24] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on computer and communications security*, 2016.

[25] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. *arXiv:2203.08124*, 2022.

[26] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] Bart van Merriënboer, Amartya Sanyal, Hugo Larochelle, and Yoshua Bengio. Multiscale sequence modeling with a learned dictionary. *arXiv:1707.00762*, 2017.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[29] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017.

A. Proof of Theorem 2

Here we prove the following statement:

Theorem 2. *Let $\mathcal{C} \subset \mathbb{R}^d$ satisfy $\mu(\mathcal{C}) > 0$, and let $N = N(\rho/2; \mathcal{C}, \|\cdot\|)$ be its covering number. For $\delta > 0$, when the number of samples satisfies $m \geq \frac{8N}{\mu(\mathcal{C})\eta} \log \frac{2N}{\delta}$. with probability $1 - \delta$ we have that*

$$\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) \geq \frac{1}{4}\mu(\mathcal{C}). \tag{4}$$

For notational convenience, we replace ρ by 2ρ in all places for the proof below.

Proof. Without loss of generality, let $\mathcal{C}_0 = \{\, \boldsymbol{x} \in \mathcal{C} : f^*(\boldsymbol{x}) = 0 \,\}$ have probability $\mu(\mathcal{C}_0) \geq \frac{1}{2}\mu(\mathcal{C})$. Let $\mu_0 = \mu|_{\mathcal{C}_0}$, normalized so that $\mu_0(\mathcal{C}_0) = 1$.

By Chernoff, with probability $1 - \exp\left(-\frac{\mu(\mathcal{C})m}{16}\right) \geq 1 - \frac{\delta}{2}$, at least $m_0 = \lfloor \frac{\mu(\mathcal{C})m}{4} \rfloor$ of the samples $\boldsymbol{z}_i$ are in $\mathcal{C}_0$. Without loss of generality, let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m_0}$ be those samples. Then

$$\begin{aligned}
\mathcal{R}_{\mathrm{Adv},2\rho}\left(f, \mu\right) &\geq \frac{1}{2}\mu(\mathcal{C})\, \mathbb{P}_{\boldsymbol{x}\sim\mu, \boldsymbol{x}\in\mathcal{C}_0}\left[\exists \boldsymbol{z} \in B_{2\rho}(\boldsymbol{x}),\ f^*(\boldsymbol{x}) \neq f(\boldsymbol{z}))\right] && (11)\\
&= \frac{1}{2}\mu(\mathcal{C})\, \mathbb{P}_{\boldsymbol{x}\sim\mu_0}\left[\exists \boldsymbol{z} \in B_{2\rho}(\boldsymbol{x}),\ f(\boldsymbol{z}) \neq 0\right] && (12)\\
&\geq \frac{1}{2}\mu(\mathcal{C})\, \mathbb{P}_{\boldsymbol{x}\sim\mu_0}\left[\exists\, i \leq m_0\ :\ \boldsymbol{z}_i \in B_{2\rho}(\boldsymbol{x}) \cap \mathcal{C}_0,\ f(\boldsymbol{z}_i) \neq 0\right] && (13)\\
&= \frac{1}{2}\mu(\mathcal{C})\, \mathbb{P}_{\boldsymbol{x}\sim\mu_0}\left[\exists\, i \leq m_0\ :\ \boldsymbol{x} \in B_{2\rho}(\boldsymbol{z_i}),\ \boldsymbol{z}_i \in \mathcal{C}_0,\ f(\boldsymbol{z}_i) \neq 0\right]. && (14)\\
&= \frac{1}{2}\mu(\mathcal{C})\, \mu_0\left(\bigcup_{i\leq m_0,\ f(\boldsymbol{z}_i)\neq 0} B_{2\rho}(\boldsymbol{z_i})\right). && (15)
\end{aligned}$$

Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ be a ρ-covering of $\mathcal{C}_0$, ordered such that

$$\mu_0(B_\rho(\boldsymbol{s}_1)) \geq \ldots \geq \mu_0(B_\rho(\boldsymbol{s}_N)) \tag{16}$$

The plan is the following: we will lower bound $\bigcup_{i\leq m_0,\ f(\boldsymbol{z}_i)\neq 0} B_{2\rho}(\boldsymbol{z}_i)$ by the union of some $B_\rho(\boldsymbol{s}_k)$, which will have large μ_0-measure in total. Moreover, each of the chosen $B_\rho(\boldsymbol{s}_k)$ will have large enough μ_0-measure. For this, we use the following lemma:

Lemma 6. *If $1 \leq K \leq N$ is the largest index such that $\mu_0(B_\rho(\boldsymbol{s}_K) \geq \frac{1}{2N}$, then*

$$\mu_0\left(\bigcup_{k=1}^{K} B_\rho(\boldsymbol{s}_i)\right) > \frac{1}{2}. \tag{17}$$

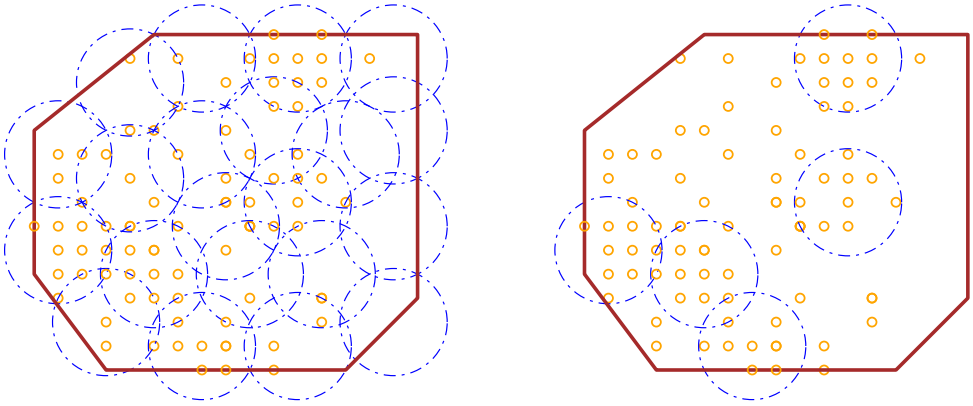
Proof. As $\bigcup_{k=1}^{N} B_\rho(\boldsymbol{s}_i)$ is a cover of $\mathcal{C}_0$,

$$\mu_0\left(\bigcup_{k=K+1}^{N} B_\rho(\boldsymbol{s}_i)\right) \geq 1 - \mu_0\left(\bigcup_{k=K+1}^{N} B_\rho(\boldsymbol{s}_i)\right) \geq 1 - \sum_{k=K+1}^{N} \mu_0(B_\rho(\boldsymbol{s}_i)) \geq 1 - \frac{N-K}{2N} > \frac{1}{2}. \tag{18}$$

■

We now show that the chosen balls are dense enough to get samples in the training set with high probability.

Lemma 7. *With probability $1 - \delta/2$, each $B_\rho(\boldsymbol{s}_k)$ for $k \leq K$ contains at least one $\boldsymbol{z}_i \in \mathcal{C}_0$ such that $f(\boldsymbol{z}_i) \neq 0$.*



a) The original cover of $\mathcal{C}$. b) The dense greedy subcover.

Figure 2: Illustration of Lemma 6. Given a cover of N balls, we can pick a subcover of balls covering at least half of the measure, with each ball having measure at least $\frac{1}{2N}$.

Proof. We have

$$\mathbb{P}\left[\boldsymbol{z}_i \in B_\rho(\boldsymbol{s}_k) \mid \boldsymbol{z}_i \in \mathcal{C}_0\right] \geq \frac{1}{2N}, \tag{19}$$

and because the label corruption is independent from everything, we also have

$$\mathbb{P}\left[f(\boldsymbol{z}_i) \neq 0 \mid \boldsymbol{z}_i \in \mathcal{C}_0\right] = \eta \tag{20}$$

$$\implies \mathbb{P}\left[f(\boldsymbol{z}_i) \neq 0 \wedge \boldsymbol{z}_i \in B_\rho(\boldsymbol{s}_k) \mid \boldsymbol{z}_i \in \mathcal{C}_0\right] \geq \frac{\eta}{2N} \tag{21}$$

Therefore,

$$\mathbb{P}\left[B_\rho(\boldsymbol{s}_k) \cap \{\, \boldsymbol{z}_i : i \leq m_0,\ f(\boldsymbol{z}_i) \neq 0 \,\} = \emptyset\right] \tag{22}$$

$$= \prod_{i=1}^{m_0} \mathbb{P}\left[z_i \notin B_\rho(\boldsymbol{s}_k) \vee z_i \notin \mathcal{C}_0 \vee f(\boldsymbol{z}_i) \neq 0\right] \tag{23}$$

$$\leq \left(1 - \frac{\eta}{2N}\right)^{m_0} \tag{24}$$

$$\leq \exp\left(-\frac{m_0 \eta}{2N}\right) = \frac{\delta}{2N}, \tag{25}$$

and hence

$$\mathbb{P}\left[\left(\bigcup_{k=1}^{K} B_\rho(\boldsymbol{s}_k)\right) \cap \{\, \boldsymbol{z}_i : i \leq m_0,\ f(\boldsymbol{z}_i) \neq 0 \,\} = \emptyset\right] \leq K \frac{\delta}{2N} \leq \frac{\delta}{2}. \tag{26}$$

■

Finally, using both Lemma 6 and Lemma 7, we can finish:

$$\mathcal{R}_{\mathrm{Adv},2\rho} = \frac{1}{2}\mu(\mathcal{C})\ \mu_0\left(\bigcup_{i \leq m_0,\ f(\boldsymbol{z}_i) \neq 0} B_{2\rho}(\boldsymbol{z}_i)\right). \tag{27}$$

$$\geq \frac{1}{2}\mu(\mathcal{C})\ \mu_0\left(\bigcup_{k=1}^{K} B_\rho(\boldsymbol{s}_k)\right) \geq \frac{1}{4}\mu(\mathcal{C}). \tag{28}$$

□

B. Proof of Proposition 4

Proposition 4. *Let μ be the uniform distribution on $\mathbb{S}^{d-1} = \{x_1, \dots, x_d \in \mathbb{R}^d : x_1^2 + \dots + x_d^2 = 1\}$, and let the ground truth classifier f^* be a threshold function on x_1: $f^*(x) = \mathbb{1}_{x_1 > \frac{1}{2}}$. Consider any adversarial radius $\rho < \frac{1}{4}$ in the Euclidean metric. Then, for any label noise $\eta < 1$: with high probability, there exists a classifier f that interpolates $m = \lfloor 1.01^d \rfloor$ samples from the label noise distribution, such that $\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) = o_d(1)$.*

Proof. Let the $m = 1.01^d \le \exp(d/80)$ samples be $\boldsymbol{z}_1, \dots, \boldsymbol{z}_m$ with labels $y_1, \dots, y_m \in \{0, 1\}$. Almost surely the $\boldsymbol{z}_i$ are distinct. Define the interpolating classifier $f : \mathbb{R}^d \to \{0, 1\}$ as

$$f(\boldsymbol{x}) = \begin{cases} y_i & \text{if } \boldsymbol{x} \in \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\}; \\ \mathbb{1}_{x_1 > \frac{1}{2}} & \text{otherwise.} \end{cases} \tag{29}$$

We want to show f is robust. Draw $\boldsymbol{x} = (x_1, \dots, x_d)$ uniformly on $\mathbb{S}^{d-1}$. There are only two ways $\boldsymbol{x}$ can contribute to the adversarial risk $\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu)$:

- $\boldsymbol{x}$ is close to a training sample $\boldsymbol{z_i}$ with label noise;
- $\boldsymbol{x}$ is close to the "decision boundary" $x_1 = \frac{1}{2}$ of $\mathbb{S}^{d-1}$.

Hence, remembering Equation (1),

$$\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) \le \mathbb{P}\left[\boldsymbol{x} \text{ is in a } \rho\text{-ball around at least one of the } \boldsymbol{z}_i\right] + \mathbb{P}\left[\frac{1}{2} - \rho \le x_1 \le \frac{1}{2} + \rho\right]. \tag{30}$$

$$\le \mathbb{P}\left[\boldsymbol{x} \text{ is in a } \rho\text{-ball around at least one of the } \boldsymbol{z}_i\right] + \mathbb{P}\left[x_1 \ge \frac{1}{2} - \rho\right]. \tag{31}$$

By the union bound,

$$\mathbb{P}\left[\boldsymbol{x} \text{ is in a } \rho\text{-ball around at least one of the } \boldsymbol{z}_i\right] \tag{32}$$

$$\le m\, \mathbb{P}\left[\|\boldsymbol{x} - \boldsymbol{z}_1\|_2 \le \rho\right] \tag{33}$$

$$\le m\, \mathbb{P}\left[\|\boldsymbol{x}\|^2 + \|\boldsymbol{z}_1\|^2 - 2\langle \boldsymbol{x}, \boldsymbol{z}_1\rangle \le \rho^2\right] \tag{34}$$

$$= m\, \mathbb{P}\left[\langle \boldsymbol{x}, \boldsymbol{z}_1\rangle \ge 1 - \rho^2/2\right]. \tag{35}$$

As μ is rotationally invariant, $\langle \boldsymbol{x}, \boldsymbol{z}_1\rangle$ is distributed the same as x_1. We have proved

$$\mathcal{R}_{\mathrm{Adv},\rho}(f, \mu) \le m\, \mathbb{P}\left[x_1 \ge 1 - \frac{\rho^2}{2}\right] + \mathbb{P}\left[x_1 \ge \frac{1}{2} - \rho\right]. \tag{36}$$

We can bound $\mathbb{P}[x_1 \ge t]$ for $t > 0$ as follows: let $g_1, \dots, g_d$ be i.i.d. standard $N(0, 1)$ random variables.

$$\mathbb{P}\left[x_1 \ge t\right] = \mathbb{P}\left[\frac{g_1}{\sqrt{g_1^2 + \dots + g_d^2}} \ge t\right] \tag{37a}$$

$$= \mathbb{P}\left[g_1^2 \ge t^2(g_1^2 + \dots + g_d^2)\right] \tag{37b}$$

$$= \mathbb{P}\left[\frac{1 - t^2}{t^2} g_1^2 \ge g_2^2 + \dots + g_d^2\right] \tag{37c}$$

$$\le \mathbb{P}\left[\frac{1 - t^2}{t^2} g_1^2 \ge \frac{d-1}{2}\right] + \mathbb{P}\left[g_2^2 + \dots + g_d^2 \le \frac{d-1}{2}\right], \tag{37d}$$

where the last inequality is because $a \le b$ implies $a \ge c$ or $b \le c$.

As $0 < \rho < \frac{1}{4}$, we can take $t = \frac{1}{4}$ in both probabilities in Equation (36). We now use the often-cited chi-square bounds from Lemma 1 in Laurent and Massart [15].

$$\mathbb{P}\left[g_2^2 + \ldots + g_d^2 \leq (d-1) - 2\sqrt{(d-1)s}\right] \leq \exp\left(-s\right) \tag{38a}$$

$$\mathbb{P}\left[g_1^2 \geq 1 + 2\sqrt{s} + 2s\right] \leq \exp\left(-s\right) \tag{38b}$$

Then for $s = \frac{d}{40}$, it's easy to see that both probabilities in Equation (37d) are less than the corresponding probabilities in Equation (38b) and Equation (38a).

Finally, as d goes to infinity,

$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) \leq m\exp(-d/40) + \exp(-d/40) \tag{39}$$

$$\leq \exp(-d/80) + \exp(-d/40) \to 0. \tag{40}$$

□

C. Proof of inductive bias

Theorem 8. *For any $\rho > 0$, there exists a distribution μ and two hypotheses classes $\mathcal{H}$ and $\mathcal{F}$, such that for any label noise rate $\eta \in (0, 1/2)$ and dataset size $m = \Theta\left(\frac{1}{\eta}\right)$, we have that: for all $h \in \mathcal{H}$ that interpolate $S_{m,\eta}$,*

$$\mathcal{R}_{\mathrm{Adv},\rho}\left(h;\mu\right) \geq \Omega\left(1\right); \tag{41}$$

whereas there exists an $f \in \mathcal{F}$ that interpolates $S_{m,\eta}$ and

$$\mathcal{R}_{\mathrm{Adv},\rho}\left(f;\mu\right) = \mathcal{O}\left(\rho\right). \tag{42}$$

Proof. For any $\rho \geq 0,\ W \gg \rho$, construct a distribution μ on $[0, W] \times \{0\}$ as follows. Distribute the covariates uniformly randomly in $[0,\ \frac{W}{2} - 2\rho] \bigcup [\frac{W}{2} + 2\rho,\ W]$ and then label then with the ground truth labelling function $f^*\left(x\right) = \mathbb{1}\{x_1 \geq \frac{W}{2}\}$ where $x = [x_1, x_2]$ is the two-dimensional covariate. Next, we construct an m dimensional dataset and flip each label independently with probability $1 - \eta$. We denote this set with $S_{m,\eta}$.

The hypothesis class $\mathcal{F}$ is the class of one-dimensional thresholds on the first coordinate of the input space (ignores the second coordinate entirely). Define the following interpolating classifier $f \in \mathcal{F} : \mathbb{R}^2 \to \{0, 1\}$ as follows

$$f(x) = \begin{cases} y_1 & \text{if } x \text{ is in } S_{m,\eta} \\ \mathbb{1}\{x_1 \geq W/2\} & \text{otherwise} \end{cases}.$$

As the sampling of the covariates and the label noise are independent events,

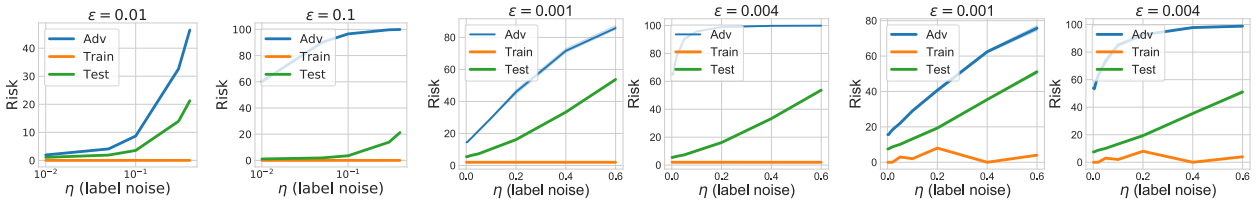
$$\mathbb{E}_{S_{m,\eta}}\left[\#\text{ of mislabelled points in } S_{m,\eta}\right] = m\eta.$$

Then the expected measure of the set of points adversarially vulnerable by an adversary of perturbation magnitude ρ on the classifier h, as defined above, is upper bounded by $2\rho m\eta$. Using the fact that the total measure of the domain is W and that $m = \Theta\left(\frac{1}{\eta}\right)$, we get that

$$\mathbb{E}_{S_{m,\eta}}\left[\mathcal{R}_{\mathrm{Adv},\rho}\left(f;\mu\right)\right] \leq \frac{2\rho m\eta}{W} = \mathcal{O}(\rho).$$

Next, consider the hypothesis class $\mathcal{H}$ defined as follows. Given a set of points $\mathcal{Z} = \{z_1, \ldots, z_k\} \in [0, W]^k$ and $\gamma > \rho$, define the hypothesis

$$h_{\mathcal{Z},\gamma}\left(x\right) = \begin{cases} 1 & \exists z \in \mathcal{Z} \mid \mathbb{1}\{x_2 < \rho\} \wedge \mathbb{1}\{x_1 = z\} \\ 1 & \exists z \in \mathcal{Z} \mid \mathbb{1}\{x_2 < \rho\} \wedge \mathbb{1}\{|x_1 - z| \leq \gamma\} \\ 0 & \text{otherwise.} \end{cases}$$



(a) MNIST (b) ResNet18 (CIFAR10) (c) DenseNet121 (CIFAR10)

Figure 3: From Sanyal et al. [21]. Adversarial error increases with increasing label noise η (x-axis) at a rate much faster than predicted by Theorem 2. This is likely due to the inductive bias of the neural network. Here, ϵ is the perturbation magnitude (ρ in the current paper). The label noise is synthetically injected in the training set with probability η.

If $\widetilde{S}$ is the set of mislabelled 1s in $S_{m,\eta}$, then for any interpolating classifier $h_{\mathcal{Z},\gamma}$, it holds that $\widetilde{S} \subseteq \mathcal{Z}$. Next, by construction, for every point $z \in \mathcal{Z}$, it holds that all points $x \in [z-\gamma, z+\gamma]$ can adversarially perturbed in the x_2 component to obtain the label 1. Thus the total measure of the adversarially vulnerable set of points is greater than the number of mislabelled points, whose original label is zero, multipled with 2γ, which is $2m\eta\gamma$.

Thus, we have that for any $h \in \mathcal{H}$ that interpolates $S_{m,\eta}$,

$$\mathbb{E}_{S_{m,\eta}}\left[\mathcal{R}_{\mathrm{Adv},\rho}\left(f;\mu\right)\right] \geq \min\left(\frac{2\gamma m\eta}{W},\ 1/2\right) = \Omega(\gamma).$$

□

Finally, note that both of the bounds in Theorem 8 can be transformed into high probability bounds using concentration inequalities. Also note that for simplicity, we do not treat the above as learning problems, but it is possible to show that there exists a learning algorithms that uses a similar number of samples as above to output $f \in \mathcal{F}$ such that the adversarial risk is $\mathcal{O}\left(\rho\right)$.

D. Existing experimental results

In Figure 3, we show results from Sanyal et al. [21]. Here, the adversarial risk is plotted against label noise in the dataset for various neural network models and datasets.