BRIDGING THE GAP BETWEEN ADVERSARIAL ROBUSTNESS AND OPTIMIZATION BIAS

Fartash Faghri[1,2*] **Cristina Vasconcelos**[3] **David J. Fleet**[1,2,3] **Fabian Pedregosa**[3]

Nicolas Le Roux[3,4]

[1]University of Toronto [2]Vector Institute [3]Google Research [4]Mila

ABSTRACT

Adversarial robustness is an open challenge in deep learning, most often tackled using adversarial training. Adversarial training is computationally costly, involving alternated optimization with a trade-off between standard generalization and adversarial robustness. We explore training robust models without adversarial training by revisiting a known result linking maximally robust classifiers and minimum norm solutions, and combining it with recent results on the implicit bias of optimizers. First, we show that, under certain conditions, it is possible to achieve both perfect standard accuracy and a certain degree of robustness without a trade-off, simply by training an overparametrized model using the implicit bias of the optimization. In that regime, there is a direct relationship between the type of the optimizer and the attack to which the model is robust. Second, we investigate the role of the architecture in designing robust models. In particular, we characterize the robustness of linear convolutional models, showing that they resist attacks subject to a constraint on the Fourier-ℓ_∞ norm. This result explains the property of ℓ_p-bounded adversarial perturbations that tend to be concentrated in Fourier space. This leads us to a novel attack in the Fourier domain that is inspired by the well-known frequency-dependent sensitivity of human perception. We evaluate Fourier-ℓ_∞ robustness of recent CIFAR-10 models with robust training and visualize adversarial perturbations.

1 PRELIMINARIES

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ denote a training dataset sampled I.I.D. from the distribution $\mathcal{D}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ are features and $y_i \in \{-1, +1\}$ are binary labels. A binary classifier is a function $\varphi : \mathbb{R}^d \to \mathbb{R}$, and its prediction on an input $\boldsymbol{x}$ is given by $\operatorname{sign}(\varphi(\boldsymbol{x})) \in \{-1, +1\}$. The aim in supervised learning is to find a classifier that accurately classifies the training data and generalizes to unseen test data. One standard framework for training a classifier is Empirical Risk Minimization (ERM), $\arg\min_{\varphi\in\Phi} \mathcal{L}(\varphi)$, $\mathcal{L}(\varphi) \coloneqq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\, \zeta(y\varphi(\boldsymbol{x}))$ where Φ is a family of classifiers and $\zeta : \mathbb{R} \to \mathbb{R}^+$ is a loss function that we assume to be strictly monotonically decreasing to 0, i.e., $\zeta' < 0$. Examples are the exponential loss, $\exp(-\hat{y}y)$, and the logistic loss, $\log(1 + \exp(-\hat{y}y))$.

Given a classifier, an adversarial perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$ is any small perturbation that changes the model prediction, i.e., $\operatorname{sign}(\varphi(\boldsymbol{x} + \boldsymbol{\delta})) \neq \operatorname{sign}(\varphi(\boldsymbol{x}))$, $\|\boldsymbol{\delta}\| \leq \varepsilon$, where $\|\cdot\|$ is a norm on $\mathbb{R}^d$, and ε is an arbitrarily chosen constant. In practice, an adversarial perturbation, $\boldsymbol{\delta}$, is found as an approximate solution to the optimization problem $\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq\varepsilon} \zeta(y\varphi(\boldsymbol{x} + \boldsymbol{\delta}))$. Madry et al. (2017) defined an adversarially robust classifier as the solution to the saddle-point optimization problem,

$$\arg\min_{\varphi\in\Phi} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq\varepsilon} \zeta(y\varphi(\boldsymbol{x} + \boldsymbol{\delta}))\,. \tag{1}$$

Adversarial Training (Goodfellow et al., 2014) refers to solving (1) using an alternated optimization that is computationally expensive because it often requires solving the inner maximization many times.

*Work done during an internship at Google Research. Code available at: `https://github.com/fartashf/robust_bias`. Correspondence to: Fartash Faghri <faghri@cs.toronto.edu>.

The main drawback of defining the adversarially robust classifier using (1), and a drawback of adversarial training, is that the parameter ε needs to be known or tuned. The choice of ε controls a trade-off between standard accuracy on samples of the distribution $\mathcal{D}$ versus the robust accuracy, i.e., the accuracy on adversarial samples.

2 Maximally Robust Classifier

In order to avoid the trade-off imposed by ε in adversarial robustness and reduce computational load of finding a desirable ε, we revisit a definition from robust optimization. Appendix B discusses additional details of this section. Proofs are provided in Appendix E.

Definition 2.1. A **Maximally Robust Classifier** (Ben-Tal et al., 2009) is a solution to

$$\arg\max_{\varphi\in\Phi}\{\varepsilon \,|\, y_i\varphi(\boldsymbol{x}_i+\boldsymbol{\delta})>0,\ \|\boldsymbol{\delta}\|\leq\varepsilon,\ \forall i\}\,. \tag{2}$$

Compared with the saddle-point formulation (1), ε in (2) is not an arbitrary constant. Rather, it is maximized as part of the optimization problem. Moreover, the maximal ε in this definition does not depend on a particular loss function. Note, a maximally robust classifier is not necessarily unique.

The downside of this formulation is that it requires the training data to be separable so that (2) is non-empty, i.e. there exists $\varphi\in\Phi$ such that $\forall i, y_i\varphi(\boldsymbol{x}_i)>0$. In most deep learning settings, this is not necessarily a concern as models are large enough that they can interpolate the training data, i.e. for any dataset there exists φ such that $\varphi(\boldsymbol{x}_i)=y_i$.

Given a dataset and a norm for the perturbation, what is the maximally robust linear classifier? Here, we revisit a result from Ben-Tal et al. (2009) for classification. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The associated *dual norm*, denoted $\|\cdot\|_*$, is defined as $\|\boldsymbol{\delta}\|_*=\sup_{\boldsymbol{x}}\{|\langle\boldsymbol{\delta},\boldsymbol{x}\rangle|\ |\ \|\boldsymbol{x}\|\leq 1\}$.

Lemma 2.1 Maximally Robust Linear Classifier (Ben-Tal et al. (2009), §12). *For linear models and linearly separable data, the following problems are equivalent; i.e., from a solution of one, a solution of the other is readily found.*

$$\textit{Maximally robust classifier:}\quad \arg\max_{\boldsymbol{w},b}\{\varepsilon \,|\, y_i(\boldsymbol{w}^\top(\boldsymbol{x}_i+\boldsymbol{\delta})+b)>0,\ \|\boldsymbol{\delta}\|\leq\varepsilon,\ \forall i\} \tag{3}$$

$$\textit{Maximum margin classifier:}\quad \arg\max_{\boldsymbol{w},b:\|\boldsymbol{w}\|_*\leq 1}\ \{\varepsilon \,|\, y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\geq\varepsilon,\ \ \forall i\} \tag{4}$$

$$\textit{Minimum norm classifier:}\quad \arg\min_{\boldsymbol{w},b}\{\|\boldsymbol{w}\|_* \,|\, y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\geq 1,\ \ \forall i\}\,. \tag{5}$$

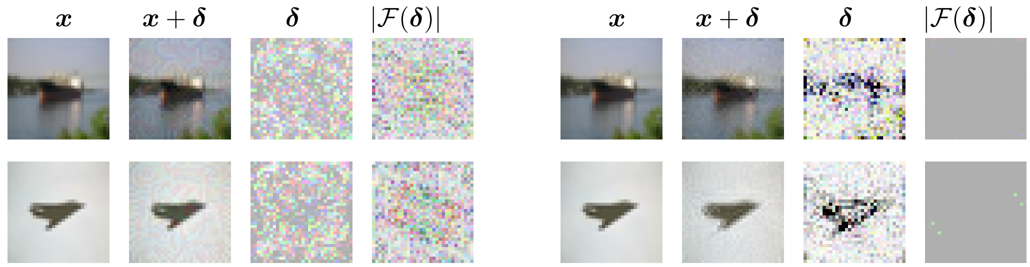
We provide a proof for general norms based on Ben-Tal et al. (2009) in Appendix E.1.

3 Implicit Robustness of Optimizers

In this section, we use recent results on the implicit bias of optimizers combined with Lemma 2.1 to characterize the robustness of overparametrized models. The most common approach to empirical risk minimization is through a gradient-based optimizer. Gunasekar et al. (2018a) showed that gradient descent, and more generally steepest descent methods, have an implicit bias towards minimum norm solutions. Even though there exist infinitely many solutions that minimize the empirical risk, we can characterize the one found by steepest descent. Using Lemma 2.1, we show that such a classifier is also maximally robust w.r.t. a specific norm. Appendix C discusses additional details of this section. Proofs are provided in Appendix E.

Corollary 1 Implicit Robustness of Steepest Descent. *For any linearly separable dataset and any norm $\|\cdot\|$, steepest descent iterates minimizing the empirical risk, $\mathcal{L}(\boldsymbol{w})$, satisfying the conditions of Theorem C.1, converge in direction to a maximally robust classifier,* $\arg\max_{\boldsymbol{w}}\{\varepsilon \,|\, y_i\boldsymbol{w}^\top(\boldsymbol{x}_i+\boldsymbol{\delta})>0,\ \|\boldsymbol{\delta}\|_*\leq\varepsilon,\ \ \forall i\}$. *In particular, a maximally robust classifier against ℓ_1, ℓ_2, and ℓ_∞ is reached, respectively, by sign gradient descent, gradient descent, and coordinate descent.*

Corollary 1 implies that overparametrized linear models are maximally robust classifiers, without the additional cost of adversarial training. One could also obtain an alternative robustness guarantee with the appropriate choice of optimization method.



(a) ℓ_∞ attack (b) Fourier-ℓ_∞ attack

Figure 1: **Adversarial attacks (ℓ_∞ and Fourier-ℓ_∞) against CIFAR-10 classification models.** Fourier-ℓ_∞ perturbations in the spatial domain are concentrated around subtle details of the object. In contrast, ℓ_∞ perturbations are perceived by people as noise. The magnitude of Fourier-ℓ_∞ perturbations at almost all frequencies is equal to ε except for a few frequencies that appear as small light colored dots in the mostly grey perturbation (darker means stronger perturbation). We use $\varepsilon = 8/255$. Amplitude spectra are displayed with the origin at the center of the image.

Next, we show that even for linear models, the choice of the architecture affects implicit robustness, which gives another alternative for achieving maximal robustness. We use a generalization of Theorem C.1 to linear convolutional models. To extend Corollary 1 to deep learning models one can use generalizations of Theorem C.1. For the special case of gradient descent, Theorem C.1 has been generalized to multi-layer fully-connected linear networks and a larger family of strictly monotonically decreasing loss functions including the logistic loss (Nacson et al., 2019, Theorem 2).

Corollary 2 Maximally Robust to Perturbations with Bounded Fourier Coefficients. *Consider the family of two-layer linear convolutional networks and the gradient descent iterates, $\boldsymbol{w}_t$, minimizing the empirical risk. For almost all linearly separable datasets under conditions of Theorem C.2, $\boldsymbol{w}_t$ converges in direction to a maximally robust classifier,* $\arg\max_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_L}\{\varepsilon \,|\, y_i\varphi_{conv}(\boldsymbol{x}_i;\{\boldsymbol{w}_l\}_{l=1}^L) > 0,\ \|\mathcal{F}(\boldsymbol{\delta})\|_\infty \le \varepsilon,\ \forall i\}$*, where $\mathcal{F}$ denotes the Discrete Fourier Transform (DFT).*

Corollary 2 implies that, at no additional cost, linear convolutional models are already maximally robust, but w.r.t. perturbations in the Fourier domain. As our visualizations in Fig. 1 show, adversarial perturbations under bounded Fourier-ℓ_∞ are more concentrated on subtle details of the main object. Appendix I depicts the various norm-balls in 3D to illustrate the significant geometrical difference between the Fourier-ℓ_∞ and other commonly used norm-balls for adversarial robustness. The robustness to Fourier-ℓ_∞ can justify standard generalization of convolutional models.

Model	Norm	Standard	Fourier-ℓ_∞
WideResNet-28-10	-	$\mathbf{94.78}$	23.40
Wu et al. (2020)	ℓ_∞	88.25	82.33
Carmon et al. (2019)	ℓ_∞	89.69	83.86
Wu et al. (2020)	ℓ_2	88.51	84.17
Augustin et al. (2020)	ℓ_2	91.08	$\mathbf{86.84}$

Figure 2: **Standard and Fourier-ℓ_∞ accuracy for CIFAR-10 models.** Fourier-ℓ_∞ uses $\varepsilon = 8/255$ which is beyond the maximal ε for any of the models. Beyond the maximal ε, there is no guarantee for robustness. Yet, standard training has non-zero accuracy against this attack which shows a degree of robustness. Models with robust training reach an accuracy higher than standard training which suggests the implicit bias of non-linear convolutional networks is not exactly Fourier-ℓ_∞. (The *Norm* column specifies the norm used for robust training.)

These results further imply that to fool a linear convolutional network, any adversarial perturbation must have at least one frequency beyond the maximal robustness of the model. This condition is satisfied for perturbations with small ℓ_1 norm in the spatial domain, i.e., only a few pixels are perturbed. This has been observed empirically, where the amplitude spectra of adversarial samples with bounded ℓ_p norms are largely band limited (Yin et al., 2019).

3.1 CIFAR-10 Experiments

Fig. 2 reports standard and robust accuracy of image classification models on CIFAR-10 (Krizhevsky et al., 2009), comparing models with robust training to standard training. We implement our attack in AutoAttack (Croce & Hein, 2020) and evaluate the robustness of recent defenses available in

RobustBench (Croce et al., 2020). The attack methods are APGD-CE and APGD-DLR with default hyperparameters in RobustBench (Croce et al., 2020) and $\varepsilon = 8/255$. Algorithm 1 provides pseudocode for our Fourier-ℓ_∞ attack. In Appendix F we derive the Fourier-ℓ_∞ attack in closed form for linear models. Theoretical results do not provide any guarantee beyond the maximally robust ε and on test data. The ℓ_2 robustly trained model of Augustin et al. (2020) achieves highest accuracy against Fourier-ℓ_∞ attack, suggesting that the implicit bias of standard training for deep non-linear models does not match the implicit bias of linear models.

Appendix H.3 provide more examples like those in Fig. 1 for robustly trained models, showing that perturbations are qualitatively different from those against the standard model. Details of the experimental setup and more figures and visualizations appear in Appendix H.

Algorithm 1 Fourier-ℓ_∞ Attack (see Appendix F)

Input: data $\boldsymbol{x}$, label y, loss function ζ, classifier φ, perturbation size ε, number of attack steps m, dimensions d, Fourier transform $\mathcal{F}$

for $k = 1$ **to** m **do**

$\quad \hat{\boldsymbol{g}} = \mathcal{F}(\nabla_{\boldsymbol{x}} \zeta(y\varphi(\boldsymbol{x})))$

$\quad [\boldsymbol{\delta}]_i = \frac{[\hat{\boldsymbol{g}}]_i}{|[\hat{\boldsymbol{g}}]_i|}, \forall i \in \{0, \dots, d-1\}$

$\quad \boldsymbol{x} = \boldsymbol{x} + \varepsilon \mathcal{F}^{-1}(\boldsymbol{\delta})$

end for

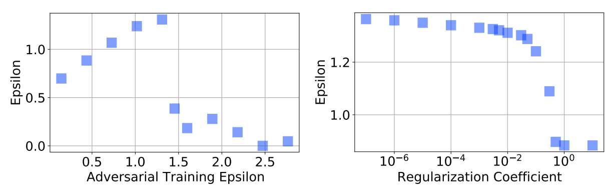
4 Explicit Regularization

Above we discussed the impact of optimization method and model architecture on robustness. Here, we discuss explicit regularization as another choice that determines maximal robustness. Appendix D discusses additional details of this section. Proofs are provided in Appendix E.

Corollary 3 Maximally Robust Classifier via Infinitesimal Regularization. *For linearly separable data, under conditions of Theorem D.1, solutions to the regularized empirical risk minimization problem,* $\hat{\boldsymbol{w}}(\lambda) = \arg\min_{\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \zeta(y\boldsymbol{w}^\top \boldsymbol{x}) + \lambda\|\boldsymbol{w}\|$, *converges in direction to a maximally robust classifier as* $\lambda \to 0$. *That is,* $\lim_{\lambda\to 0} \hat{\boldsymbol{w}}(\lambda)/\|\hat{\boldsymbol{w}}(\lambda)\|$ *converges to a solution of* $\arg\max_{\boldsymbol{w}} \{\varepsilon \,|\, y_i \boldsymbol{w}^\top(\boldsymbol{x}_i + \boldsymbol{\delta}) > 0,\ \|\boldsymbol{\delta}\|_* \leq \varepsilon,\ \forall i\}$.

Assuming the solution to the regularized problem is unique, the regularization term replaces other implicit biases in minimizing the empirical risk. The regularization coefficient controls the tradeoff between robustness and standard accuracy. The advantage of this formulation compared with adversarial training is that we do not need the knowledge of the maximally robust ε to find a maximally robust classifier. It suffices to choose an infinitesimal regularization coefficient. Wei et al. (2019, Theorem 4.1) generalized Theorem D.1 for a family of classifiers that includes fully-connected networks with ReLU non-linearities, which allows for potential extension of Corollary 3 to non-linear models.

Fig. 3 illustrates the tradeoff between standard accuracy and adversarial robustness for adversarial training and explicit regularization methods. Adversarial training finds the maximally robust classifier only if it is trained with the knowledge of the maximally robust ε. Without this knowledge, we have to search for the maximal ε by training multiple models. This adds further computational complexity to adversarial training which performs an alternated optimization. In contrast, explicit regularization converges to a maximally robust classifier for a small enough regularization constant.



(a) Adversarial training (b) ℓ_1 regularization

Figure 3: **Trade-off in robustness against ℓ_∞ attack.** We plot the maximally robust ε for adversarial training and explicit regularization. Robustness is controlled using ε in adversarial training (left) and regularization coefficient in explicit regularization (right). Using adversarial training we have to search for the maximal ε but for explicit regularization it suffices to choose a small regularization coefficient.

5 Conclusion

We bridge the two literature on adversarial robustness and optimization bias by revisiting a known result linking maximally robust classifiers and minimum norm solutions, and combining it with recent results on the implicit bias of optimizers. We discussed implicit optimization, architectural bias and explicit regularization as approaches to finding maximally robust classifiers. We designed an optimal Fourier-ℓ_∞ attack and evaluated the robustness of CIFAR-10 models. Our framework facilitates future study of the implicit robustness of novel architectures and optimizers to more general attacks.

ACKNOWLEDGMENTS

The authors would like to thank Nicholas Carlini, Nicolas Papernot, and Courtney Paquette for helpful discussions and invaluable feedback. NLR is supported by the Canada CIFAR AI Chair program.

REFERENCES

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2020.

Muhammad Awais, Fahad Shamshad, and Sung-Ho Bae. Towards an adversarially robust normalization approach. *CoRR*, abs/2006.11007, 2020.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, pp. 11190–11201, 2019.

Josue Ortega Caro, Yilong Ju, Ryan Pyle, and Ankit Patel. Using learning dynamics to explore the role of implicit regularization in adversarial examples. *arXiv preprint arXiv:2006.11440*, 2020.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A Model of Double Descent for High-dimensional Binary Linear Classification. *arXiv e-prints*, art. arXiv:1911.05822, November 2019.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv e-prints*, art. arXiv:2006.05161, June 2020.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in neural information processing systems*, pp. 1178–1187, 2018a.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018b.

Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W. Taylor. Batch normalization is a cause of adversarial vulnerability. *CoRR*, abs/1905.02161, 2019.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial Spheres. *arXiv e-prints*, art. arXiv:1801.02774, January 2018.

Rafael C. González and Richard E. Woods. *Digital image processing, 3rd Edition*. Pearson Education, 2008.

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

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. *arXiv e-prints*, art. arXiv:1810.12715, October 2018.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples. *arXiv e-prints*, art. arXiv:2010.03593, October 2020.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018b.

Yiwen Guo, Long Chen, Yurong Chen, and Changshui Zhang. On connections between regularizations for improving dnn robustness. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints*, art. arXiv:1903.08560, March 2019.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in neural information processing systems*, pp. 2266–2276, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Adel Javanmard and Mahdi Soltanolkotabi. Precise Statistical Analysis of Classification Accuracies for Adversarial Training. *arXiv e-prints*, art. arXiv:2010.11213, October 2020.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2034–2078. PMLR, 2020.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *NeurIPS*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *ICLR (Poster)*. OpenReview.net, 2019.

Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. A unified gradient regularization family for adversarial examples. In *2015 IEEE International Conference on Data Mining*, pp. 301–309. IEEE, 2015.

Avery Ma, Fartash Faghri, and Amir-massoud Farahmand. Adversarial Robustness through Regularization: A Second-Order Approach. *arXiv e-prints*, art. arXiv:2004.01832, April 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pp. 6640–6650. PMLR, 2020.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019.

Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Hold me tight! influence of discriminative features on deep network boundaries. *arXiv preprint arXiv:2002.06349*, 2020.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 2013.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pp. 13847–13856, 2019.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Saharon Rosset, Ji Zhu, and Trevor J Hastie. Margin maximizing loss functions. In *Advances in neural information processing systems*, pp. 1237–1244, 2004.

Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. In *Advances in Neural Information Processing Systems*, pp. 9835–9846, 2019.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Ayon Sen, Xiaojin Zhu, Liam Marshall, and Robert Nowak. Should Adversarial Attacks Use Pixel p-Norm? *arXiv e-prints*, art. arXiv:1906.02439, June 2019.

Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. In *CVPR Workshops*, pp. 1605–1613. IEEE Computer Society, 2018.

Yash Sharma, Gavin Weiguang Ding, and Marcus Brubaker. On the effectiveness of low frequency perturbations. *arXiv preprint arXiv:1903.00073*, 2019.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5):1926–1940, 1998.

Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In *International Conference on Machine Learning*, pp. 5809–5817. PMLR, 2019.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel RD Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Matus Telgarsky. Margins, shrinkage, and boosting. In *ICML (2)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 307–315. JMLR.org, 2013.

Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *NeurIPS*, pp. 5858–5868, 2019.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On Adaptive Attacks to Adversarial Example Defenses. *arXiv e-prints*, art. arXiv:2002.08347, February 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Yusuke Tsuzuku and Issei Sato. On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 51–60, 2019.

Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Nicolas Le Roux, and Ross Goroshin. An effective anti-aliasing approach for residual networks. *CoRR*, abs/2011.10675, 2020.

Brian Wandell and Stephen Thomas. Foundations of vision. *Psyccritiques*, 42(7), 1997.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9712–9724, 2019.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2018.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020.

Cihang Xie, Mingxing Tan, Boqing Gong, Alan L. Yuille, and Quoc V. Le. Smooth adversarial training. *CoRR*, abs/2006.14536, 2020.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A Fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32:13276–13286, 2019.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A Unifying View on Implicit Bias in Training Linear Neural Networks. *arXiv e-prints*, art. arXiv:2010.02501, October 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*. OpenReview.net, 2017.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. *arXiv e-prints*, art. arXiv:1901.08573, January 2019.

Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Adversarial Color Enhancement: Generating Unrestricted Adversarial Images by Optimizing a Color Filter. *arXiv e-prints*, art. arXiv:2002.01008, February 2020.

A INTRODUCTION

Deep neural networks achieve high accuracy on standard test sets, yet Szegedy et al. (2013) showed that any natural input correctly classified by a neural network can be modified with adversarial perturbations that fool the network into misclassification; even when such perturbations are constrained to be small enough that do not significantly affect human perception. Adversarially perturbed inputs are referred to as Adversarial samples. Adversarial training (Goodfellow et al., 2014) was proposed to improve the robustness of models through training on adversarial samples and further generalized as a saddle-point optimization problem (Madry et al., 2017). In practice, adversarial training refers to methods that solve the saddle-point problem approximately by minimizing the loss on adversarial samples.

Adversarial training is the state-of-the-art approach to adversarial robustness (Gowal et al., 2020; Croce et al., 2020), as alternative approaches are more probable to exhibit spurious robustness (Tramer et al., 2020). Nevertheless, adversarial training is also computationally more expensive than standard training as it involves an alternated optimization. Adversarial training also exhibits a trade-off between standard generalization and adversarial robustness. That is, it achieves improved *robust accuracy*, on adversarially perturbed data, at the expense of *standard accuracy*, the probability of correct predictions on natural data (Tsipras et al., 2018). The adversarial robustness trade-off has been shown to be intrinsic in a number of toy examples (Fawzi et al., 2018b), even independent of the learning algorithm in some examples (Schmidt et al., 2018). Alternatives to adversarial training have been proposed to reduce this trade-off but a gap still remains in practice (Zhang et al., 2019).

Here we consider connections between the adversarial robustness trade-off and *standard generalization* of overparametrized models. Deep learning models can often achieve interpolation; i.e., they have the capacity to exactly fit the training data (Zhang et al., 2016). Their ability to generalize well in such cases has been attributed to an implicit bias toward simple solutions (Gunasekar et al., 2018a; Hastie et al., 2019).

Our first contribution is to establish a link between the two bodies of work on adversarial robustness and standard generalization. Focusing on models that achieve interpolation, we use the formulation of a *Maximally Robust Classifier* from robust optimization (Ben-Tal et al., 2009). We analyze models that achieve perfect training accuracy together with maximal robustness to a given set of perturbations. We consider training procedures that yield maximally robust classifiers through an implicit bias in the optimization method, or through explicit regularization. We observe that, in contrast to adversarial training, under certain conditions we can find maximally robust classifiers at no additional cost.

The link between adversarial robustness and the implicit bias of optimization methods also yields a natural way to characterize the robustness of linear convolutional models to Fourier-ℓ_∞ perturbations. This result provides a justification for recent observations about Fourier properties of adversarial perturbations (Tsuzuku & Sato, 2019; Sharma et al., 2019; Yin et al., 2019). As a result, we design a novel attack with bounded Fourier-ℓ_∞ norm and illustrate adversarial attacks against CIFAR-10 models.

B ADDITIONAL DETAILS OF MAXIMALLY ROBUST CLASSIFICATION

One can also show that adversarial training, i.e., solving the saddle-point problem (1), does not necessarily find a maximally robust classifier. To see this, suppose we are given the maximal ε in (2). Further assume the minimum of (1) is non-zero. Then the cost in the saddle-point problem does not distinguish between the following two models: 1) a model that makes no misclassification but has low confidence, i.e. $\forall i,\ 0 < \max_{\boldsymbol{\delta}} y_i \varphi(\boldsymbol{x}_i + \boldsymbol{\delta}) \leq c_1$ for some small c_1 2) a model that classifies a training point, $\boldsymbol{x}_j$, incorrectly but is highly confident on all other training data and adversarially perturbed ones, i.e. $\forall i \neq j,\ 0 < c_2 < \max_{\boldsymbol{\delta}} y_i \varphi(\boldsymbol{x}_i + \boldsymbol{\delta})$. The second model can incur a loss $n\zeta(c_1) - (n-1)\zeta(c_2)$ on $\boldsymbol{x}_j$ while being no worse than the first model according to the cost of the saddle-point problem. The reason is another trade-off between standard and robust accuracy caused by taking the expectation over data points.

Definition B.1 Dual norm. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The associated *dual norm*, denoted $\|\cdot\|_*$, is defined as

$$\|\boldsymbol{\delta}\|_* = \sup_{\boldsymbol{x}} \{ |\langle \boldsymbol{\delta}, \boldsymbol{x} \rangle| \mid \|\boldsymbol{x}\| \leq 1 \} \ . \tag{6}$$

Definition B.2 Linear Separability. We say a dataset is linearly separable if there exists $\boldsymbol{w}, b$ such that $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0$ for all i.

For a dataset to be linearly separable, it is sufficient to have $d \geq n$ and no i, j exist such that $\boldsymbol{x}_i = \boldsymbol{x}_j$.

The expression $\min_i y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)/\|\boldsymbol{w}\|$ is the margin of a classifier $\boldsymbol{w}$ that is the distance of the nearest training point to the classification boundary, i.e. the line $\{\boldsymbol{v} : \boldsymbol{w}^\top \boldsymbol{v} = -b\}$.

Each formulation in Lemma 2.1 is connected to a wide array of results that can be transferred to other formulations. Maximally robust classification is one example of a problem in robust optimization. Other problems such as robust regression as well as robustness to correlated input perturbations have been studied prior to deep learning (Ben-Tal et al., 2009). In the case of maximally robust classification, the problem can be reduced and solved efficiently.

On the other hand, maximum margin and minimum norm classification have long been popular because of their generalization guarantees. Recent theories for overparametrized models link the margin and the norm of a model to generalization (Hastie et al., 2019). Although the tools are different, connecting the margin and the norm of a model has also been the basis of generalization theories for Support Vector Machines and AdaBoost (Shawe-Taylor et al., 1998; Telgarsky, 2013). Maximum margin classification does not require linear separability, because there can exist a classifier with $\varepsilon < 0$ that satisfies the margin constraints. Minimum norm classification is the easiest formulation to work with in practice as it does not rely on ε nor $\boldsymbol{\delta}$ and minimizes a function of the weights subject to a set of constraints.

C Additional details of Implicit Robustness of Optimizers

Recall that empirical risk minimization is defined as $\arg\min_{\varphi \in \Phi} \mathcal{L}(\varphi)$, where $\mathcal{L}(\varphi) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \zeta(y\varphi(\boldsymbol{x}))$. Here we assume $\mathcal{D}$ is a finite dataset of size n. For the linear family of functions, we write $\mathcal{L}(\boldsymbol{w}, b)$. Hereafter, we rewrite the loss as $\mathcal{L}(\boldsymbol{w})$ and augment the input with a constant 1 dimension. For linearly separable data and overparametrized models ($d > n$), there exist infinitely many linear classifiers that minimize the empirical risk (Gunasekar et al., 2018a). We will find it convenient to ignore the scaling and focus on the normalized vector $\boldsymbol{w}/\|\boldsymbol{w}\|$, namely the direction of $\boldsymbol{w}$. We will say that the sequence $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots$ converges in direction to a vector $\boldsymbol{v}$ if $\lim_{t\to\infty} \boldsymbol{w}_t/\|\boldsymbol{w}_t\| = \boldsymbol{v}$.

Definition C.1 Steepest Descent. Let $\|\cdot\|$ denote a norm, f a function to be minimized, and γ a step size. The steepest descent method associated with this norm finds

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \gamma \Delta \boldsymbol{w}_t, \quad \text{where}$$

$$\Delta \boldsymbol{w}_t \in \arg\min_{\boldsymbol{v}} \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{v} \rangle + \frac{1}{2}\|\boldsymbol{v}\|^2 \,. \tag{7}$$

The steepest descent step, $\Delta \boldsymbol{w}_t$, can be equivalently written as $-\|\nabla f(\boldsymbol{w}_t)\|_* \, g_{\text{nst}}$, where

$$g_{\text{nst}} \in \arg\min \left\{ \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{v} \rangle \mid \|\boldsymbol{v}\| = 1 \right\} \,. \tag{8}$$

A proof can be found in Boyd & Vandenberghe (2004, §9.4).

Remark. For some p-norms, steepest descent steps have closed form expressions. Gradient Descent (GD) is steepest descent w.r.t. ℓ_2 norm where $-\nabla f(\boldsymbol{w}_t)$ is a steepest descent step. Sign gradient descent is steepest descent w.r.t. ℓ_∞ norm where $-\|\nabla f(\boldsymbol{w}_t)\|_1 \operatorname{sign}(\nabla f(\boldsymbol{w}_t))$ is a steepest descent step. Coordinate Descent (CD) is steepest descent w.r.t. ℓ_1 norm where $-\nabla f(\boldsymbol{w}_t)_i \boldsymbol{e}_i$ is a steepest descent step (i is the coordinate for which the gradient has the largest absolute magnitude).

Theorem C.1 Implicit Bias of Steepest Descent (Gunasekar et al. (2018a) (Theorem 5)). *For any separable dataset $\{\boldsymbol{x}_i, y_i\}$ and any norm $\|\cdot\|$, consider the steepest descent updates from (7) for minimizing the empirical risk $\mathcal{L}(\boldsymbol{w})$ with the exponential loss, $\zeta(z) = \exp(-z)$. For all initializations $\boldsymbol{w}_0$, and all bounded step-sizes satisfying a known upper bound, the iterates $\boldsymbol{w}_t$ satisfy*

$$\lim_{t\to\infty} \min_i \frac{y_i \boldsymbol{w}_t^\top \boldsymbol{x}_i}{\|\boldsymbol{w}_t\|} = \max_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq 1} \min_i y_i \boldsymbol{w}^\top \boldsymbol{x}_i \,. \tag{9}$$

In particular, if a unique maximum margin classifier $\boldsymbol{w}^*_{\|\cdot\|} = \arg\max_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq 1} \min_i y_i \boldsymbol{w}^\top \boldsymbol{x}_i$ exists, the limit direction converges to it, i.e. $\lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|} = \boldsymbol{w}^*_{\|\cdot\|}$.

In other words, the margin converges to the maximum margin and if the maximum margin classifier is unique, the iterates converge in direction to $\boldsymbol{w}^*_{\|\cdot\|}$.

We use this result to derive our Corollary 1.

We also note that Theorem C.1 and Corollary 1, characterize linear models, but do not account for the bias b. We can close the gap with an augmented input representation, to include the bias explicitly. Or one could preprocess the data, removing the mean before training.

Definition C.2 Linear convolutional network. An L-layer convolutional network with 1-D circular convolution is parameterized using weights of $L-1$ convolution layers, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{L-1} \in \mathbb{R}^d$, and weights of a final linear layer, $\boldsymbol{w}_L \in \mathbb{R}^d$, such that the linear mapping of the network is

$$\varphi_{\text{conv}}(\boldsymbol{x}; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_L) \coloneqq \boldsymbol{w}_L^\top (\boldsymbol{w}_{L-1} \star \cdots (\boldsymbol{w}_1 \star \boldsymbol{x}))\,.$$

Here, circular convolution is defined as $[\boldsymbol{w} \star \boldsymbol{x}]_i \coloneqq \frac{1}{\sqrt{d}} \sum_{k=0}^{d-1} [\boldsymbol{w}]_{\overline{-k}} [\boldsymbol{x}]_{\overline{i+k}}$, where $[\boldsymbol{v}]_i$ denotes the i-th element of a vector $\boldsymbol{v}$ for $i = 0, \ldots, d-1$, and $\overline{i} = i \bmod d$. [1]

A linear convolutional network is equivalent to a linear model with weights $\boldsymbol{w} = \boldsymbol{w}_L \star (\cdots \star (\boldsymbol{w}_2 \star \boldsymbol{w}_1))$ because of the associative property of convolution. In particular, for two-layer linear convolutional networks $\boldsymbol{w} = \boldsymbol{w}_2 \star \boldsymbol{w}_1$.

Definition C.3 Discrete Fourier Transform. $\mathcal{F}(\boldsymbol{w}) \in \mathbb{C}^d$ denotes the Fourier coefficients of $\boldsymbol{w}$ where $[\mathcal{F}(\boldsymbol{w})]_d = \frac{1}{\sqrt{d}} \sum_{k=0}^{d-1} [\boldsymbol{w}]_k \exp(-\frac{2\pi j}{d} kd)$ and $j^2 = -1$.

Theorem C.2 Implicit Bias towards Fourier Sparsity (Gunasekar et al. (2018b), Theorem 2, 2.a). *Consider the family of L-layer linear convolutional networks and the sequence of gradient descent iterates, $\boldsymbol{w}_t$, minimizing the empirical risk, $\mathcal{L}(\boldsymbol{w})$, with the exponential loss, $\exp(-z)$. For almost all linearly separable datasets under known conditions on the step size and convergence of iterates, $\boldsymbol{w}_t$ converges in direction to the classifier minimizing the norm of the Fourier coefficients given by*

$$\operatorname*{arg\,min}_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_L} \{\|\mathcal{F}(\boldsymbol{w})\|_{2/L} \mid y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 1,\, \forall i\}. \tag{10}$$

In particular, for two-layer linear convolutional networks the implicit bias is towards the solution with minimum ℓ_1 norm of the Fourier coefficients, $\|\mathcal{F}(\boldsymbol{w})\|_1$. For $L > 2$, the convergence is to a first-order stationary point.

We use this result to derive our Corollary 2.

D Additional details of Explicit Regularization

Definition D.1 Regularized Classification. The regularized empirical risk minimization problem for linear classification is defined as

$$\hat{\boldsymbol{w}}(\lambda) = \operatorname*{arg\,min}_{\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \zeta(y \boldsymbol{w}^\top \boldsymbol{x}) + \lambda \|\boldsymbol{w}\|\,, \tag{11}$$

where λ denotes a regularization constant, ζ is a monotone loss function, and $\mathcal{D}$ is a dataset. For simplicity we assume this problem has a unique solution.

Theorem D.1 Maximum Margin Classifier using Regularization (Rosset et al., 2004), Theorem 2.1. *Consider linearly separable finite datasets and monotonically non-increasing loss functions. Then as $\lambda \to 0$, the sequence of solutions, $\hat{\boldsymbol{w}}(\lambda)$, to the regularized problem in Definition D.1, converges in direction to a maximum margin classifier as defined in (4). Moreover, if the maximum margin classifier is unique,*

$$\lim_{\lambda \to 0} \frac{\hat{\boldsymbol{w}}(\lambda)}{\|\hat{\boldsymbol{w}}(\lambda)\|} = \operatorname*{arg\,max}_{\boldsymbol{w}: \|\boldsymbol{w}\| \leq 1} \min_i y_i \boldsymbol{w}^\top \boldsymbol{x}_i\,. \tag{12}$$

[1] We use the usual definition of circular convolution in signal processing, rather than cross-correlation, $\boldsymbol{w}^\downarrow \star \boldsymbol{x}$ with $[\boldsymbol{v}^\downarrow]_i = [\boldsymbol{v}]_{\overline{-i}}$, which is used in deep learning literature, but not associative.

The original proof was given specifically for ℓ_p norms, however we observe that their proof only requires convexity of the norm, so we state it more generally. Quasi-norms such as ℓ_p for $p < 1$ are not covered by this theorem. In addition, the condition on the loss function is weaker than our strict monotonic decreasing condition as shown in (Nacson et al., 2019, Appendix A).

We use this result to derive our Corollary 3.

Explicit regularization has been explored as an alternative approach to adversarial training (Hein & Andriushchenko, 2017; Sokolić et al., 2017; Zhang et al., 2019; Qin et al., 2019; Guo et al., 2020). To be clear, we do not propose a new regularization method but rather provide a framework for deriving and guaranteeing the robustness of existing regularization methods.

E PROOFS

E.1 PROOF OF LEMMA 2.1

Lemma 2.1 Maximally Robust Linear Classifier (Ben-Tal et al. (2009), §12). *For linear models and linearly separable data, the following problems are equivalent; i.e., from a solution of one, a solution of the other is readily found.*

$$\text{\textit{Maximally robust classifier:}}\quad \operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \,|\, y_i(\boldsymbol{w}^\top(\boldsymbol{x}_i+\boldsymbol{\delta})+b)>0,\ \|\boldsymbol{\delta}\|\le\varepsilon,\ \forall i\} \tag{3}$$

$$\text{\textit{Maximum margin classifier:}}\quad \operatorname*{arg\,max}_{\boldsymbol{w},b:\|\boldsymbol{w}\|_*\le 1}\{\varepsilon \,|\, y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\ge\varepsilon,\ \forall i\} \tag{4}$$

$$\text{\textit{Minimum norm classifier:}}\quad \operatorname*{arg\,min}_{\boldsymbol{w},b}\{\|\boldsymbol{w}\|_* \,|\, y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\ge 1,\ \forall i\}\,. \tag{5}$$

Proof. We first show that the maximally robust classifier is equivalent to a robust counterpart by removing $\boldsymbol{\delta}$ from the problem,

$$\begin{aligned}
&\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid y_i(\boldsymbol{w}^\top(\boldsymbol{x}_i+\boldsymbol{\delta})+b)>0,\ \|\boldsymbol{\delta}\|\le\varepsilon,\ \forall i\}\\
&\qquad\text{(homogeneity of p-norm)}\\
&=\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid y_i(\boldsymbol{w}^\top(\boldsymbol{x}_i+\varepsilon\boldsymbol{\delta})+b)>0,\ \|\boldsymbol{\delta}\|\le 1,\ \forall i\}\\
&\qquad\text{(if it is true for all $\boldsymbol{\delta}$ it is true for the worst of them)}\\
&=\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid \inf_{\|\boldsymbol{\delta}\|\le 1} y_i(\boldsymbol{w}^\top(\boldsymbol{x}_i+\varepsilon\boldsymbol{\delta})+b)>0,\ \forall i\}\\
&=\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)+\varepsilon\inf_{\|\boldsymbol{\delta}\|\le 1}\boldsymbol{w}^\top\boldsymbol{\delta}\ge 0,\ \forall i\}\\
&\qquad\text{(definition of dual norm)}\\
&=\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)>\varepsilon\|\boldsymbol{w}\|_*,\ \forall i\}
\end{aligned}$$

Assuming $\boldsymbol{w}\neq 0$, which is a result of linear separability assumption, we can divide both sides by $\|\boldsymbol{w}\|_*$ and change variables,

$$=\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\ge\varepsilon,\ \forall i, \|\boldsymbol{w}\|_*\le 1\}\,,$$

where we are also allowed to change $>$ to $\ge$ because any solution to one problem gives an equivalent solution to the other given $\boldsymbol{w}\neq 0$.

Now we show that the robust counterpart is equivalent to the minimum norm classification problem by removing ε. When the data is linearly separable there exists a solution with $\varepsilon > 0$,

$$\begin{aligned}
&\operatorname*{arg\,max}_{\boldsymbol{w},b}\{\varepsilon \mid y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)>\varepsilon\|\boldsymbol{w}\|_*,\ \forall i\}\\
&=\operatorname*{arg\,max}_{\boldsymbol{w},b}\left\{\varepsilon \mid y_i\left(\frac{\boldsymbol{w}^\top}{\varepsilon\|\boldsymbol{w}\|_*}\boldsymbol{x}_i+\frac{b}{\varepsilon\|\boldsymbol{w}\|_*}\right)\ge 1,\ \forall i\right\}
\end{aligned}$$

This problem is invariant to any non-zero scaling of $(\boldsymbol{w}, b)$, so with no loss of generality we set $\|\boldsymbol{w}\|_* = 1$.

$$= \arg\max_{\boldsymbol{w},b} \left\{ \varepsilon \mid y_i \left(\frac{\boldsymbol{w}^\top}{\varepsilon} \boldsymbol{x}_i + b \right) \geq 1, \|\boldsymbol{w}\|_* = 1, \forall i \right\}$$

Let $\boldsymbol{w}' = \boldsymbol{w}/\epsilon$, then the solution to the following problem gives a solution for $\boldsymbol{w}$,

$$\arg\max_{\boldsymbol{w}',b} \left\{ \frac{1}{\|\boldsymbol{w}'\|_*} \mid y_i(\boldsymbol{w}'^\top \boldsymbol{x}_i + b) \geq 1, \forall i \right\}$$
$$= \arg\min_{\boldsymbol{w}',b} \left\{ \|\boldsymbol{w}'\|_* \mid y_i(\boldsymbol{w}'^\top \boldsymbol{x}_i + b) \geq 1, \forall i \right\}.$$

□

E.2 Proof of Maximally Robust with Steepest Descent (Corollary 1)

Corollary 1 Implicit Robustness of Steepest Descent. *For any linearly separable dataset and any norm $\|\cdot\|$, steepest descent iterates minimizing the empirical risk, $\mathcal{L}(\boldsymbol{w})$, satisfying the conditions of Theorem C.1, converge in direction to a maximally robust classifier,* $\arg\max_{\boldsymbol{w}} \{\varepsilon \mid y_i \boldsymbol{w}^\top (\boldsymbol{x}_i + \boldsymbol{\delta}) > 0, \|\boldsymbol{\delta}\|_* \leq \varepsilon, \ \forall i\}$. *In particular, a maximally robust classifier against ℓ_1, ℓ_2, and ℓ_∞ is reached, respectively, by sign gradient descent, gradient descent, and coordinate descent.*

Proof. By Theorem C.1, the margin of the steepest descent iterates, $\min_i \frac{y_i \boldsymbol{w}_t^\top \boldsymbol{x}_i}{\|\boldsymbol{w}_t\|}$, converges as $t \to \infty$ to the maximum margin, $\max_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq 1} \min_i y_i \boldsymbol{w}^\top \boldsymbol{x}_i$. By Lemma 2.1, any maximum margin classifier w.r.t. $\|\cdot\|$ gives a maximally robust classifier w.r.t. $\|\cdot\|_*$. □

E.3 Proof of Maximally Robust to Perturbations Bounded in Fourier Domain (Corollary 2)

Corollary 2 Maximally Robust to Perturbations with Bounded Fourier Coefficients. *Consider the family of two-layer linear convolutional networks and the gradient descent iterates, $\boldsymbol{w}_t$, minimizing the empirical risk. For almost all linearly separable datasets under conditions of Theorem C.2, $\boldsymbol{w}_t$ converges in direction to a maximally robust classifier,* $\arg\max_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_L} \{\varepsilon \mid y_i \varphi_{conv}(\boldsymbol{x}_i; \{\boldsymbol{w}_l\}_{l=1}^L) > 0, \|\mathcal{F}(\boldsymbol{\delta})\|_\infty \leq \varepsilon, \ \forall i\}$, *where $\mathcal{F}$ denotes the Discrete Fourier Transform (DFT).*

The proof mostly follows from the equivalence for linear models in Appendix E.1 by substituting the dual norm of Fourier-ℓ_1. Here, $\boldsymbol{A}^*$ denotes the complex conjugate transpose, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}^*$ is the complex inner product, $[\boldsymbol{F}]_{ik} = \frac{1}{\sqrt{D}} \omega_D^{ik}$ the DFT matrix where $\omega_D = e^{-j2\pi/D}$, $j = \sqrt{-1}$.

Let $\|\cdot\|$ be a norm on $\mathbb{C}^n$ and $\langle \cdot, \cdot \rangle$ be the complex inner product. Similar to $\mathbb{R}^n$, the associated dual norm is defined as $\|\boldsymbol{\delta}\|_* = \sup_{\boldsymbol{x}} \{|\langle \boldsymbol{\delta}, \boldsymbol{x} \rangle| \mid \|\boldsymbol{x}\| \leq 1\}$.

$$\|\mathcal{F}(\boldsymbol{w})\|_1$$
$$= \sup_{\|\boldsymbol{\delta}\|_\infty \leq 1} |\langle \mathcal{F}(\boldsymbol{w}), \boldsymbol{\delta} \rangle|$$

(Expressing DFT as a linear transformation.)

$$= \sup_{\|\boldsymbol{\delta}\|_\infty \leq 1} |\langle \boldsymbol{F}\boldsymbol{w}, \boldsymbol{\delta} \rangle|$$
$$= \sup_{\|\boldsymbol{\delta}\|_\infty \leq 1} |\langle \boldsymbol{w}, \boldsymbol{F}^* \boldsymbol{\delta} \rangle|$$

(Change of variables and $\boldsymbol{F}^{-1} = \boldsymbol{F}^*$.)

$$= \sup_{\|\boldsymbol{F}\boldsymbol{\delta}\|_\infty \leq 1} |\langle \boldsymbol{w}, \boldsymbol{\delta} \rangle|$$
$$= \sup_{\|\mathcal{F}(\boldsymbol{\delta})\|_\infty \leq 1} |\langle \boldsymbol{w}, \boldsymbol{\delta} \rangle| .$$

E.4 Proof of Maximally Robust with Explicit Regularization (Corollary 3)

Corollary 3 Maximally Robust Classifier via Infinitesimal Regularization. *For linearly separable data, under conditions of Theorem D.1, solutions to the regularized empirical risk minimization problem, $\hat{\boldsymbol{w}}(\lambda) = \arg\min_{\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\zeta(y\boldsymbol{w}^\top\boldsymbol{x}) + \lambda\|\boldsymbol{w}\|$, converges in direction to a maximally robust classifier as $\lambda \to 0$. That is, $\lim_{\lambda\to 0} \hat{\boldsymbol{w}}(\lambda)/\|\hat{\boldsymbol{w}}(\lambda)\|$ converges to a solution of* $\arg\max_{\boldsymbol{w}}\{\varepsilon \,|\, y_i\boldsymbol{w}^\top(\boldsymbol{x}_i + \boldsymbol{\delta}) > 0,\ \|\boldsymbol{\delta}\|_* \le \varepsilon,\ \forall i\}$.

Proof. By Theorem D.1, the margin of the sequence of regularized classifiers, $\min_i y_i \frac{\hat{\boldsymbol{w}}(\lambda)^\top}{\|\hat{\boldsymbol{w}}(\lambda)\|}\boldsymbol{x}_i$, converges to the maximum margin, $\max_{\boldsymbol{w}:\|\boldsymbol{w}\|\le 1}\min_i y_i\boldsymbol{w}^\top\boldsymbol{x}_i$. By Lemma 2.1, any maximum margin classifier w.r.t. $\|\cdot\|$ gives a maximally robust classifier w.r.t. $\|\cdot\|_*$. □

F Linear Operations in Discrete Fourier Domain

Finding an adversarial sample with bounded Fourier-ℓ_∞ involves ℓ_∞ complex projection to ensure adversarial samples are bounded, as well as the steepest ascent direction w.r.t the Fourier-ℓ_∞ norm. We also use the complex projection onto ℓ_∞ simplex for proximal gradient method that minimizes the regularized empirical risk.

F.1 ℓ_∞ Complex Projection

Let $\boldsymbol{v}$ denote the ℓ_2 projection of $\boldsymbol{x} \in \mathbb{C}^d$ onto the ℓ_∞ unit ball. It can be computed as,

$$\arg\min_{\|\boldsymbol{v}\|_\infty\le 1} \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{x}\|_2^2 \tag{13}$$

$$= \{\boldsymbol{v} : \forall i,\ \boldsymbol{v}_i = \arg\min_{|\boldsymbol{v}_i|\le 1}\frac{1}{2}|\boldsymbol{v}_i - \boldsymbol{x}_i|^2\}\,, \tag{14}$$

that is independent projection per coordinate which can be solved by 2D projections onto ℓ_2 the unit ball in the complex plane.

F.2 Steepest Ascent Direction w.r.t. Fourier-ℓ_∞

Consider the following optimization problem,

$$\arg\max_{\boldsymbol{v}:\|\boldsymbol{F}\boldsymbol{v}\|_\infty\le 1} f(\boldsymbol{v})\,, \tag{15}$$

where $\boldsymbol{F} \in \mathbb{C}^{d\times d}$ is the Discrete Fourier Transform (DFT) matrix and $\boldsymbol{F}^* = \boldsymbol{F}^{-1}$ and $\boldsymbol{F}^*$ is the conjugate transpose.

Normalized steepest descent direction is defined as (See Boyd & Vandenberghe (2004, Section 9.4)),

$$\arg\min_{\boldsymbol{v}}\{\nabla\langle f(\boldsymbol{w}), \boldsymbol{v}\rangle : \|\boldsymbol{v}\| = 1\}\,. \tag{16}$$

Similarly, we can define the steepest ascent direction,

$$\arg\max_{\boldsymbol{v}\in\mathbb{R}^d}\{|\langle\nabla f(\boldsymbol{w}), \boldsymbol{v}\rangle| : \|\boldsymbol{F}\boldsymbol{v}\|_\infty = 1\} \tag{17}$$

$$\text{(Assuming } f \text{ is linear.)} \tag{18}$$

$$\arg\max_{\boldsymbol{v}\in\mathbb{R}^d}\{|\langle\boldsymbol{g}, \boldsymbol{F}^*\boldsymbol{F}\boldsymbol{v}\rangle| : \|\boldsymbol{F}\boldsymbol{v}\|_\infty = 1\} \tag{19}$$

$$\arg\max_{\boldsymbol{v}\in\mathbb{R}^d}\{|\langle\boldsymbol{F}\boldsymbol{g}, \boldsymbol{F}\boldsymbol{v}\rangle| : \|\boldsymbol{F}\boldsymbol{v}\|_\infty = 1\} \tag{20}$$

where $\boldsymbol{g} = \nabla f(\boldsymbol{w})$.

Consider the change of variable $\boldsymbol{u} = \boldsymbol{F}\boldsymbol{v} \in \mathbb{C}^{d\times d}$. Since $\boldsymbol{v}$ is a real vectors, is DFT is Hermitian, i.e. $\boldsymbol{u}_i^* = [\boldsymbol{u}]_{\overline{-i}}$ for all coordinates i where $\overline{j} = j \mod d$. Similarly, $\boldsymbol{F}\boldsymbol{g}$ is Hermitian.

$$\underset{\boldsymbol{u}\in\mathbb{C}^d:\|\boldsymbol{u}\|_\infty=1}{\arg\max} \quad \{|\langle \boldsymbol{F}\boldsymbol{g}, \boldsymbol{u}\rangle| : \boldsymbol{u}_i^* = [\boldsymbol{u}]_{\overline{-i}}\} \tag{21}$$

$$\underset{\boldsymbol{u}\in\mathbb{C}^d:\forall i,|\boldsymbol{u}_i|=1}{\arg\max} \quad \{|[\boldsymbol{F}\boldsymbol{g}]_i\boldsymbol{u}_i| + |[\boldsymbol{F}\boldsymbol{g}]_{\overline{-i}}\,\boldsymbol{u}_i^*| : \boldsymbol{u}_i^* = [\boldsymbol{u}]_{\overline{-i}}\} \tag{22}$$

$$\underset{\boldsymbol{u}\in\mathbb{C}^d:\forall i,|\boldsymbol{u}_i|=1}{\arg\max} \quad \{|[\boldsymbol{F}\boldsymbol{g}]_i\boldsymbol{u}_i| + |[\boldsymbol{F}\boldsymbol{g}]_i^*\boldsymbol{u}_i^*| : \boldsymbol{u}_i^* = [\boldsymbol{u}]_{\overline{-i}}\} \tag{23}$$

$$\underset{\boldsymbol{u}\in\mathbb{C}^d:\forall i,|\boldsymbol{u}_i|=1}{\arg\max} \quad \{|[\boldsymbol{F}\boldsymbol{g}]_i\boldsymbol{u}_i| : \boldsymbol{u}_i^* = [\boldsymbol{u}]_{\overline{-i}}\} \tag{24}$$

$$\boldsymbol{u}_i = [\boldsymbol{F}\boldsymbol{g}]_i / |[\boldsymbol{F}\boldsymbol{g}]_i| \,. \tag{25}$$

and the steepest ascent direction is $\boldsymbol{v}_i = \boldsymbol{F}^{-1}\boldsymbol{u}_i$ which is a real vector. In practice, there can be non-zero small imaginary parts because of numerical issues which we remove.

G Related work

This paper bridges two bodies of work, on adversarial robustness and optimization bias.

Hypotheses. Goodfellow et al. (2014) proposed the linearity hypothesis that informally suggests ℓ_p adversarial samples exist because deep learning models converge to functions similar to linear models. To improve robustness, they argued models have to be more non-linear. Based on our framework, linear models are not inherently weak. When trained, regularized, and parametrized appropriately they can be robust to some degree, the extent of which depends on the dataset. Gilmer et al. (2018) proposed adversarial spheres as a toy example where a two layer neural network exists with perfect standard and robust accuracy for non-zero perturbations. Yet, training a randomly initialized model with gradient descent and finite data does not converge to a robust model. Based on our framework, we interpret this as an example where the implicit bias of gradient descent is not towards the ground-truth model, even though there is no misalignment in the architecture. It would be interesting to understand this implicit bias in future work.

Fourier Analysis of Robustness. Fourier domain is a natural representation for analysis of human perception and describing image operations (Wandell & Thomas, 1997; González & Woods, 2008). Various observations have been made about Fourier properties of adversarial perturbations against deep non-linear models (Ilyas et al., 2019; Tsuzuku & Sato, 2019; Sharma et al., 2019). Yin et al. (2019) showed that adversarial training increases robustness to perturbations concentrated at high frequencies and reduces robustness to perturbations concentrated at low frequencies. Ortiz-Jimenez et al. (2020) also observed that the measured margin of classifiers in high frequencies is larger than the margin in low frequencies hence the robustness to high frequency perturbations is higher than low frequency perturbations. Our result does not distinguish between low and high frequencies but we establish an exact characterization of robustness. Caro et al. (2020) hypothesized about the implicit robustness to ℓ_1 perturbations in the Fourier domain while we prove maximal robustness to Fourier-ℓ_∞ bounded perturbations. Still a gap remains between our results and deep non-linear convolutional networks.

Architectural Robustness. An implication of our results is that robustness can be achieved at a lower computational cost compared with adversarial training by various architectural choices as recently explored (Xie et al., 2020; Galloway et al., 2019; Awais et al., 2020). Morever, for architectural choices that align with human biases, standard generalization can also improve (Vasconcelos et al., 2020). Another potential future direction is to rethink ℓ_p robustness as an architectural bias and find inspiration in human visual system for appropriate architectural choices.

Certified Robustness. Adversarially trained models are empirically harder to attack than standard models. But their robustness is not often provable. Certifiably robust models seek to close this gap (Hein & Andriushchenko, 2017; Wong & Kolter, 2018; Cohen et al., 2019; Gowal et al., 2018; Salman et al., 2019). A model is certifiably robust if for any input, it also provides an ε-certificate that guarantees robustness to any perturbation within the ε-ball of the input. In contrast, a maximally robust classifier finds a classifier that is guaranteed to be robust to maximal ε while classifying all training data correctly. That allows for data dependent robustness guarantees at test time. In this work, we have not explored standard generalization guarantees.

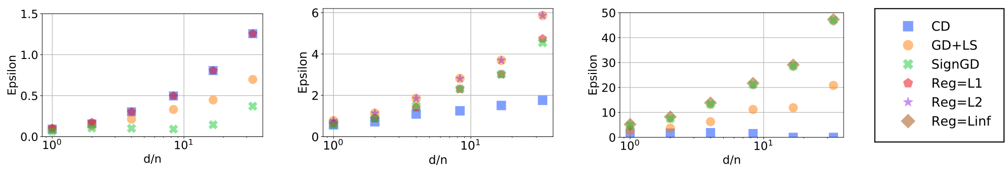
Robustness Trade-offs Most prior work define the metric for robustness and generalization using an expectation over the loss. We define robustness based on the accuracy and express it as a set of conditions to be satisfied. Our approach better matches the security perspective that even a single inaccurate prediction is a vulnerability. On the other hand, explicit constraints only ensure perfect accuracy on the training set. As such, standard generalization remains to be studied using other approaches in deep learning such as using additional assumptions on the data distribution. Existing work has used assumptions of the data distribution to achieve explicit trade-offs between robustness and standard generalization (Dobriban et al., 2020; Javanmard & Soltanolkotabi, 2020; Javanmard et al., 2020; Raghunathan et al., 2020; Tsipras et al., 2018; Zhang et al., 2019; Schmidt et al., 2018; Fawzi et al., 2018a;b).

Robust Optimization A robust counterpart to an optimization problem considers uncertainty in the data and optimizes for the worst-case. Ben-Tal et al. (2009) provided extensive formulations and discussions on robust counterparts to various convex optimization problems. Adversarial robustness is one such robust counterpart and many other robust counterparts could also be considered in deep learning. An example is adversarial perturbations with different norm-ball constraints at different training inputs. Madry et al. (2017) observed the link between robust optimization and adversarial robustness where the objective is a min-max problem that minimizes the worst-case loss. However, they did not consider the more challenging problem of maximally robust optimization that we revisit.

Implicit bias of optimization methods. Minimizing the empirical risk for an overparametrized model with more parameters than the training data has multiple solutions. Zhang et al. (2017) observed that overparametrized deep models can even fit to randomly labeled training data, yet given correct labels they consistently generalize to test data. This behavior has been explained using the implicit bias of optimization methods towards particular solutions. Gunasekar et al. (2018a) proved that minimizing the empirical risk using steepest descent and mirror descent have an implicit bias towards minimum norm solutions in overparametrized linear classification. Characterizing the implicit bias in linear regression proved to be more challenging and dependent on the initialization. Ji & Telgarsky (2018) proved that training a deep linear classifier using gradient descent not only implicitly converges to the minimum norm classifier in the space of the product of parameters, each layer is also biased towards rank-1 matrices aligned with adjacent layers. Gunasekar et al. (2018b) proved the implicit bias of gradient descent in training linear convolutional classifiers is towards minimum norm solutions in the Fourier domain that depends on the number of layers. Ji & Telgarsky (2020) has established the directional alignment the training of deep linear networks using gradient flow as well as the implicit bias of training deep 2-homogeneous networks. In the case of gradient flow (gradient descent with infinitesimal step size) the implicit bias of training multi-layer linear models is towards rank-1 layers that satisfy directional alignment with adjacent layers (Ji & Telgarsky, 2020, Proposition 4.4). Recently, Yun et al. (2020) has proposed a unified framework for implicit bias of neural networks using tensor formulation that includes fully-connected, diagonal, and convolutional networks and weakened the convergence assumptions.

Recent theory of generalization in deep learning, in particular the double descent phenomenon, studies the generalization properties of minimum norm solutions for finite and noisy training sets (Hastie et al., 2019). Characterization of the double descent phenomenon relies on the implicit bias of optimization methods, uses additional assumptions about the data distribution. In contrast, our results only rely on the implicit bias of optimization and hence are independent of the data distribution.

Robustness to ℓ_p-bounded attacks. Robustness is achieved when any perturbation to natural inputs that changes a classifiers prediction also confuses a human. ℓ_p-bounded attacks are the first step in achieving adversarial robustness. Tramer et al. (2020) has recently shown many recent robust models only achieve spurious robustness against ℓ_∞ and ℓ_1 attacks. Croce & Hein (2020) shows that on image classification datasets there is still a large gap in adversarial robustness to ℓ_p-bounded attacks and standard accuracy. Robustness to multiple ℓ_p-bounded perturbations through adversarial training and its trade-offs has also been analyzed (Tramèr & Boneh, 2019; Maini et al., 2020). Sharif et al. (2018); Sen et al. (2019) argue that none of ℓ_0, ℓ_1, ℓ_∞, or SSIM are a perfect match for human perception of similarity. That is for any such norm, for any ε, there exists a perturbation such that humans classify it differently. Attacks based on other perceptual similarity metrics exist (Zhao et al., 2020; Liu et al., 2019). This shows that the quest for adversarial robustness should also be seen as a quest for understanding human perception.



(a) ℓ_∞ attack (b) ℓ_2 attack (c) ℓ_1 attack

Figure 4: **Maximally robust ε for linear classification against ℓ_∞, ℓ_2, and ℓ_1 attacks.** For each attack, there exists one optimizer and one regularization method that finds a maximally robust classifier. We compare Coordinate Descent (CD), Gradient Descent with Line Search (GD+LS), Sign Gradient Descent (SignGD), and explicit ℓ_1, ℓ_2, and ℓ_∞ regularization. The margin of these classifiers and additional methods are shown in Fig. 5.

Robustness through Regularization Various regularization methods have been proposed for adversarial robustness that penalize the gradient norm and can be studied using the framework of maximally robust classification. Lyu et al. (2015) proposed general ℓ_p norm regularization of gradients. Hein & Andriushchenko (2017) proposed the Cross-Lipschitz penalty by regularizing the norm of the difference between two gradient vectors of the function. Ross & Doshi-Velez (2018) proposed ℓ_2 regularization of the norm of the gradients. Sokolić et al. (2017) performed regularization of Frobenius norm of the per-layer Jacobian. Moosavi-Dezfooli et al. (2019) proposed penalizing the curvature of the loss function. Qin et al. (2019) proposed encouraging local linearity by penalizing the error of local linearity. Simon-Gabriel et al. (2019) proposed regularization of the gradient norm where the dual norm of the attack norm is used. Ma et al. (2020) proposed Hessian regularization. Guo et al. (2020) showed that some regularization methods are equivalent or perform similarly in practice. Strong gradient or curvature regularization methods can suffer from gradient masking Ma et al. (2020).

H Extended Experiments

H.1 Details of Linear Classification Experiments

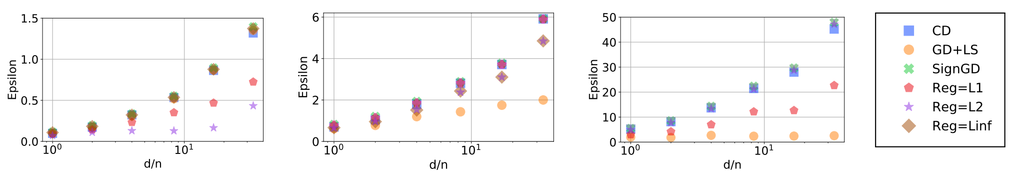
For experiments with linear classifiers, we sample n training data points from the $\mathcal{N}(0, \mathbb{I}_d)$, d-dimensional standard normal distribution centered at zero. We label data points $y = \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x})$, using a ground-truth linear separator sampled from $\mathcal{N}(0, \mathbb{I}_d)$. For $n < d$, the generated training data is linearly separable. This setting is similar to a number of recent theoretical works on the implicit bias of optimization methods in deep learning and specifically the double descent phenomenon in generalization (Montanari et al., 2019; Deng et al., 2019). We focus on robustness against norm-bounded attacks centered at the training data, in particular, ℓ_2, ℓ_∞, ℓ_1 and Fourier-ℓ_∞ bounded attacks.

Because the constraints and the objective in the minimum norm linear classification problem are convex, we can use off-the-shelf convex optimization toolbox to find the solution for small enough d and n. We use the CVXPY library (Diamond & Boyd, 2016). We evaluate the following approaches based on the implicit bias of optimization: Gradient Descent (GD), Coordinate Descent (CD), and Sign Gradient Descent (SignGD) on fully-connected networks as well as GD on linear two-layer convolutional networks (discussed in Section 3). We also compare with explicit regularization methods (discussed in Section 4) trained using proximal gradient methods (Parikh & Boyd, 2013). We do not use gradient descent because ℓ_p norms can be non-differentiable at some points (e.g. ℓ_1 and ℓ_∞) and we seek a global minima of the regularized empirical risk. We also compare with adversarial training. As we discussed in Section 1 we need to provide the value of maximally robust ε to adversarial training for finding a maximally robust classifier. In our experiments, we give an advantage to adversarial training by providing it with the maximally robust ε. We also use the steepest descent direction corresponding to the attack norm to solve the inner maximization.

For regularization methods a sufficiently small regularization coefficient achieves maximal robustness. Adversarial training given the maximal ε also converges to the same solution. We tune all hyperparameters for all methods including learning rate regularization coefficient and maximum step size in line search. We provide a list of values in Table 1.

Hyperparameter	Values
Random seed	0,1,2
d	100
d/n	$1, 2, 4, 8, 16, 32$
Training steps	10000
Learning rate	$1e{-}5$, $3e{-}5$, $1e{-}4$, $3e{-}4$, $1e{-}3$, $3e{-}3$, $1e{-}2$, $3e{-}2$, $1e{-}1$, $3e{-}1$, 1, 2, 3, 6, 9, 10, 20, 30, 50
Reg. coefficient	$1e{-}7$, $1e{-}6$, $1e{-}5$, $1e{-}4$, $1e{-}3$, $1e{-}2$, $1e{-}1$, 1, 10, $3e{-}3$, $5e{-}3$, $3e{-}2$, $5e{-}2$, $3e{-}1$, $5e{-}1$
Line search max step	1, 10, 100, 1000
Adv. Train steps	10
Adv. Train learning rate	0.1
Runtime (line search/prox. method)	$<$ 20 minutes
Runtime (others)	$<$ 2 minutes

Table 1: **Range of Hyperparameters.** Each run uses 2 CPU cores.



(a) ℓ_∞ attack (b) ℓ_2 attack (c) ℓ_1 attack

Figure 5: **Margin of models in Fig. 4.** Models are trained to be robust against ℓ_∞, ℓ_2, ℓ_1 attacks. For each attack, there exists one optimizer and one regularization method that finds the maximally robust classifier. Adversarial training also finds the solution given the maximal ε.

H.2 Margin Figures

A small gap exists between the solution found using CVXPY compared with coordinate descent. That is because of limited number of training iterations. The convergence of coordinate descent to minimum ℓ_1 norm solution is slower than the convergence of gradient descent to minimum ℓ_2 norm solution. There is also a small gap between the solution of ℓ_1 regularization and CVXPY. The reason is the regularization coefficient has to be infinitesimal but in practice numerical errors prevent us from training using very small regularization coefficients.

H.3 Visualization of Fourier Adversarial Attacks

In Figs. 7 to 9 we visualize adversarial samples for models available in RobustBench (Croce et al., 2020). Fourier-ℓ_∞ adversarial samples are qualitatively different from ℓ_∞ adversarial samples as they concentrate on the object.

I Visualization of Norm-balls

To reach an intuition of the norm-ball for Fourier ℓ_∞ norm, we visualize a number of common norm-balls in 3D in Fig. 10. Norm-balls have been visualized in prior work (Bach et al., 2012) but we are not aware of any visualization of Fourier-ℓ_∞.

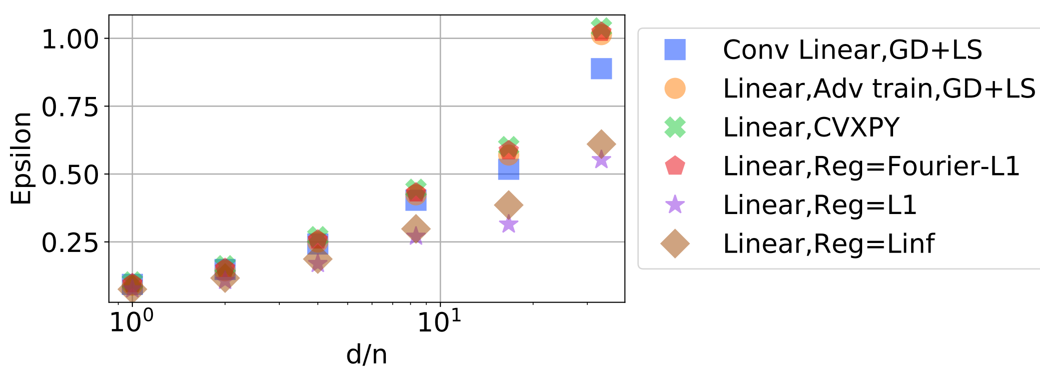
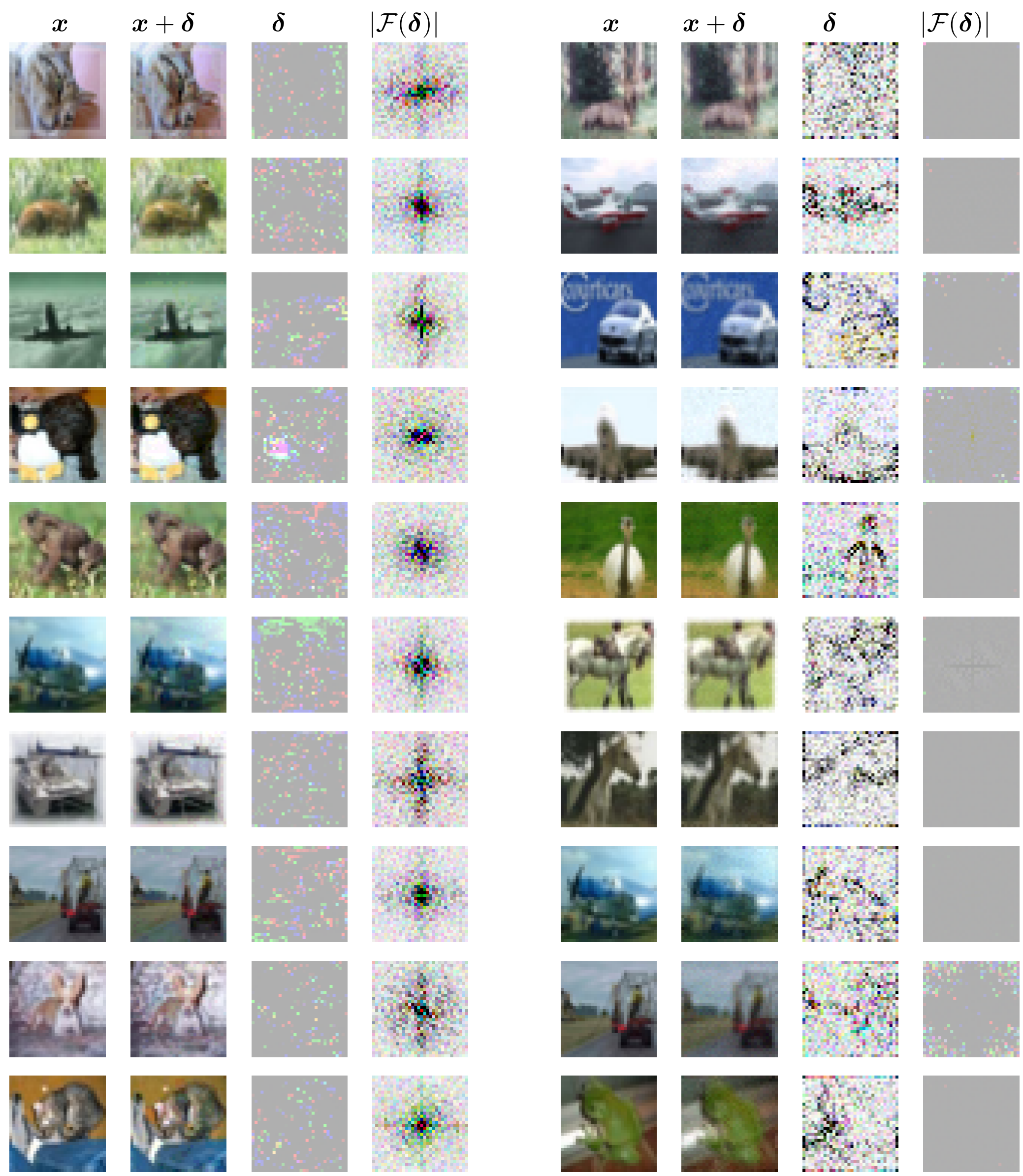


Figure 6: **Fourier-ℓ_1 margin of Linear Convolutional Models.**

(a) ℓ_∞ attack

(b) Fourier-ℓ_∞ attack

Figure 7: **Adversarial attacks (ℓ_∞ and Fourier-ℓ_∞) against CIFAR-10 with standard training.** WideResNet-28-10 model with standard training. The attack methods are APGD-CE and APGD-DLR with default hyperparameters in RobustBench. We use $\varepsilon = 8/255$ for both attacks. Fourier-ℓ_∞ perturbations are more concentrated on the object.



(a) ℓ_∞ attack (b) Fourier-ℓ_∞ attack

Figure 8: **Adversarial attacks (ℓ_∞ and Fourier-ℓ_∞) against CIFAR-10 ℓ_∞ model of (Carmon et al., 2019).** Adversarially trained model against ℓ_∞ attacks. The attack methods are APGD-CE and APGD-DLR with default hyperparameters in RobustBench. We use $\varepsilon = 8/255$ for both attacks. Fourier-ℓ_∞ perturbations are more concentrated on the object.

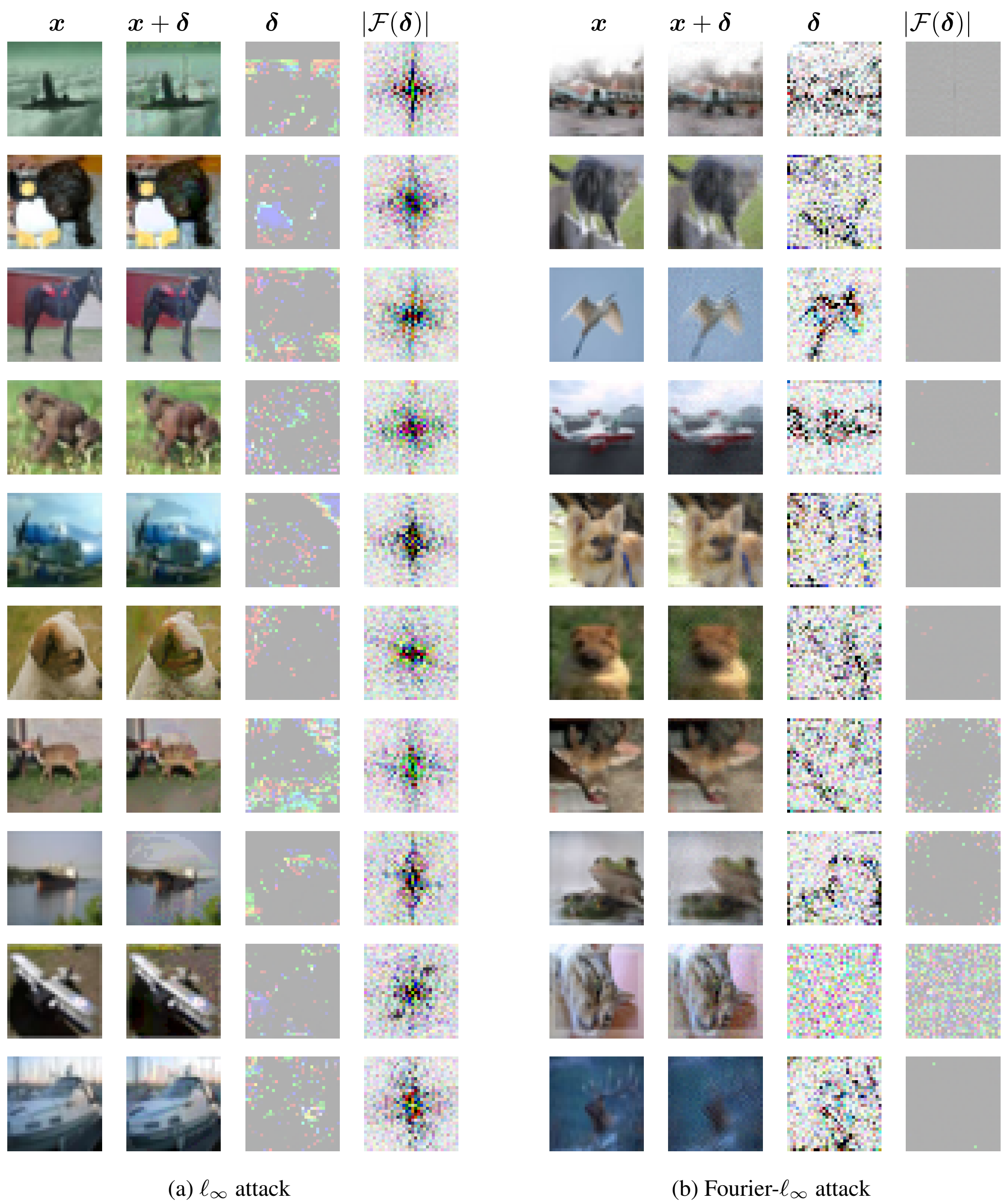
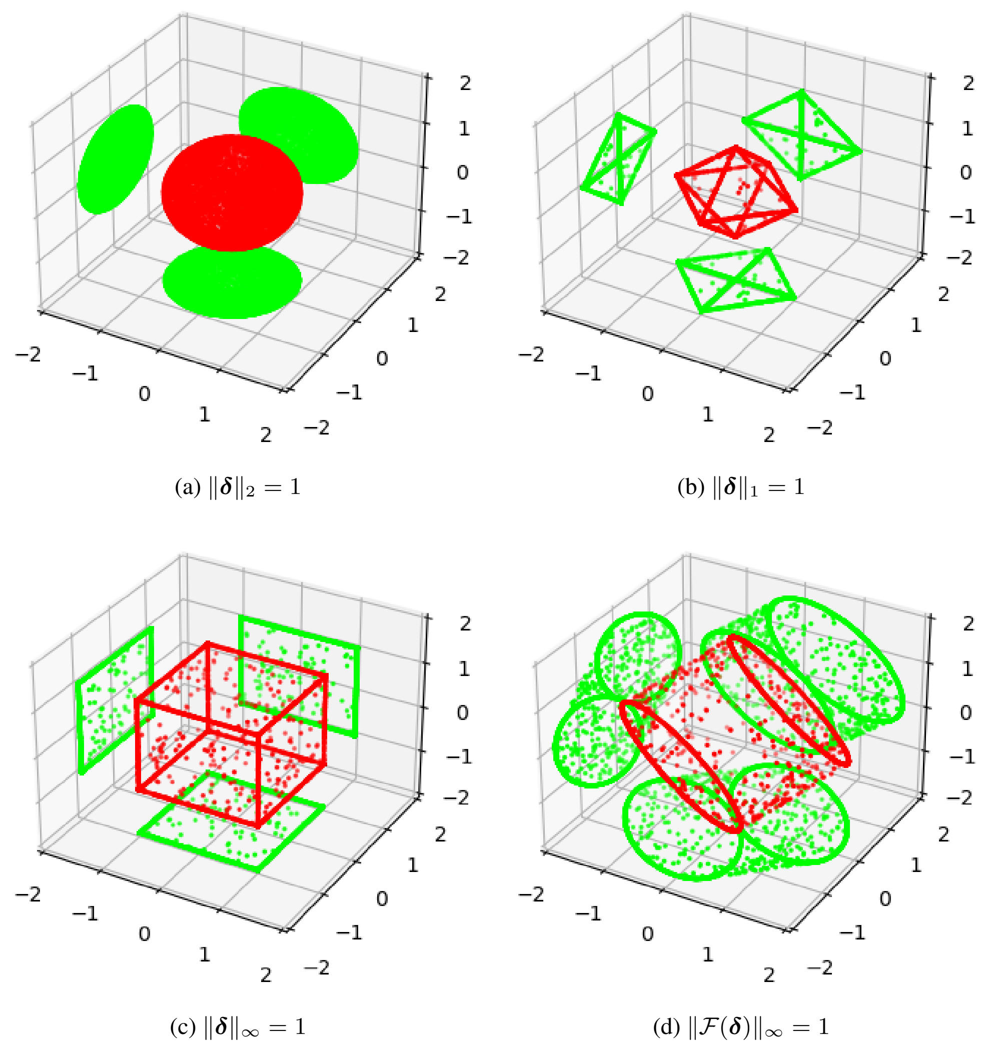


Figure 9: **Adversarial attacks (ℓ_∞ and Fourier-ℓ_∞) against CIFAR-10 ℓ_2 model of (Augustin et al., 2020).** Adversarially trained model against ℓ_2 attacks. The attack methods are APGD-CE and APGD-DLR with default hyperparameters in RobustBench. We use $\varepsilon = 8/255$ for both attacks. Fourier-ℓ_∞ perturbations are more concentrated on the object.



(a) $\|\boldsymbol{\delta}\|_2 = 1$

(b) $\|\boldsymbol{\delta}\|_1 = 1$

(c) $\|\boldsymbol{\delta}\|_\infty = 1$

(d) $\|\mathcal{F}(\boldsymbol{\delta})\|_\infty = 1$

Figure 10: Unit norm balls in 3-D (red) and their 2-D projections (green). Linear models trained with gradient descent are maximally robust to ℓ_2 perturbations. Two-layer linear convolutional networks trained with gradient descent are maximally robust to perturbations with bounded Fourier-ℓ_∞.