The Price of Robustness: Stable Classifiers Need Overparameterization

Jonas von Berg BERG@MATH.LMU.DE
Ludwig-Maximilians-Universität München,
Munich Center for Machine Learning (MCML)

Adalbert Fono FONO@MATH.LMU.DE
Ludwig-Maximilians-Universität München,
Munich Center for Machine Learning (MCML)

Massimiliano Datres DATRES@MATH.LMU.DE
Ludwig-Maximilians-Universität München

Sohir Maskey SOHIR.MASKEY@ALEPH-ALPHA-RESEARCH.COM
Ludwig-Maximilians-Universität München
Aleph Alpha Research

Gitta Kutyniok KUTYNIOK@MATH.LMU.DE
Ludwig-Maximilians-Universität München
University of Tromsø
DLR-German Aerospace Center
Munich Center for Machine Learning (MCML)

Abstract

In this work, we show that *class stability*, the expected distance of an input to the decision boundary, captures what classical capacity measures, such as weight norms, fail to explain. In particular, we prove a generalization bound that improves inversely with the class stability. As a corollary, interpreting class stability as a quantifiable notion of robustness, we derive a *law of robustness for classification* that extends results by Bubeck and Selke beyond smoothness assumptions to discontinuous functions. Specifically, any interpolating model with $p \approx n$ parameters on n data points must be *unstable*, implying that high stability requires substantial overparameterization. Preliminary experiments support this theory: empirical stability increases with model size, while traditional norm-based measures remain uninformative.

1. Introduction

The generalization behavior of overparameterized neural networks presents fundamental challenges to classical statistical learning theory. Traditional complexity measures, such as parameter counts or spectral norms of weights, form the basis of many generalization bounds, including those derived from VC dimension theory and Rademacher complexity. However, these quantities do not exhaustively explain several empirical phenomena, e.g., *double descent* [3] and *benign overfitting* [2]. The occurrence of double descent illustrates that the test error, after initially increasing near the interpolation threshold, can improve as the the model size continues to grow. Similarly, the phenomenon of benign overfitting demonstrates that models that perfectly interpolate noisy training

data can nonetheless achieve strong generalization. These findings expose the limitations of norm- and size-based complexity measures as predictors of generalization.

An emerging perspective suggests that generalization in modern networks is not only governed by model capacity but also by the *stability*, aka *robustness*, of predictions under input perturbations [8, 22, 24]. Similar insights arise from the literature on stability for learning algorithms [5] and flat minima [11]. In the context of regression under mild assumptions on the data distribution, a link between robustness, generalization, and overparameterization can be formally proven: small Lipschitz constants or local smoothness of a function correlate with strong generalization [4, 7]. However, this "law of robustness" crucially relies on the assumption that the function class is Lipschitz, making it inadequate for classifiers, whose co-domain is discrete by design. One could attempt to circumvent the issue by studying the Lipschitz constant of the underlying score function g, where the classifier f is obtained as $f := \arg\max \circ g$. This approach is not informative since g can be arbitrarily rescaled without altering the class prediction of f, and thus its Lipschitz constant may not reflect the geometry of the decision boundary [14]. This renders existing smoothness-based theories vacuous in the classification setting, a fundamental problem in deep learning.

1.1. Our contributions

Our contributions are twofold. First, we establish that the data-dependent Rademacher complexity of a finite hypothesis class of classifiers can be bounded in terms of the minimum class stability (see Definition 1). This leads to a new generalization bound for discontinuous classifiers, which sharpens with increasing stability of the function class. Second, we further show that in the classically parameterized regime (# parameters $\approx$ # training samples), any interpolating classifier must be unstable. Hence, achieving almost-perfect fitting as well as high class stability requires significant overparameterization. Together, these results extend the law of robustness to classifiers and provide a unified theoretical foundation for understanding generalization in modern deep networks.

2. Preliminaries and Notation

Next, we provide background on the key concepts relevant to our analysis, including stability, generalization, and isoperimetry. We consider a binary classification setting: Let $(\mathcal{X} \times \{-1, 1\}, \mu)$ be a probability measure space with $\mathcal{X} \subset \mathbb{R}^d$ bounded and $\mathcal{F} \subset \{f \mid f : \mathcal{X} \to \{-1, 1\}\}$ a finite set of classifiers. The goal is to find a stable function $f \in \mathcal{F}$ minimizing a bounded loss function $\ell : \{-1, 1\}^2 \to \mathbb{R}_+$ on n i.i.d. samples $(x_i, y_i) \sim \mu$. A natural loss in the classification setting is the 0-1 loss $\ell_{0\text{-}1}(y, y') := \mathbb{1}_{y \neq y'}$. In this setup, following a similar approach as in [14], we define the *class stability* of f as the expected distance of a sample to the decision boundary in $\mathcal{X}$, thereby capturing the average robustness of a classifier f to input perturbations.

Definition 1 (Local and Class Stability) *For $f : \mathcal{X} \to \{-1, 1\}$, denote by $\bar{f} : \mathbb{R}^d \to \{-1, 0, 1\}$ its* extension *given by $\bar{f}(x) = f(x)$ for $x \in \mathcal{X}$, and $\bar{f}(x) = 0$ otherwise. We define the* local stability *of f at $x \in \mathcal{X}$ as:*

$$h_f(x) := \inf\{\|x - z\|_2 : \bar{f}(x) \neq \bar{f}(z),\, z \in \mathbb{R}^d\}.$$

The class stability *of f on $\mathcal{X}$ is defined as the expected local stability with respect to the marginal distribution of μ, that is,*

$$S(f) := \int_{\mathcal{X}} h_f(x)\, d\mu = \mathbb{E}[h_f].$$

Our goal is to relate the class stability to the Rademacher complexity of a function class, which, in turn, connects to *generalization* bounds through classical results [1]. In particular, for a bounded loss $|\ell| \leq a$, the difference between the *population risk* $R_\ell(f) := \mathbb{E}[\ell(f(x), y)]$ and the *empirical risk* $\hat{R}_\ell(f) := \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$ is bounded with probability at least $1-\delta$ over the samples by

$$\sup_{f\in\mathcal{F}}\big(R_\ell(f) - \hat{R}_\ell(f)\big) \leq 2\mathcal{R}_{n,\mu}(\ell\circ\mathcal{F}) + a\sqrt{\frac{2\log(2/\delta)}{n}}, \tag{1}$$

where $\mathcal{R}_{n,\mu}(\mathcal{G})$ denotes the *Rademacher complexity* of a general function class $\mathcal{G}$, defined as

$$\mathcal{R}_{n,\mu}(\mathcal{G}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n \sigma_i g(x_i)\right|\right],$$

with $(\sigma_i)_{i=1}^n$ i.i.d. Rademacher random variables. To obtain a bound in (1) in terms of $\mathcal{R}_{n,\mu}(\mathcal{F})$, note that $\mathcal{R}_{n,\mu}(\ell\circ\mathcal{F}) \leq C\mathcal{R}_{n,\mu}(\mathcal{F})$ holds under certain conditions on the loss, e.g., we have

$$\mathcal{R}_{n,\mu}(\ell_{0\text{-}1}\circ\mathcal{F}) \leq \frac{1}{2}\mathcal{R}_{n,\mu}(\mathcal{F}), \quad \text{i.e., } C = \frac{1}{2}, \tag{2}$$

whereas for L-Lipschitz losses $C = L$ holds, see [1, 21] for detailed explanations. Overall, it therefore suffices to bound $\mathcal{R}_{n,\mu}(\mathcal{F})$ in terms of the class stability of functions $f \in \mathcal{F}$ to link generalization to stability. In other words, deriving the desired generalization bound, requires tightly controlling how well stable functions can fit random labels, which demands structural assumptions on the input distribution. We refer to Appendix A for further discussions on why certain assumptions regarding the underlying distribution are necessary. A natural and widely used condition is *isoperimetry*, which ensures sharp concentration for bounded Lipschitz-continuous functions [7].

Definition 2 (Isoperimetry) *A probability measure μ on $\mathcal{X} \subset \mathbb{R}^d$ satisfies* c-isoperimetry *if for any bounded L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, and any $t \geq 0$,*

$$\mathbb{P}(|f(x) - \mathbb{E}[f]| \geq t) \leq 2e^{-\frac{dt^2}{2cL^2}}. \tag{3}$$

Isoperimetry is, for instance, satisfied by Gaussian measures and the volume measure on Riemannian manifolds with positive curvature, such as the uniform measure on the sphere [7, 23].

3. A Law of Robustness for Classification

In this section, we establish a law of robustness for classification, extending stability-generalization trade-offs to discontinuous functions. Classical results for smooth functions characterize robustness via the Lipschitz constant, which is ill-defined for classifiers with discrete outputs. To address this, we adapt the general strategy of [7], presented in more detail in Appendix A, but substitute Lipschitz continuity with the notion of class stability introduced in Definition 1, i.e., we proceed under the following assumptions:

(H1) $(\mathcal{X}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric[1] measure μ;

1. It is worth noting that our framework can be readily extended to mixtures of c-isoperimetric distributions.

(H2) the considered hypothesis class $\mathcal{F}$ of classifiers $f : \mathcal{X} \to \{-1, 1\}$ is finite, that is $|\mathcal{F}| < \infty$.

Indeed, by exploiting class stability, we construct a Lipschitz-continuous restriction of a classifier to which [7, Lemma 4.1] applies. Controlling the measure of the complement via isoperimetry then yields the following result; see Appendix B for details.

Theorem 3 (Rademacher Bound) *Under assumptions (H1), (H2), suppose that* $\min_{f\in\mathcal{F}} S(f) > S > 0$ *and* $\log|\mathcal{F}| \geq n$. *Then, we have*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \leq K \max\left\{\sqrt{\frac{1}{n}}, \frac{1}{S}\frac{\sqrt{c}\,\log|\mathcal{F}|}{n\sqrt{d}}\right\}, \tag{4}$$

where $K > 0$ *is an absolute constant independent of* $\log|\mathcal{F}|, n, d, S, c$.

Remark 4 *In comparison to [7], where the stability is measured by the minimum Lipschitz constant of the function class, we obtain a scaling which is worse by a factor* $\sqrt{\frac{\log|\mathcal{F}|}{n}}$ *in the relevant regime* $\log|\mathcal{F}| \geq n$. *Intuitively, this gap arises from the stronger regularity imposed by Lipschitz continuity in contrast to class stability, which allows for jump-discontinuities; see Appendix B.*

The key insight of Theorem 3, combined with the classical generalization bound (1), is that *good generalization* can still be achieved in the highly *overparameterized* regime—provided the classifiers exhibit sufficiently *high class stability*. Indeed, the presence of $\frac{1}{S}$ in (4) shows that class stability impacts the effective complexity of the model class, possibly mitigating the risks of overfitting in large models. Note that, using a uniform discretization, a finite approximation of an infinite function class parameterized with p parameters over a bounded subset of $\mathbb{R}^p$ satisfies $\log|\mathcal{F}| \in \mathcal{O}(p)$. In this sense, $\log|\mathcal{F}|$ reflects the number of model parameters. Therefore, when the number of parameters $p \approx \log|\mathcal{F}|$ is much larger than n, the second term in the maximum in (4) dominates, and the bound becomes small if S scales at least in the order of $\frac{p}{n\sqrt{d}}$.

Finally, we are in a position to deduce our law of robustness for discontinuous functions, which follows as a corollary of the Rademacher bound in Theorem 3.

Corollary 5 (Law of Robustness for Discontinuous Functions) *Under Assumptions (H1), (H2), and* $p := \log|\mathcal{F}| \geq n$, *let us fix* $\varepsilon, \delta \in (0, 1)$ *and consider the 0-1 loss* $\ell_{0\text{-}1}$. *Then there exists an absolute constant* $K > 0$ *such that under the additional conditions*

1. *the minimal risk, defined as* $\sigma^2 := \min_{f\in\mathcal{F}} R_{0\text{-}1}(f)$, *satisfies* $\sigma^2 \geq \varepsilon$,
2. *the number of samples* n *is large enough such that* (i) $\frac{K}{\sqrt{n}} < \frac{\varepsilon}{3}$ *and* (ii) $\sqrt{\frac{2\log(2/\delta)}{n}} < \frac{\varepsilon}{2}$,

we have: With probability at least $1 - \delta$ *with respect to the sample, the following holds uniformly for all* $f \in \mathcal{F}$:

$$\hat{R}_{0\text{-}1}(f) \leq \sigma^2 - \varepsilon \implies S(f) < \frac{3K\sqrt{c}}{\varepsilon}\frac{p}{n\sqrt{d}}. \tag{5}$$

Remark 6 *Unlike the setting in [7], which assumes Lipschitz-continuous losses, our analysis treats the discontinuous 0-1 loss, which is more appropriate for classification problems. However, the general proof strategy remains valid for any loss as long as one can establish a suitable bound on the Rademacher complexity of the composed function class as in (2). Furthermore, this new result also covers intrinsically discontinuous classifiers such as quantized neural networks and spiking neural networks by design.*

Proof Let $K > 0$ be an absolute constant such that (4) holds and define the threshold stability

$$S_* = S_*(p, n, d, \varepsilon) := \frac{3K\sqrt{c}}{\varepsilon} \frac{p}{n\sqrt{d}}.$$

Then, Theorem 3, together with condition $2(i)$, implies that

$$\mathcal{R}_{n,\mu}(\mathcal{F}_{S_*}) \leq K \max\left\{ \sqrt{\frac{1}{n}}, \frac{1}{S_*} \frac{\sqrt{c}\, p}{n\sqrt{d}} \right\} \leq \varepsilon/3,$$

where $\mathcal{F}_{S_*} := \{f \in \mathcal{F} : S(f) \geq S_*\}$ is the subset of functions in $\mathcal{F}$ with stability at least S_*. Hence, applying the generalization inequality (1), together with condition $2(ii)$, gives with probability $1-\delta$:

$$\sup_{f \in \mathcal{F}_{S_*}} \left(R_{0\text{-}1}(f) - \hat{R}_{0\text{-}1}(f) \right) \leq 2\mathcal{R}_{n,\mu}(\ell_{0\text{-}1} \circ \mathcal{F}_{S_*}) + \sqrt{\frac{2\log(2/\delta)}{n}} \leq \mathcal{R}_{n,\mu}(\mathcal{F}_{S_*}) + \frac{\varepsilon}{2} < \varepsilon,$$

where we additionally used (2) in the second step. In particular, we can bound the probability

$$\mathbb{P}(\forall f \in \mathcal{F}_{S_*} : \hat{R}_{0\text{-}1}(f) > \sigma^2 - \varepsilon) \geq \mathbb{P}(\forall f \in \mathcal{F}_{S_*} : R_{0\text{-}1}(f) - \hat{R}_{0\text{-}1}(f) < \varepsilon) \geq 1 - \delta,$$

where the first inequality follows from

$$R_{0\text{-}1}(f) - \hat{R}_{0\text{-}1}(f) < \varepsilon \overset{\text{condition 1.}}{\implies} \sigma^2 - \hat{R}_{0\text{-}1}(f) < \varepsilon \implies \hat{R}_{0\text{-}1}(f) > \sigma^2 - \varepsilon.$$

Decomposing this probability into two disjoint events

$$\begin{aligned} 1 - \delta \leq \mathbb{P}(\forall f \in \mathcal{F}_{S_*} : \hat{R}_{0\text{-}1}(f) > \sigma^2 - \varepsilon) = \mathbb{P}(\forall f \in \mathcal{F} : \hat{R}_{0\text{-}1}(f) > \sigma^2 - \varepsilon) \\ + \mathbb{P}(\exists f \in \mathcal{F}_{S_*}^c : \hat{R}_{0\text{-}1}(f) \leq \sigma^2 - \varepsilon). \end{aligned} \tag{6}$$

enables us to easily recognize that the expression exactly characterizes the probability that the following implication, and thereby the result, holds uniformly for all $f \in \mathcal{F}$:

$$\hat{R}_{0\text{-}1}(f) \leq \sigma^2 - \varepsilon \implies S(f) < S_*.$$

Indeed, the implication above holds if, for a given data sample $(x_i, y_i)_{i=1}^n$, either

- no function $f \in \mathcal{F}$ satisfies $\hat{R}_{0\text{-}1}(f) \leq \sigma^2 - \varepsilon$, or
- any such f lies in $\mathcal{F}_{S_*}^c$, that is, $S(f) < S_*$,

which is the case with probability at least $1 - \delta$ due to (6). ■

We conclude via (5) that achieving both low training error and high stability requires parametrization at the scale $p \approx n\sqrt{d}$, which indicates the necessity of overparameterization in high dimensions. This reinforces our central message: *overparameterization may not harm generalization, but on the contrary, is necessary for achieving robustness and good fitting in classification*. Notably, modern neural networks, including large language models (LLMs) [6], are trained in heavily overparameterized regimes, where the parameter count far exceeds the sample size, aligning with the proposed theory. Therefore, our result may help to understand why these models can still generalize effectively.

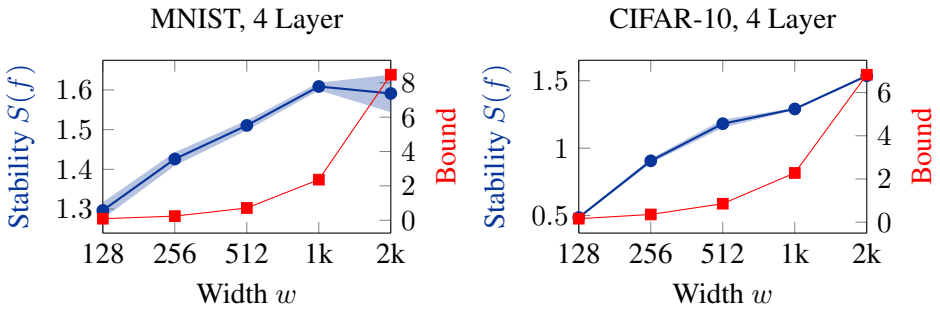

Figure 1: Empirical class stability $S(f)$ (blue, left axis) and theoretical bound from Theorem 5 (red, right axis) as a function of network width $w \propto \sqrt{p}$.

4. Experiments

We empirically validate our theoretical prediction that class stability $S(f)$ increases with model size in interpolating networks.

Setup. We train fully connected MLPs with 4 hidden layers and hidden dimensions $w \in \{128, 256, 512, 1024, 2048\}$ on MNIST and CIFAR-10. All models are trained until achieving at least 99.9% training accuracy, ensuring (almost) interpolation.

To estimate the empirical class stability $S(f)$, we perform adversarial attacks on each input. For an increasing sequence of perturbation radii $\mathbf{r} = (r_1, \ldots, r_n)$, we gradually increase r until the classifier's prediction changes. The smallest successful radius is recorded as the estimated distance to the decision boundary for that sample. We report $S(f)$ as the average of these distances.

Results. Figure 1 shows that, for MLPs, $S(f)$ consistently increases with model size. The growth trend matches the theoretical prediction $S(f) \sim p/n\sqrt{d}$, supporting our law of robustness. In contrast, standard weight norms show no consistent correlation with either model size or generalization performance. Additional details are provided in Appendix C.

5. Limitations and Future Work

On the empirical side, it would be valuable to investigate the validity of Assumption (H1), specifically the isoperimetric concentration properties of real-world datasets. Empirically testing these properties could not only shed light on the practical relevance of our theoretical assumptions but may also motivate alternative concentration inequalities beyond isoperimetry, allowing for broader applicability of our results.

A key limitation of our theoretical analysis lies in Assumption (H2), which restricts our results to finite function classes. Extending to infinite classes is challenging due to the discontinuity of parameterized classifiers with respect to their parameters. However, we believe that discretization of specific (infinite) classes of classifiers, where the discontinuities can be controlled, is still achievable.

Lastly, to better interpret our results as statements about robustness, it is interesting to relate class stability to the robust generalization error introduced in [15], as well as related notions of adversarial robustness. Exploring these connections offers a promising direction for future work.

Acknowledgements

J. von Berg and G. Kutyniok acknowledge support by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

S. Maskey acknowledges support by the NSF-Simons Research Collaboration on the Mathematical and Scientific Foundations of Deep Learning (MoDL) (NSF DMS 2031985).

J. von Berg, M. Datres and G. Kutyniok acknowledge support by the gAIn project, which is funded by the Bavarian Ministry of Science and the Arts (StMWK Bayern) and the Saxon Ministry for Science, Culture and Tourism (SMWK Sachsen).

G. Kutyniok, J. von Berg and A. Fono acknowledge support by the Munich Center for Machine Learning (MCML).

G. Kutyniok acknowledges support by the German Research Foundation under Grants DFG-SPP-2298, KU 1446/31-1 and KU 1446/32-1. Furthermore, G. Kutyniok is supported by LMUexcellent, funded by the Federal Ministry of Research, Technology and Space (BMFTR) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria.

References

[1] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[2] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[4] Simone Bombari, Shayan Kiyani, and Marco Mondelli. Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels, 2023. URL https://arxiv.org/abs/2302.01629.

[5] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[7] Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

[8] Nikhil Ghosh and Mikhail Belkin. A universal trade-off between the model size, test loss, and training loss of linear predictors, 2023. URL https://arxiv.org/abs/2207.11621.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6572.

[10] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[11] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

[13] Mojżesz Dawid Kirszbraun. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22:77–108, 1934. URL https://api.semanticscholar.org/CorpusID:117250450.

[14] Z. N. D. Liu and A. C. Hansen. Do stable neural networks exist for classification problems? – a new view on stability in ai, 2024. URL https://arxiv.org/abs/2401.07874.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1706.06083.

[16] BY E. J. McSHANE and Edward James McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40:837–842, 1934. URL https://api.semanticscholar.org/CorpusID:38462037.

[17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. doi: 10.1109/CVPR.2016.282.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/

9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[19] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017. URL https://arxiv.org/abs/1707.04131.

[20] Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics, 2023. URL https://arxiv.org/abs/2310.19244.

[21] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[22] Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Building a stable classifier with the inflated argmax, 2025. URL https://arxiv.org/abs/2405.14064.

[23] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[24] Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective, 2022. URL https://arxiv.org/abs/2210.01787.

Appendix A. The need for isoperimetry

Concentration inequalities are essential tools in high-dimensional probability theory, providing bounds on the tail behavior of random variables. Next, we outline the key strategy from Bubeck & Selke [7] for proving the law of robustness for regression, highlighting the importance of an additional assumption on the measure μ. The authors employ the Lipschitz constant of a function as a measure of robustness, where a small Lipschitz constant (i.e., ≈ 1) of the realization indicates a robust model. The basic idea is to leverage the Lipschitz continuity of functions $f : \mathcal{X} \to \mathbb{R}$ in conjunction with isoperimetric inequalities to bound the probability

$$\mathbb{P}(\exists f \in \mathcal{F} : \hat{R}_\ell(f) \approx 0 \quad \wedge \quad L(f) \leq L_*) < \delta. \tag{7}$$

That is, we aim to bound the probability of observing a model that is both robust (i.e., has a small Lipschitz constant $L(f)$) and fits the data well (i.e., $\hat{R}(f) \approx 0$, meaning it nearly interpolates).
By contraposition, this implies that with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$:

$$\hat{R}_\ell(f) \approx 0 \implies L(f) > L_*(p, n, d). \tag{8}$$

Here, $L_*(p, n, d)$ is an algebraic function of the number of parameters $p \approx \log|\mathcal{F}|$ (see the paragraph below Theorem 3 for details), the number of training samples n, and the input dimension d. It satisfies $L_*(p, n, d) \gg 1$ in the non-overparameterized regime $p \approx n$, thereby implying non-robust behavior.

A key ingredient in [7] for proving (a variant of) (7) is the isoperimetry assumption on the measure μ. Isoperimetry, originating in geometry, provides an upper bound on a set's volume in terms of its boundary's surface area. In high dimensions, the principle of isoperimetry induces a concentration of measure, where the measure of the ε-neighborhood A_ε of any set A with $\mu(A) > 0$ has measure $\mu(A_\varepsilon) \to 1$, and the complementary measure decays in the order of $\exp(-d\varepsilon^2)$. This is equivalent to the sub-Gaussian behavior of every bounded Lipschitz-continuous function as stated in Definition 2, yielding a concentration property for $|f(x) - \mathbb{E}(f)|$ that depends on the Lipschitz constant $L(f)$.

The induced concentration property allows us to bound the probability in (7), leveraging the intuition that a smaller Lipschitz constant limits the function's capacity to align with random labels. However, it is important to note that (8) provides information about robustness within $\mathcal{F}$ only if

$$\mathbb{P}(\nexists f \in \mathcal{F} : \hat{R}_\ell(f) \approx 0) \leq 1 - \delta \quad \iff \quad \mathbb{P}(\exists f \in \mathcal{F} : \hat{R}_\ell(f) \approx 0) \geq \delta.$$

Otherwise, the implication becomes vacuous, as almost no function in $\mathcal{F}$ generalizes well, i.e., achieves near-zero empirical risk, to begin with. Without imposing any assumptions on μ, Hoeffding's inequality already suffices to derive a Lipschitz-independent bound for any function $f : \mathcal{X} \to [-1, 1]$:

$$\mathbb{P}(|f(x) - \mathbb{E}(f)| \geq t) \leq 2\exp\left(-\frac{t^2}{2}\right) \quad \forall t > 0. \tag{9}$$

Thus, to ensure that the probability in (7) remains below δ while simultaneously allowing for $\mathbb{P}(\exists f \in \mathcal{F} : \hat{R}_\ell(f) \approx 0) > \delta$, any concentration inequality relying on the Lipschitz constant must exhibit a sufficiently fast decay (in comparison with (9)) in the regime $L(f) \gtrsim 1$. This is necessary to yield a non-vacuous bound in (8), which allows to assess robustness by the increase of the minimal Lipschitz constant L_* even for $L_* > 1$.

For instance, McDiarmid's inequality applied to Lipschitz functions yields a tail bound of the order $\exp(-\frac{2t^2}{\operatorname{diam}(\mathcal{X})^2 L(f)^2})$, which is insufficient as it decays faster than the Hoeffding bound only for $L(f) < 2/\operatorname{diam}(\mathcal{X})$, i.e., at least $\operatorname{diam}(\mathcal{X}) < 2$ is required to include the (relevant) range $L(f) > 1$ of Lipschitz constants. This indicates that a certain restriction of the admissible measures is indeed necessary to obtain non-vacuous statements, i.e., they can not be derived in full generality.

Notably, the c-isoperimetry condition (3) leads to a faster decay than the Hoeffding bound in (9) when $L(f) < \sqrt{dc^{-1}}$, making it effective for functions with moderate Lipschitz constants in high-dimensional settings. Our goal is to generalize this strategy to handle discontinuous functions, addressing the inherent challenges of classification tasks.

Appendix B. Proof of the Rademacher bound (Theorem 3)

In the regression setting, one can assume without loss of generality that the considered regressors are Lipschitz continuous and thereby derive insightful statements about the expected and feasible robustness of models in a given setting. In contrast, this approach is not meaningful anymore in the classification setting as the considered classifiers are (except for trivial cases) discontinuous by design, i.e., they can not be captured by a finite Lipschitz constant. Thus, statements about the robustness of classification models can not be derived via Lipschitz constants. This motivates the use of class stability as a replacement measure in the classification setting, which, however, is (inversely) related to Lipschitzness as highlighted and exploited in the subsequent proof of Theorem 3. For convenience, we repeat the statement with the corresponding assumptions.

(H1) $(\mathcal{X}, \mu)$ is a probability space with bounded sample space $\mathcal{X}$ and c-isoperimetric measure μ;

(H2) the considered hypothesis class $\mathcal{F}$ of classifiers $f : \mathcal{X} \to \{-1, 1\}$ is finite, that is $|\mathcal{F}| < \infty$.

Theorem *Under assumptions (H1), (H2), suppose that* $\min_{f\in\mathcal{F}} S(f) > S > 0$ *and* $\log|\mathcal{F}| \geq n$. *Then, we have*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \leq K \max\left\{\sqrt{\frac{1}{n}}, \frac{1}{S}\frac{\sqrt{c}\,\log|\mathcal{F}|}{n\sqrt{d}}\right\},$$

where $K > 0$ *is an absolute constant independent of* $\log|\mathcal{F}|, n, d, S, c$.

Proof : To begin, we explore the relationship between two measures of robustness: the Lipschitz constant $L(f)$ and the class stability $S(f)$ of a $f \in \mathcal{F}$ on the set

$$A_t(f) := \{x \in \mathcal{X} : h_f(x) > S(f) - t\} \quad \text{for } 0 \leq t \leq S(f).$$

Observe that for $x_1 \in A_t(f)$ and $x_2 \in \mathcal{X}$

$$\|f(x_1) - f(x_2)\| \leq \begin{cases} 0, & \text{if } f(x_1) = f(x_2) \\ 2 \cdot \overbrace{\frac{\|x_1 - x_2\|}{S(f) - t}}^{\geq 1}, & \text{if } f(x_1) \neq f(x_2) \end{cases} \leq \frac{2}{S(f) - t}\|x_1 - x_2\|,$$

i.e., f is $\frac{2}{S(f)-t}$-Lipschitz on $A_t(f)$ and, therefore, according to the assumption $S(f) > S$, any $f \in \mathcal{F}$ is at least $\frac{2}{S-t}$-Lipschitz on $A_t(f)$. Our strategy now is to apply the Rademacher bound

based on Lipschitz functions of Bubeck & Selke in [7] to the restriction $f_{|A_t(f)}$, and additionally exploit isoperimetry to control the measure of the complement $A_t(f)^c$. We rely on two key facts:

- **Fact 1**: Every Lipschitz continuous function $g : A \to \mathbb{R}$, defined on a subset $A \subset \mathcal{X}$ of a metric space, can be extended to a function $G_g : \mathcal{X} \to \mathbb{R}$, preserving the same Lipschitz constant ([16], [13]). $\implies$ *This allows us to apply isoperimetry and thereby the result in [7, Lemma 4.1] to the $\frac{2}{S-t}$-Lipschitz extension F_f of $f_{|A_t(f)}$ (by w.l.o.g. restricting its co-domain to $[-1,1]$) to obtain*

$$\frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i F_f(x_i)\right|\right] \leq C_1\frac{1}{\sqrt{n}} + C_2\frac{1}{S-t}\sqrt{\frac{c\,\log|\mathcal{F}|}{nd}}$$

  for some absolute constants $C_1, C_2 > 0$.

- **Fact 2**: The local stability $h_f(x) : \mathcal{X} \to \mathbb{R}$, is 2-Lipschitz continuous with respect to the ℓ_2-norm ([14, Prop. 7.5]. $\implies$ *This allows us to control $\mathbb{P}(A_t(f)^c)$ via isoperimetry:*

$$\mathbb{P}(A_t(f)^c) = \mathbb{P}(\overbrace{S(f)}^{=\mathbb{E}[h_f]} - h_f(x) \geq t) \leq \exp\left(-\frac{dt^2}{2cL(h_f)^2}\right) = \exp\left(-\frac{dt^2}{2^3c}\right). \tag{10}$$

Via Fact 1, we can bound the Rademacher complexity by

$$\begin{aligned}
\mathcal{R}_{n,\mu}(\mathcal{F}) &= \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] \\
&\leq \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i F_f(x_i)\right|\right] + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right|\right] \\
&\leq C_1\frac{1}{\sqrt{n}} + C_2\frac{1}{S-t}\sqrt{\frac{c\,\log|\mathcal{F}|}{nd}} + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right|\right],
\end{aligned} \tag{11}$$

To control the last term, we subdivide $\mathcal{X}^n$ into subsets on which specific samples achieve a minimum local stability. To that end, we fix $t = \frac{S}{2}$ (the exact value is not crucial since it will be subsumed into the absolute constants) and define, for $I \subset [n]$,

$$A^I(f) = A^I_{\frac{S}{2}}(f) := \left\{x \in \mathcal{X}^n : i \in I \iff h_f(x_i) \geq \frac{S}{2}\right\}.$$

Note, that $A^{[n]}(f) = A_{\frac{S}{2}}(f)^n$ and $\cup_{I\in\mathcal{P}([n])} A^I(f)$ is a disjoint partition of $\mathcal{X}^n$. Thus, applying a union bound yields for $r > 0$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right| > r\right) \leq \sum_{f\in\mathcal{F}}\sum_{I\in\mathcal{P}([n])} \mathbb{P}\left(\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right| > r \ \wedge\ x \in A^I(f)\right) \\
&= \sum_{f\in\mathcal{F}}\sum_{I\in\mathcal{P}([n])} \mathbb{P}\left(\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right| > r \,\middle|\, x \in A^I(f)\right)\mathbb{P}(A^I(f)).
\end{aligned} \tag{12}$$

We make the following observations:

- By construction $F_f = f$ holds on $A^I(f)$ for all $f \in \mathcal{F}$.
- As a mean-zero and bounded random variable with range $[-2, 2]$, $\sigma_i(F_f - f)(x_i)$ is (via Hoeffding's inequality) subgaussian with variance proxy $\frac{(2-(-2))^2}{4} = 4$ for every $i \in [n], f \in \mathcal{F}$.

Using the additional fact that the sum of k independent subgaussian random variables with variance proxy σ^2 is again subgaussian with variance proxy $k\sigma^2$ [20], implies for $I \subsetneq [n]$ (for $I = [n]$ the probability is trivially zero) that

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \sigma_i(f - F_f)(x_i)\right| > r \,\middle|\, x \in A^I(f)\right) \leq \mathbb{P}\left(\left|\sum_{i \in I^c} \sigma_i(f - F_f)(x_i)\right| > r \,\middle|\, x \in A^I(f)\right) \leq 2\exp\left(-\frac{r^2}{2 \cdot 4(n - |I|)}\right).$$

On the other hand, we get for $I \subset [n]$ via (10) that

$$\mathbb{P}\left(A^I(f)\right) \leq \mathbb{P}\left(\forall j \in I^c : \, x_j \in A_{\frac{S}{2}}(f)^c\right) = \mathbb{P}\left(x \in A_{\frac{S}{2}}(f)^c\right)^{n-|I|} \leq \exp\left(-\frac{dS^2}{2^5 c}\right)^{n-|I|}.$$

Inserting in (12) and replacing the constants independent of the parameters of interest ($n, |\mathcal{F}|, d, r, S$, and $|I|$) by $c_1, c_2 > 0$ then gives

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \sigma_i(f - F_f)(x_i)\right| > r\right) \leq \sum_{f \in \mathcal{F}} \sum_{I \in \mathcal{P}([n]) \setminus [n]} 2\exp\left(-\frac{r^2 c_1}{n - |I|}\right) \exp\left(-\frac{(n - |I|)dS^2 c_2}{c}\right).$$

To simplify the above expression, we want to find the maximal term in the sum and use this worst case as an upper bound over all terms in the sum. To that end, we introduce $g : [0, n) \to \mathbb{R}_+$ by

$$g(x) = \frac{r^2 c_1}{n - x} + \frac{1}{c}(n - x)S^2 d c_2,$$

aiming to find its minima, which correspond to an upper bound on the sought worst-case term. Differentiating g yields the extrema

$$g'(x) = \frac{r^2 c_1}{(n - x)^2} - \frac{1}{c}S^2 d c_2 \stackrel{!}{=} 0$$
$$\implies x_{+/-} = n \pm \frac{r}{S}\sqrt{\frac{c_1 c}{c_2 d}} =: n \pm \alpha(r) \tag{13}$$

We calculate the second derivatives to be $g''(x_-) > 0$ and $g''(x_+) < 0$, thus only x_- is a minimum. Now, there are two cases associated with the location of x_- (taking into account that $\alpha(r) > 0$ for every $r > 0$).

- **Case I**: $\alpha(r) \leq n$.
  Then, x_- is a valid minimum in the considered range and therefore

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right|>r\right)&\\ \leq \sum_{f\in\mathcal{F}}\sum_{I\in\mathcal{P}([n])\setminus[n]} 2\exp\left(-\frac{r^2c_1}{\alpha(r)}\right)\exp\left(-\frac{\alpha(r)dS^2c_2}{c}\right)&\\ \leq 2|\mathcal{F}|2^n\exp\left(-2rS\sqrt{\frac{dc_2c_1}{c}}\right) := \mathbb{P}_{(I)}(r).&\end{aligned}$$

- **Case II**: $\alpha(r) > n$.
  Then, $x_- < 0$ is outside of the domain of g. However, the derivative satisfies $g'(x) > 0$ for any $0 \leq x < n$ since $x_+ > n$. Therefore, g necessarily takes its minimal value at $x = 0$ so that

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F_f)(x_i)\right|>r\right)&\\ \leq \sum_{f\in\mathcal{F}}\sum_{I\in\mathcal{P}([n])\setminus[n]} 2\exp\left(-\frac{r^2c_1}{n}\right)\exp\left(-\frac{ndS^2c_2}{c}\right)&\\ \leq 2|\mathcal{F}|2^n\exp\left(-\frac{r^2c_1}{n}\right)\exp\left(-\frac{ndS^2c_2}{c}\right) =: \mathbb{P}_{(II)}(r).&\end{aligned}$$

Using (13), condition $\alpha(r) > n$ is equivalent to $r > nS\sqrt{\frac{c_2d}{c_1c}}$. In this range, we have $\mathbb{P}_{(II)}(r) \leq \mathbb{P}_{(I)}(r)$ since

$$\mathbb{P}_{(II)}\left(nS\sqrt{\frac{c_2d}{c_1c}}\right) = 2|\mathcal{F}|2^n\exp\left(-2nS^2dc^{-1}c_2\right) = \mathbb{P}_{(I)}\left(nS\sqrt{\frac{c_2d}{c_1c}}\right)$$

and one verifies that $\mathbb{P}_{(II)}(r)$ decays faster than $\mathbb{P}_{(I)}(r)$ when further increasing r. Therefore, we conclude that for all $r > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F)(x_i)\right|>r\right) \leq \mathbb{P}_{(I)}(r) = 2|\mathcal{F}|2^n\exp\left(-2rS\sqrt{\frac{dc_2c_1}{c}}\right). \tag{14}$$

Further rewriting the expression, distinguishing between two cases with respect to the magnitude of $|\mathcal{F}|2^n$ yields the upper bounds:

- **Case 1**: $|\mathcal{F}|2^n \leq \exp\left(rS\sqrt{\frac{dc_2c_1}{c}}\right)$.
  We immediately obtain via (14) that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right| > r\right) \leq 2|\mathcal{F}|2^n \exp\left(-2rS\sqrt{\frac{dc_2c_1}{c}}\right)$$
$$\leq 2\exp\left(-rS\sqrt{\frac{dc_2c_1}{c}}\right)$$
$$\leq 2\exp\left(-\underbrace{\frac{2}{3\log(|\mathcal{F}|2^n)}}_{<1} rS\sqrt{\frac{dc_2c_1}{c}}\right).$$

- **Case 2**: $|\mathcal{F}|2^n > \exp\left(rS\sqrt{\frac{dc_2c_1}{c}}\right)$.
  In this case, the probability is trivially bounded by

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right| > r\right) \leq 1 < 2\exp\left(-\frac{2}{3}\right) < 2\exp\left(-\frac{2}{3}\underbrace{\frac{rS\sqrt{\frac{dc_2c_1}{c}}}{\log(|\mathcal{F}|2^n)}}_{<1}\right)$$

Putting both cases together, we proved that for all $r > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right| > r\right) \leq 2\exp\left(-\frac{2S\sqrt{\frac{dc_2c_1}{c}}}{3\log(|\mathcal{F}|2^n)}r\right).$$

This tail bound shows that $\sup_{f\in\mathcal{F}}|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)|$ is sub-exponential. Since the expected value of any sub-exponential random variable is up to an absolute constant given by its sub-exponential norm, which corresponds (up to a constant) to the parameter $\frac{3\log(|\mathcal{F}|2^n)}{2S\sqrt{\frac{dc_2c_1}{c}}}$ in the tail bound [23], we obtain for a constant $C_3 > 0$ that

$$\frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right|\right] \leq C_3\frac{1}{S}\left(\frac{\log|\mathcal{F}| + n\log 2}{n\sqrt{\frac{d}{c}}}\right)$$

Finally, the desired bound on the Rademacher complexity follows via (11):

$$\mathcal{R}_{n,\mu}(\mathcal{F}) = \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right]$$
$$\leq C_1\frac{1}{\sqrt{n}} + C_2\frac{1}{S}\sqrt{\frac{c\,\log|\mathcal{F}|}{nd}} + C_3\frac{1}{S}\frac{\sqrt{c}\log|\mathcal{F}|}{n\sqrt{d}} + C_3\frac{1}{S}\sqrt{\frac{c}{d}},$$

which, with the additional assumption $\log|\mathcal{F}| \geq n$, gives the result. ■

B.1. Comparison to standard bound without accounting for stability

Note that the crucial expectation in the derivation, i.e., the last term in (11), can be treated without linking it to the minimum class stability. Indeed, the expectation of the maximum of N subgaussians $X_1, \dots, X_N$ with variance proxy σ^2 scales as

$$\mathbb{E}\left[\max_{1\le i\le N} |X_i|\right] \le \sigma\sqrt{2\log(2N)},$$

see for instance [20]. Hence, in our case, as $\sigma_i(f - F_f)(x_i)$ is subgaussian with variance proxy 4 and therefore $\sum_{i=1}^n \sigma_i(f - F_f)(x_i)$ is subgaussian with variance proxy $4n$, we obtain

$$\frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i(f - F_f)(x_i)\right|\right] \le \frac{1}{n}2\sqrt{n}\sqrt{2\log(2|\mathcal{F}|)} \le C_4\left(\sqrt{\frac{1}{n}} + \sqrt{\frac{\log|\mathcal{F}|}{n}}\right).$$

for some absolute constant $C_4 > 0$. Neglecting the constants, this leads to the following comparison to our bound in (4):

$$\frac{1}{S}\frac{\sqrt{c}\log|\mathcal{F}|}{n\sqrt{d}} \le \sqrt{\frac{\log|\mathcal{F}|}{n}} \quad \iff \quad S \ge \sqrt{\frac{c\log|\mathcal{F}|}{nd}}.$$

Thus, our result tightens the bound on the Rademacher complexity of the function class under the isoperimentry condition provided that S is at least of the order $\sqrt{\frac{cp}{nd}}$, i.e., in the relevant overparameterized range $n \le p = \log|\mathcal{F}| \le n\sqrt{d}$ we only need S of the order $\sqrt{\frac{c}{d}}$.

Appendix C. Appendix: Experimental Details for Stability Measurement

Training setup. To empirically validate our robustness law, we trained fully connected MLPs on MNIST and CIFAR-10 datasets. Each model has 4 hidden layers with widths $w \in \{128, 256, 512, 1024, 2048\}$. All models use ReLU activations, batch normalization, and were initialized with standard parametrization. Training was conducted using the Adam optimizer [12] for the embedding and output layers, and the Muon optimizer [10] for the hidden layers. Models were trained with a batch size of 256 and learning rate 10^{-3}, until at least 99.9% training accuracy was achieved, ensuring (near) interpolation.

Parameter counts and normalization. For each model, we recorded the total number of trainable parameters p, input dimension d, and total number of training samples n. The theoretical bound $p/(n\sqrt{d})$ was used as a reference scale for comparison with the measured stability.

Stability estimation. Class stability $S(f)$ was computed using adversarial perturbation analysis. We performed a suite of ℓ_2-based attacks (FGSM, PGD, DeepFool, and L2PGD [9, 15, 17]) using the Foolbox library [19]. For each input x, we recorded the minimum perturbation norm required to change the classifier's prediction, over a grid of radii $\mathbf{r} = (0.002, 0.01, 0.05, 0.1)$. The final stability score $S(f)$ was taken as the average ℓ_2 distance across the dataset.

Implementation. Training and evaluation code is implemented in PyTorch [18]. For MLPs, images were flattened to vectors. Attack evaluations were conducted over the full dataset (train and test).

Reproducibility. All experiments were run with multiple random seeds $\{0, 1, 2, 3, 4\}$, and mean with standard deviation are reported.