THE PRICE OF ROBUSTNESS: STABLE CLASSIFIERS NEED OVERPARAMETERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The relationship between overparameterization, stability, and generalization remains incompletely understood in the setting of discontinuous classifiers. We address this gap by establishing a generalization bound for finite function classes that improves inversely with *class stability*, defined as the expected distance to the decision boundary in the input domain (margin). Interpreting class stability as a quantifiable notion of robustness, we derive as a corollary a *law of robustness for classification* that extends the results of Bubeck and Sellke beyond smoothness assumptions to discontinuous functions. In particular, any interpolating model with $p \approx n$ parameters on n data points must be *unstable*, implying that substantial overparameterization is necessary to achieve high stability. We obtain analogous results for (parameterized) infinite function classes by analyzing a stronger robustness measure derived from the margin in the codomain, which we refer to as the *normalized co-stability*. Experiments support our theory: stability increases with model size and correlates with test performance, while traditional norm-based measures remain largely 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 (Sain, 1996) and Rademacher complexity (Bartlett & Mendelson, 2002). However, these approaches do not adequately explain several empirical phenomena, e.g., *double descent* (Belkin et al., 2019) and *benign overfitting* (Bartlett et al., 2020). The occurrence of double descent illustrates that the test error, after initially increasing near the interpolation threshold, can improve as 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. Such findings expose the limitations of norm- and size-based complexity measures as predictors of generalization.

A large-scale empirical study evaluating more than forty complexity measures found that many norm-based quantities not only fail to correlate with generalization, but often even correlate negatively (Jiang et al., 2019). Beyond optimization-related metrics, one of the few quantities that consistently correlated with generalization was the margin, i.e., the distance to the decision boundary, closely related to the notion of (co-)stability we develop in this work. This aligns with an emerging perspective: generalization in modern networks is governed less by model size or norms, and more by the *stability* / *robustness* of predictions under input perturbations (Soloff et al., 2025; Ghosh & Belkin, 2023; Zhang et al., 2022). Related insights also arise from the literature on algorithmic stability (Bousquet & Elisseeff, 2002) and flat minima (Keskar et al., 2017). However, most theoretical results in this direction are restricted to linear models.

An exception is the *universal law of robustness* of Bubeck & Sellke (2021), which, under mild distributional assumptions, establishes a formal link between robustness, generalization, and overparameterization: smoothness and overparameterization need to balance in order to ensure good generalization while overfitting. The *law of robustness* relies on the assumption that the function class is Lipschitz, which makes it inadequate for classifiers whose codomain is discrete by design. We therefore take a step toward the open challenge posed in Bubeck & Sellke (2021, p. 4): "[...]

it is an interesting challenge to understand for which notions of smoothness there is a tradeoff with size." Specifically, we introduce *class stability* and *normalized co-stability* as geometric smoothness measures that extend robustness laws to classification. In fact, replacing Lipschitz continuity is essential: simply focusing on the Lipschitz constant of an underlying score function g, where the classifier is of type $f := \arg\max \circ g$, is not informative. In particular, since g can be arbitrarily rescaled without changing the predictions of f, its Lipschitz constant does not need to reflect the geometry of the decision boundary (Liu & Hansen, 2024).

Paper Roadmap. We discuss related work in Section 2. Section 3 introduces class stability and the isoperimetry assumption, a concentration property of the data that underlies our analysis. Section 4 presents a generalization bound for finite hypothesis classes and examines its implications for overparameterization. In Section 5, we extend the framework to infinite function classes via the notion of normalized co-stability. Our theoretical predictions are tested experimentally on MNIST and CIFAR-10 in Section 6. Finally, Section 7 concludes with a discussion of open directions.

Contributions. We provide a summary of our main results.

1) We prove that, under an isoperimetry assumption on the data distribution, the data-dependent Rademacher complexity of a finite hypothesis class of classifiers can be bounded in terms of the minimum *class stability*. This yields an improved generalization bound for discontinuous classifiers (Theorem 4),which tightens as stability increases.

2) We show that in the classically parameterized regime (#parameters $\approx$ #samples), any interpolating classifier must be unstable (Corollary 6) with high probability. Consequently, achieving both near-perfect fitting and high class stability requires substantial overparameterization of order $p \approx nd$.

3) We extend the framework to infinite function classes by considering classifiers of the form $f(x) := \arg\max \circ g_w(x)$, where g_w is a parameterized Lipschitz-continuous (in both x and w) score function. This enables us to define a robustness measure – the *normalized co-stability* –, based on output score margins, and derive a corresponding generalization bound (Theorem 13). The added regularity also results in a law of robustness for infinite function classes (Corollary 15).

4) We empirically validate our predictions on MNIST and CIFAR-10, observing that stability and normalized co-stability grow with network width and closely track test error, supporting our claim that generalization in overparameterized regimes is driven by (normalized co-)stability.

Taken together, our results extend the law of robustness to discontinuous classifiers and highlight stability as a central factor in understanding generalization in modern networks.

2 Related Work

Smoothness-based generalization. Our work is inspired by the *law of robustness* of Bubeck & Sellke (2021), which shows that regression with Lipschitz predictors generalizes when smoothness and overparameterization are properly balanced. Subsequent works have extended this perspective: for example, Zhu et al. (2023) investigate how width, depth, and initialization affect robustness, while more recent studies Das et al. (2025) establish refined smoothness–generalization trade-offs for a wider range of loss landscapes.

Margin-based generalization. Classical generalization bounds combine a margin term, defined with respect to a score function, with a capacity measure – for example, spectrally-normalized margin bounds (Bartlett et al., 2017) or path-norm bounds (Neyshabur et al., 2018). Recent extensions include multi-class margin bounds in terms of margin-normalized geometric complexity (Munn et al., 2024). These approaches are closely aligned with our normalized co-stability perspective: both control a codomain margin while coupling it to a regularity property of the score function, and both recover inverse-margin scaling.

Input-space margin bounds have also been studied, yielding that generalization is controlled by the minimum robustness radius (Sokolic et al., 2017), while sample-complexity lower bounds show that adversarial robustness increases the VC dimension (Gao et al., 2019). Our notion of *class stability* differs: it is the *expected input margin* – the average distance to the decision boundary under the

data distribution – rather than a minimum or an empirical quantile. This measure is closely tied to robustness (Fawzi et al., 2016; Gilmer et al., 2018) and induces data-dependent bounds that track generalization.

Limits of uniform generalization bounds. Uniform convergence–based bounds are often vacuous in overparameterized networks (Nagarajan & Kolter, 2021), since SGD appears to find solutions at a macroscopic level (supporting generalization) but with microscopic fluctuations that break uniform analyses. Our bounds remain uniform but depend on macroscopic, distribution-dependent quantities: the Rademacher complexity—our applied technique to derive generalization bounds—is controlled by stability (or co-stability). Whether this structure avoids the vacuity identified by Nagarajan & Kolter (2021) remains open.

Stability, robustness, and implicit bias. Algorithmic stability (Bousquet & Elisseeff, 2002) and the flat minima literature (Keskar et al., 2017) argue that robustness under perturbations drives generalization. More recently, Zou et al. (2024) derive out-of-distribution generalization bounds based on the sharpness of the learned minima. Our contribution is to extend a stability-based perspective to discontinuous neural classifiers, both theoretically and empirically. Complementary work on implicit bias shows that gradient descent favors solutions with a small number of connected decision regions, a proxy for large input-space margin (Li et al., 2025). This suggests that optimization dynamics may implicitly favor the same geometric simplicity that our stability-based bounds capture.

Out-of-Distribution Generalization. Classically and also in our analysis, generalization is based on the independently and identically distributed assumption on the data, in particular, the test data are generated from the same distribution as the training data coined In-Distribution (ID) generalization. In contrast, Out-of-Distribution (OOD) generalization aims to study the generalization performance under distributional shifts. To make the problem tractable the potential shifts are constrained to, for instance, spurious correlations or covariate shifts. In the OOD setting the connection between overparameterization and generalization has been studied in a series of theoretical works with positive Hao et al. (2024) and negative results Sagawa et al. (2020); Wald et al. (2024).

Adversarial robustness can be viewed as a special case of OOD generalization, where the distributional shift is constrained to lie within a perturbation set Sinha et al. (2020). In this sense, our stability-based analysis is conceptually connected to OOD generalization. However, our results do not provide explicit bounds on OOD error; instead, we focus on ID generalization under the assumption that the classifier satisfies a given level of adversarial robustness expressed as the margin.

3 Preliminaries and Notation

In the following, we provide background on the key concepts underlying our analysis, namely stability, generalization, and isoperimetry. For clarity of exposition, we present our results in the binary classification setting. The extension to multi-class problems follows by a one-vs-all reduction; see Appendix F for details. Thus, 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 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 Liu & Hansen (2024), 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 (Margin and Class Stability)**.** *Let $f : \mathcal{X} \to \{-1, 1\}$. The* signed distance function d_f *of f at $x \in \mathcal{X}$ is defined as*

$$d_f(x) := \begin{cases} d(x, f^{-1}(\{-1\})), & \text{if } f(x) = 1, \\ -d(x, f^{-1}(\{1\})), & \text{if } f(x) = -1, \end{cases}$$

where $d(x, A) := \inf_{y \in A} \|x - y\|_2$. *We define the (unsigned)* margin h_f *at* x *as the absolute value of the signed distance function,*

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

The class stability $S(f)$ *of* f *is its expected margin under the data distribution:*

$$S(f) := \mathbb{E}[h_f].$$

Remark 2. *The signed distance function* d_f *is* 1*-Lipschitz if* $\mathcal{X}$ *is path-connected. Moreover, if* $\operatorname{sgn}(0) = 1$ *and* $f^{-1}(\{1\})$ *is closed in* $\mathcal{X}$*, then* f *admits the representation* $f = \operatorname{sgn} \circ d_f$ *(see Appendix B for details).*

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 (Bartlett & Mendelson, 2002). 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 Equation 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, 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 Bartlett & Mendelson (2002); Shalev-Shwartz & Ben-David (2014) for detailed explanations. Overall, it therefore suffices to bound $\mathcal{R}_{n,\mu}(\mathcal{F})$ in terms of the class stability of functions $f \in \mathcal{F}$ in order to link generalization to stability. Equivalently, the key step is to control how well stable functions can fit random labels, which requires structural assumptions on the input distribution. We discuss in detail in Appendix A why such assumptions are unavoidable. A natural and widely used condition is *isoperimetry*, which guarantees sharp concentration for bounded Lipschitz-continuous functions (Bubeck & Sellke, 2021).

Definition 3 (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 (Vershynin, 2018; Bubeck & Sellke, 2021). Consequently, under the manifold hypothesis, the relevant dimension in our bounds can be interpreted as the intrinsic manifold dimension rather than the ambient dimension.

4 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 follow the general strategy of Bubeck & Sellke (2021) (see Appendix A for details), but replace their use of Lipschitz continuity with our notion of *class stability* (Definition 1). Formally, we proceed under the following assumptions:

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

[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 $|\mathcal{F}| < \infty$.

These conditions ensure concentration of measure in the input space and allow complexity control via a union bound. With this structure in place, class stability can be related to the Rademacher complexity, leading to the bound stated below.

Theorem 4 (Rademacher Bound)**.** *Suppose Assumptions (H1) and (H2) hold, and that* $\min_{f\in\mathcal{F}} S(f) > S > 0$ *with* $\log|\mathcal{F}| \geq n$.

1. *The Rademacher complexity satisfies*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \;\leq\; K_1 \max\left\{\frac{1}{\sqrt{n}},\; \frac{\sqrt{c}}{S}\cdot\frac{\log|\mathcal{F}|}{n\sqrt{d}}\right\}, \tag{4}$$

   for an absolute constant $K_1 > 0$.

2. *If, in addition,* $f^{-1}(\{1\})$ *is closed and* $\mathcal{X}$ *path-connected, the bound sharpens to*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \;\leq\; K_2 \max\left\{\frac{1}{\sqrt{n}},\; \frac{\sqrt{c}}{S}\sqrt{\frac{\log|\mathcal{F}|}{nd}},\; 2\exp\Big(-\frac{dS^2}{8c}\Big)\right\}, \tag{5}$$

   for another absolute constant $K_2 > 0$.

Proof sketch. Equation 4 is obtained via a Lipschitz surrogate argument combined with isoperimetry. The refined bound in Equation 5 further leverages the representation $f = \operatorname{sgn}\circ d_f$ (Remark 2), using that large stability ensures d_f remains well separated from the discontinuity at 0. Complete details are provided in Appendix C. □

Remark 5. *In contrast to Bubeck & Sellke (2021), where stability is measured by the minimal Lipschitz constant of the function class, our initial bound in Theorem 4 incurred an additional factor* $\sqrt{\log|\mathcal{F}|/n}$ *in the regime* $\log|\mathcal{F}| \geq n$. *By assuming mild regularity conditions, we can eliminate this gap and recover the same scaling as in Bubeck & Sellke (2021).*

The key insight of Theorem 4, combined with the classical generalization bound in Equation 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 front of $\sqrt{\log|\mathcal{F}|}$ in Equation 4 and Equation 5 indicates that class stability affects the effective complexity of the model class, potentially 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 Equation 5 dominate, and the bounds becomes small if S scales at least in the order of $\sqrt{\frac{p}{nd}}$.

We are now ready to state our *law of robustness for discontinuous functions*, obtained as a direct corollary of the refined Rademacher bound in Equation 5 of Theorem 4.

Corollary 6 (Law of Robustness for Discontinuous Functions)**.** *Assume (H1), (H2), and the additional conditions in 2. of Theorem 4 hold. Let* $p := \log|\mathcal{F}| \geq n$. *Fix* $\varepsilon, \delta \in (0, 1)$ *and consider the* 0–1 *loss* $\ell_{0\text{–}1}$. *There exists an absolute constant* $K > 0$ *such that, if*

1. *the minimal risk* $\sigma^2 := \min_{f\in\mathcal{F}} R_{0\text{–}1}(f)$ *satisfies* $\sigma^2 \geq \varepsilon$, *and*

2. *the sample size* n *is large enough to ensure (i)* $\frac{K}{\sqrt{n}} < \frac{\varepsilon}{3}$ *and (ii)* $\sqrt{\frac{2\log(2/\delta)}{n}} < \frac{\varepsilon}{2}$,

then with probability at least $1 - \delta$ *(over the sample), the following holds uniformly for all* $f \in \mathcal{F}$*:*

$$\hat{R}_{0\text{–}1}(f) \leq \sigma^2 - \varepsilon \quad\Longrightarrow\quad S(f) < \max\left\{\frac{3K}{\varepsilon}\sqrt{\frac{c\log|\mathcal{F}|}{nd}},\; \sqrt{\frac{8c}{d}\log\Big(\frac{6K}{\varepsilon}\Big)}\right\}. \tag{6}$$

Proof sketch. Apply the Rademacher bound (Theorem 4) to the high-stability subset $\mathcal{F}_{S_*} := \{f \in \mathcal{F} : S(f) \geq S_*\}$. For S_* chosen large enough, such functions cannot achieve empirical risk below $\sigma^2 - \varepsilon$, so any interpolating classifier with risk $\leq \sigma^2 - \varepsilon$ must lie outside $\mathcal{F}_{S_*}$, i.e., must satisfy $S(f) < S_*$. The full proof is provided in Appendix D. □

Remark 7. *Unlike Bubeck & Sellke (2021), which assume Lipschitz-continuous losses, our analysis directly addresses the discontinuous* 0–1 *loss, making it more natural for classification tasks. The overall proof strategy, however, extends to arbitrary losses provided one can derive an appropriate bound on the Rademacher complexity of the composed function class, as in Equation 2.*

Remark 8. *Importantly, this result also covers intrinsically discontinuous classifiers, such as quantized neural networks and spiking neural networks. Moreover, since self-attention is in general not Lipschitz-continuous Kim et al. (2021), our framework appears particularly well-suited to the analysis of overparameterization of transformers, which underlie most state-of-the-art language models.*

From Equation 6 we conclude that achieving both low training error and high stability requires parameterization on the order $p \approx nd$. This necessity arises in the high-dimensional regime, since when d is large the first term in the maximum dominates for $p \geq n$. 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) (Brown et al., 2020), are trained in heavily overparameterized regimes: Even though recent scaling laws Hoffmann et al. (2022) suggest a balance between model and data size, these models remain functionally overparameterized in that their capacity far exceeds what is required to fit the training data. Therefore, our result may help to understand why such models still do generalize effectively.

5 A Law of Robustness for Infinite Function Classes

In Theorem 4, our analysis does not straightforwardly extend to infinite function classes. The usual proof strategy via a covering-number argument requires closeness in parameter space to imply closeness in function space. In Bubeck & Sellke (2021), this is enforced via Lipschitz continuity in the parameters of the function class, but such a condition is in general meaningless for discontinuous classifiers.

To overcome this, we restrict our attention to function classes with additional structure and introduce a strengthened stability notion. Specifically, we impose a representation analogous to Remark 2, namely,

(H3) The hypothesis class has the form $\mathcal{F} = \operatorname{sgn} \circ \mathcal{G}$, where $\mathcal{G} = \{g_w : \mathcal{X} \to [-1, 1] : w \in \mathcal{W}\}$ is a parameterized family of Lipschitz functions. The parameter space $\mathcal{W} \subset \mathbb{R}^p$ is bounded with $\operatorname{diam}(\mathcal{W}) \leq W$, and the parameterization is Lipschitz:

$$\|g_{w_1} - g_{w_2}\|_\infty \;\leq\; J\,\|w_1 - w_2\|.$$

The extension from finite to infinite classes requires not only (i) Lipschitz continuity in w, but also (ii) that the scores $g_w(x)$ stay quantitatively away from zero, so that small parameter perturbations cannot cause arbitrary label flips. Class stability alone does not suffice to ensure (ii), as the following example demonstrates.

Example 9 (Class stability does not prevent discontinuity)**.** *Let* $\mathcal{G} = \{g_w(x) = w \tanh(x) : w \in [-1, 1]\}$. *The parameterization is Lipschitz since*

$$\|g_{w_1} - g_{w_2}\| \;\leq\; \|w_1 - w_2\|.$$

For $w_1 = \frac{\varepsilon}{2}$ *and* $w_2 = -w_1$, $\|w_1 - w_2\| \leq \varepsilon$, *yet*

$$\|\operatorname{sgn}(g_{w_1}(x)) - \operatorname{sgn}(g_{w_2}(x))\| = 2$$

for almost all x. *Each classifier has a single boundary (hence high class stability), but parameter proximity does not imply classifier proximity.*

To guarantee property (ii), we introduce a new robustness measure in the codomain.

Definition 10 (Co-margin and Co-stability). *Let $f = \operatorname{sgn} \circ g : \mathcal{X} \to \{-1, 1\}$. The* co-margin *at x is*

$$h_g^*(x) := |g(x)|,$$

and we denote the normalized co-margin *as*

$$\bar{h}_g^*(x) := \frac{|g(x)|}{L(g)},$$

where $L(g)$ is the Lipschitz constant of g. The co-stability *is then the expected co-margin*

$$S^*(g) := \mathbb{E}[h_g^*(x)],$$

and the normalized co-stability *is accordingly defined as the expected normalized co-margin*

$$\bar{S}^*(g) := \mathbb{E}[\bar{h}_g^*(x)].$$

Remark 11 (Representation dependence). *Unlike class stability $S(f)$, which depends only on the decision boundary of f, the co-stability $S^*(g)$ and its normalized form $\bar{S}^*(g)$ depend on the particular representation $f = \operatorname{sgn} \circ g$. Different score functions g inducing the same classifier f can yield different values of $S^*(g)$ and $\bar{S}^*(g)$. For the specific representation $f = \operatorname{sgn} \circ d_f$ from Lemma 18, however, the quantities coincide: $S^*(g) = \bar{S}^*(g) = S(f)$.*

Imposing $S^*(g) \geq S^* > 0$ ensures that scores remain, on average, a non-trivial distance away from zero. Together with (H3), co-stability provides the continuity and separation properties required for infinite-class generalization bounds.

Before turning to the formal statement of this fact, we want to discuss the relation of class stability and co-stability. The connection between input- and codomain-based margins is immediate since

$$h_g(x) \;\geq\; \frac{h_g^*(x)}{L(g)} \;=\; \bar{h}_g^*(x).$$

By $L(g)$-Lipschitz continuity, moving x by r changes $g(x)$ by at most $L(g)r$, so flipping the prediction requires $r \geq |g(x)|/L(g)$. Taking expectations yields

$$S(f) \;\geq\; \bar{S}^*(g). \tag{7}$$

Thus normalized co-stability lower-bounds class stability. This inequality highlights two levers for improving generalization: increasing $S^*(g)$ or decreasing $L(g)$. Importantly, $\bar{S}^*(g)$, like $S(f)$, is invariant to input rescaling and therefore serves as a natural robustness measure.

Remark 12. *A related ratio, $\frac{\gamma}{\mathcal{R}_f}$, appears in Bartlett et al. (2017), where γ is the minimum margin and $\mathcal{R}_f$ a spectral complexity term controlling Lipschitzness. Empirically, Lipschitz margin training, which enforces*

$$\bar{S}^*(g) \;\geq\; c,$$

improves adversarial robustness (Tsuzuku et al., 2018). Moreover, Béthune et al. (2022, Corollary 2) show that among maximally accurate classifiers, there exists a 1-Lipschitz solution that achieves maximal co-margins and satisfies $S(f) = S^*(g)$. In particular, the Bayes classifier admits the representation $b = \operatorname{sgn} \circ d_b$, which fulfills these properties.

Combining Theorem 4 with Equation 7, the Rademacher complexity of a finite function class $\mathcal{F} = \operatorname{sgn} \circ \mathcal{G}$ can be bounded in terms of normalized co-stability as

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \;\leq\; K_2 \max\left\{ \frac{1}{\sqrt{n}},\; \sqrt{c}\,\frac{L}{S^*}\sqrt{\frac{\log|\mathcal{F}|}{nd}},\; 2\exp\Big(-\frac{dS^{*2}}{L^2 8c}\Big) \right\},$$

where $S^* > 0$ and $L > 0$ are bounds on the minimal co-stability and maximal Lipschitz constant, respectively. Under condition (H3), the statement can be extended to infinite function classes.

Theorem 13. *Suppose (H1) and (H3) hold, and that $S^*(g) > S^* > 0$ and $L(g) \leq L$ for all $g \in \mathcal{G}$. Assume further that $p \geq n$. Then, for any covering precision $\tilde{\varepsilon} > 0$,*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \;\leq\; K \max\left\{ \sqrt{\frac{1}{n}},\; \frac{L}{S^*}\sqrt{\frac{p}{nd}}\sqrt{c\log\Big(1 + 60WJ\tilde{\varepsilon}^{-1}\Big)},\; 2\exp\Big(-\frac{dS^{*2}}{8cL^2}\Big),\; \frac{J}{S^*}\,\tilde{\varepsilon} \right\}, \tag{8}$$

where $K > 0$ is an absolute constant independent of $p, n, d, S^*, c, L, J, \tilde{\varepsilon}, W$.

Proof sketch. The proof follows the previously mentioned ε-net approach, standard in infinite-class settings. The Lipschitz continuity in w (from (H3)) controls the covering number of $\mathcal{G}$ at scale $\tilde{\varepsilon}$, while co-stability ensures that small perturbations in w do not induce flips through the sgn mapping. The additional term $\frac{J}{S^*}\tilde{\varepsilon}$ reflects the residual error introduced by the discretization. See Appendix E for more details. □

Remark 14. *The factor $\frac{L}{S^*}$ shows that generalization depends jointly on the average prediction confidence $S^*(g)$ and the Lipschitz constant $L(g)$, the latter quantifying robustness of predicted probabilities. This aligns with empirical findings (Khromov & Singh, 2024; Gamba et al., 2025; Gouk et al., 2020; Sanyal et al., 2020; Béthune et al., 2022), which report that smaller Lipschitz constants typically improve generalization, and in some cases exhibit a double-descent behavior.*

We obtain with the same reasoning as in Corollary 6 the following law of robustness for Lipschitz-regular infinite function classes.

Corollary 15 (Law of Robustness for Infinite Function Classes)**.** *Assume (H1) and (H3), and fix $\varepsilon, \delta \in (0,1)$. Consider the 0–1 loss $\ell_{0\text{–}1}$. There exists an absolute constant $K > 0$ such that, if*

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

then with probability at least $1-\delta$, for all $\tilde{\varepsilon} > 0$, the following holds uniformly for all $g \in \mathcal{G}$ and $f_g = sgn \circ g$:

$$\hat{R}_{0\text{–}1}(f_g) \leq \sigma^2 - \varepsilon \quad \Longrightarrow \quad \frac{S^*(g)}{L(g)} < \max\left\{ \frac{3K}{\varepsilon}\sqrt{\frac{p}{nd}}\sqrt{c\log(1+60WJ\tilde{\varepsilon}^{-1})},\ \sqrt{\frac{8c}{d}\log\left(\frac{6K}{\varepsilon}\right)} \right\}.$$

Remark 16. *As in Bubeck & Sellke (2021), we require W and J to be at most polynomial in (n, d, p) so that they do not affect the asymptotic scaling. In the case of feedforward neural networks, Bubeck & Sellke (2021) further show that when the data distribution is concentrated in a ball of radius R, it suffices to assume that W is polynomially bounded.*

Analogous to the finite-class case, we conclude that Lipschitz-regular classifiers must be overparameterized of order nd to achieve both low training error and high normalized co-stability. Without sufficient parameter capacity relative to sample size and ambient dimension, robustness cannot be guaranteed: models may fit the training data, but will necessarily exhibit either large Lipschitz constants of the score function or low co-stability, reflecting weak confidence in their predictions. Thus, overparameterization emerges as a necessary condition for robustness, not a byproduct of current training practice, but a structural limitation dictated by geometry and probability.

6 Experiments

We empirically validate our theoretical prediction that class stability $S(f)$ and co-stability $S(f)^*$ increase with model size in interpolating networks.

Setup. We train fully connected MLPs with four hidden layers and widths $w \in \{128, 256, 512, 1024, 2048\}$ on MNIST and up to $w = 1024$ for CIFAR-10. All models are trained until reaching at least 99% training accuracy, ensuring (near-)interpolation so that test accuracy effectively coincides with generalization performance.

Class Stability. We estimate empirical class stability $S(f)$ via adversarial perturbations. For each input, we increase the perturbation radius r along a predefined grid $\mathbf{r} = (r_1, \ldots, r_n)$ until the classifier's prediction changes. The minimal successful radius is recorded as the distance to the decision boundary for that sample, and $S(f)$ is reported as the average over the dataset.

Normalized Co-Stability. The empirical co-stability $S^*(g)$ is computed via the multi-class margin

$$g_j(x) - \max_{i\neq j} g_i(x), \qquad j = \arg\max_i g_i(x),$$

averaged over the dataset; see Appendix F for details about the multi-class setting. We estimate the Lipschitz constant $L(g)$ using the efficient ECLIPSE method (Xu & Sivaranjani, 2024), and report the normalized ratio $S^*(g)/L(g)$ as a function of model size.

Results. Figure 1 shows that, for MLPs, both class stability $S(f)$ and normalized co-stability $S^*(g)/L(g)$ increase consistently with model size. The observed saturation of (normalized co-) stability aligns with theoretical intuition: the Bayes classifier admits a finite (normalized co-) stability level, and pushing beyond this level necessarily reduces accuracy - an instance of the robustness/accuracy trade-off extensively discussed in the literature (Zhang et al., 2019; Tsipras et al., 2019; Béthune et al., 2022). Accordingly, we expect stability to plateau once models approach the Bayes decision boundary. For CIFAR-10, although test accuracy remains far below the Bayes optimal (around 50%), the same reasoning applies relative to the best classifier achievable within the restricted MLP architecture.

Empirically, class stability closely tracks test accuracy, whereas standard weight norms show no systematic correlation with model size or generalization performance. On MNIST, however, we observe that normalized co-stability exhibits large seed-to-seed fluctuations and no consistent trend with model size. We conjecture that this reflects the simplicity of MNIST, which admits many local minima with highly variable score functions. To probe this hypothesis, we train 4- MLPs width widths $w \in \{128, 256, 512, 1024\}$ using sharpness-aware optimization (SAM) (Foret et al., 2021; Kwon et al., 2021), which biases training toward flatter minima. As shown in Figure 2, this reduces variance across seeds and restores a clear monotonic dependence on model size. We note that the absolute values of stability are smaller for SAM-trained models, but this is explained by the absence of spectral normalization in SAM, which results in larger Lipschitz constants. What matters for our purposes is the monotonic trend, not the absolute scale. These findings suggest a quantitative link between sharpness and stability, and motivate further study of how optimization bias interacts with the geometric structure underlying our robustness laws.

Additional details and plots are provided in Appendix G. Moreover, our code is available here: `https://anonymous.4open.science/r/ICLR26-Stability-AC53/README.md`.

7 Discussion and Future Work

Our results identify class stability and its codomain analogue, normalized co-stability, as principled quantities linking overparameterization, generalization, and robustness for discontinuous classifiers. While we provide geometric laws of robustness for finite and infinite hypothesis classes, and our experiments support their validity, several directions remain open.

Empirical directions. Computing class stability $S(f)$ and Lipschitz constants $L(g)$ of neural networks is NP-hard (Katz et al., 2017; Weng et al., 2018; Scaman & Virmaux, 2019), limiting the

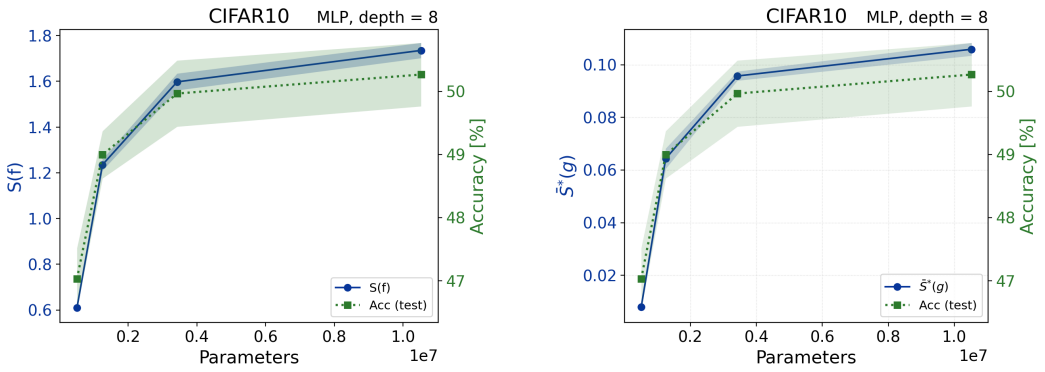

Figure 1: Stability measures for MLPs trained on CIFAR-10. Both class stability $S(f)$ and normalized co-stability $\bar{S}^*(g) = S^*(g)/L(g)$ increase systematically with model size and closely follow test accuracy, in line with our theoretical predictions.

direct use of (normalized co-)stability in training. However, practical relaxations exist: normalized co-stability underlies *Lipschitz margin training* (Tsuzuku et al., 2018), while input-space stability is related to adversarial training (Madry et al., 2018; Goodfellow et al., 2015). Biasing optimization explicitly toward (co-)stable solutions is therefore a promising empirical direction. Another avenue is to probe isoperimetry and related concentration phenomena on real data. This connects to the manifold hypothesis and raises the question of whether robustness laws fail empirically when the effective dimension of the data manifold is small.

Theoretical directions. Our framework motivates exploring alternative geometric measures, too. Do quantities such as sharpness of the loss landscape obey robustness laws analogous to those for (normalized co-)stability? Our experiments suggest a link, calling for deeper analysis. Another question concerns sufficiency: we establish that overparameterization is necessary for generalization, but is it also sufficient under suitable optimization? Bombari et al. (2023) prove sufficiency for Lipschitz regression in the NTK regime but show that it fails for a random features model. Extending such results to discontinuous classifiers may reveal qualitative differences.

Finally, the role of implicit bias remains unclear. Does gradient descent or SGD exhibit a bias toward classifiers with higher (normalized co-)stability, as suggested by analogous results on region counts (Li et al., 2025)? Establishing such a bias would explain why stable solutions emerge in practice.

Overall, our findings suggest that stability-based laws capture a core structural constraint of modern overparameterized learning. Developing efficient estimators, stronger empirical validation, and deeper theoretical connections (e.g., with sharpness and optimization bias) are promising next steps toward a unified understanding of generalization and robustness.

Ethics Statement

This work focuses on the theoretical analysis of generalization in machine learning and does not involve experiments on human subjects, sensitive personal data, or applications with direct societal risks. The datasets referenced are publicly available, and no private or restricted data was used. Potential ethical concerns related to misuse are minimal, as the contributions are mainly theoretical and methodological.

Acknowledgment of LLM Use. We explicitly acknowledge that large language models (LLMs) were used solely for polishing code, improving sentence clarity, and refining grammar. They were not used for generating research ideas, proofs, or results.

Reproducibility Statement

We have taken multiple steps to ensure reproducibility of our results. All theoretical claims are accompanied by rigorous proofs, presented in detail in the appendix. Assumptions underlying the theorems are explicitly stated, and definitions are given in full to allow independent verification. In addition, we provide open-source code to reproduce illustrative experiments and examples, which is available anonymously at anonymous GitHub.

References

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks, 2017. URL https://arxiv.org/abs/1706.08498.

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.

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.

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.

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.

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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

Louis Béthune, Thibaut Boissin, Mathieu Serrurier, Franck Mamalet, Corentin Friedrich, and Alberto González-Sanz. Pay attention to your loss: understanding misconceptions about 1-lipschitz neural networks, 2022. URL https://arxiv.org/abs/2104.05097.

Santanu Das, Jatin Batra, and Piyush Srivastava. A direct proof of a unified law of robustness for bregman divergence losses, 2025. URL https://arxiv.org/abs/2405.16639.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations, 2016. URL https://arxiv.org/abs/1502.02590.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Matteo Gamba, Hossein Azizpour, and Mårten Björkman. On the lipschitz constant of deep networks and double descent, 2025. URL https://arxiv.org/abs/2301.12309.

Ruiqi Gao, Tianle Cai, Haochuan Li, Liwei Wang, Cho-Jui Hsieh, and Jason D. Lee. Convergence of adversarial training in overparametrized neural networks, 2019. URL https://arxiv.org/abs/1906.07916.

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.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres, 2018. URL https://arxiv.org/abs/1801.02774.

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.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing lipschitz continuity, 2020. URL https://arxiv.org/abs/1804.04368.

Yifan Hao, Yong Lin, Difan Zou, and Tong Zhang. On the benefits of over-parameterization for out-of-distribution generalization, 2024. URL https://arxiv.org/abs/2403.17592.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them, 2019. URL https://arxiv.org/abs/1912.02178.

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/.

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks, 2017. URL https://arxiv.org/abs/1702.01135.

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.

Grigory Khromov and Sidak Pal Singh. Some fundamental aspects about lipschitz continuity of neural networks, 2024. URL https://arxiv.org/abs/2302.10886.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention, 2021. URL https://arxiv.org/abs/2006.04710.

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.

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

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kwon21b.html.

Jingwei Li, Jing Xu, Zifan Wang, Huishuai Zhang, and Jingzhao Zhang. Understanding nonlinear implicit bias via region counts in input space, 2025. URL https://arxiv.org/abs/2505.11370.

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.

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.

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.

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)*, pp. 2574–2582, 2016. doi: 10.1109/CVPR.2016.282.

Michael Munn, Benoit Dherin, and Javier Gonzalvo. A margin-based multiclass generalization bound via geometric complexity, 2024. URL https://arxiv.org/abs/2405.18590.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning, 2021. URL https://arxiv.org/abs/1902.04742.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks, 2018. URL https://arxiv.org/abs/1805.12076.

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*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

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.

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

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Stephan R Sain. The nature of statistical learning theory, 1996.

Amartya Sanyal, Philip H. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1enKkrFDB.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation, 2019. URL https://arxiv.org/abs/1805.10965.

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

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training, 2020. URL https://arxiv.org/abs/1710.10571.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, August 2017. ISSN 1941-0476. doi: 10.1109/tsp.2017.2708039. URL http://dx.doi.org/10.1109/TSP.2017.2708039.

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.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019. URL https://arxiv.org/abs/1805.12152.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks, 2018. URL https://arxiv.org/abs/1802.04034.

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

Yoav Wald, Gal Yona, Uri Shalit, and Yair Carmon. Malign overfitting: Interpolation can provably preclude invariance, 2024. URL https://arxiv.org/abs/2211.15724.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S. Dhillon, and Luca Daniel. Towards fast computation of certified robustness for relu networks, 2018. URL https://arxiv.org/abs/1804.09699.

Yuezhu Xu and S. Sivaranjani. Eclipse: Efficient compositional lipschitz constant estimation for deep neural networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10414–10441. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1419d8554191a65ea4f2d8e1057973e4-Paper-Conference.pdf.

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.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019. URL https://arxiv.org/abs/1901.08573.

Zhenyu Zhu, Fanghui Liu, Grigorios G Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization), 2023. URL https://arxiv.org/abs/2209.07263.

Yingtian Zou, Kenji Kawaguchi, Yingnan Liu, Jiashuo Liu, Mong-Li Lee, and Wynne Hsu. Towards robust out-of-distribution generalization bounds via sharpness, 2024. URL https://arxiv.org/abs/2403.06392.

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 & Sellke (Bubeck & Sellke, 2021) 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

$$\mathbb{P}(\exists f \in \mathcal{F} : \hat{R}_\ell(f) \approx 0 \quad \wedge \quad L(f) \leq L_*) < \delta. \tag{9}$$

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 $f \in \mathcal{F}$:

$$\hat{R}_\ell(f) \approx 0 \implies L(f) > L_*(p, n, d). \tag{10}$$

Here, $L_*(p, n, d)$ is an algebraic function of the number of parameters $p \approx \log |\mathcal{F}|$ (see the paragraph below Theorem 4 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 Bubeck & Sellke (2021) for proving (a variant of) Equation 9 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 3, 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 Equation 9, 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 Equation 10 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 $f : \mathcal{X} \to [-1, 1]$:

$$\mathbb{P}(|f(x) - \mathbb{E}(f)| \geq t) \leq 2 \exp\left(-\frac{t^2}{2}\right) \quad \forall t > 0. \tag{11}$$

Thus, to ensure that the probability in Equation 9 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 Equation 11) in the regime $L(f) \gtrsim 1$. This is necessary to yield a non-vacuous bound in Equation 10, 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 $\exp(-\frac{2t^2}{\text{diam}(\mathcal{X})^2 L(f)^2})$, which is insufficient as it decays faster than the Hoeffding bound only for $L(f) < 2/\text{diam}(\mathcal{X})$, i.e., at least $\text{diam}(\mathcal{X}) < 2$ 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 in Equation 3 leads to a faster decay than the Hoeffding bound in Equation 11 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.

B THE SIGNED DISTANCE FUNCTION (REMARK 2)

We collect the main properties of the signed distance function

$$d_f(x) := \begin{cases} d(x, f^{-1}(\{-1\})), & \text{if } f(x) = 1, \\ -d(x, f^{-1}(\{1\})), & \text{if } f(x) = -1, \end{cases}$$

where $d(x, A) := \inf_{y \in A} \|x - y\|_2$.

Lemma 17. *Let $\mathcal{X} \subset \mathbb{R}^d$ be bounded and path-connected, and let $f : \mathcal{X} \to \{-1, 1\}$. Then the signed distance function d_f is 1-Lipschitz.*

This is a classical fact, a special case of the Eikonal equation. For completeness, we include a direct proof inspired by Liu & Hansen (2024, Prop. 7.5).

Proof. **Case 1: $f(x) = f(y)$.** Assume w.l.o.g. $f(x) = f(y) = 1$. Let $(z_n)_n$ be a sequence in $f^{-1}(\{-1\})$ with $|d(y, z_n) - d_f(y)| \leq \frac{1}{n}$. Then

$$\begin{aligned} d_f(x) &= d(x, f^{-1}(\{-1\})) \\ &\leq d(x, z_n) \\ &\leq \|x - y\|_2 + d(y, z_n) \\ &\leq \|x - y\|_2 + d_f(y) + \tfrac{1}{n}. \end{aligned}$$

Letting $n \to \infty$ and exploiting symmetry yields $|d_f(x) - d_f(y)| \leq \|x - y\|_2$.

Case 2: $f(x) \neq f(y)$. Assume w.l.o.g. $f(x) = 1$, $f(y) = -1$. Consider the line segment $L = \{(1 - t)x + ty : t \in [0, 1]\} \subset \mathcal{X}$ and define

$$\begin{aligned} w_1 &= (1 - t_1)x + t_1 y, \quad t_1 := \inf\{t : f((1 - t)x + ty) = -1\}, \\ w_2 &= (1 - t_2)x + t_2 y, \quad t_2 := \sup\{t : f((1 - t)x + ty) = 1\}. \end{aligned}$$

Path-connectedness ensures $t_1 \leq t_2$, otherwise the midpoint between w_1 and w_2 would be labeled both 1 and -1, a contradiction.

Thus,

$$\begin{aligned} |d_f(x) - d_f(y)| &= d(x, f^{-1}(\{-1\})) + d(y, f^{-1}(\{1\})) \\ &\leq \|x - w_1\|_2 + \|y - w_2\|_2 \\ &\leq \|x - y\|_2. \end{aligned}$$

□

Lemma 18. *Let $\mathcal{X} \subset \mathbb{R}^d$ and $f : \mathcal{X} \to \{-1, 1\}$ with $f^{-1}(\{1\})$ closed. Then f can be represented as*

$$f(x) = \operatorname{sgn}(d_f(x)),$$

where we adopt the convention $\operatorname{sgn}(0) = 1$.

Proof. If $d_f(x) \neq 0$, the claim follows directly from the definition of d_f. If $d_f(x) = 0$, then $x \in f^{-1}(\{1\})$ by closedness, so $f(x) = 1 = \operatorname{sgn}(0)$. □

Remark 19. *Lemma 18 justifies the representation $f = \operatorname{sgn} \circ d_f$ used in the proof of Theorem 4. This link between classifiers and their signed distance functions is what allows stability arguments to be combined with smoothness-based tools.*

C PROOF OF THE RADEMACHER BOUND (THEOREM 4)

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 4. 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 (Rademacher Bound)**.** *Suppose Assumptions (H1) and (H2) hold, and that* $\min_{f\in\mathcal{F}} S(f) > S > 0$ *with* $\log|\mathcal{F}| \geq n$.

1. *The empirical Rademacher complexity satisfies*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \;\leq\; K_1 \max\left\{\frac{1}{\sqrt{n}},\; \frac{\sqrt{c}}{S}\cdot\frac{\log|\mathcal{F}|}{n\sqrt{d}}\right\}, \tag{12}$$

   for an absolute constant $K_1 > 0$.

2. *If, in addition,* $f^{-1}(\{1\})$ *is closed and* $\mathcal{X}$ *path connected, the bound sharpens to*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \;\leq\; K_2 \max\left\{\frac{1}{\sqrt{n}},\; \frac{\sqrt{c}}{S}\sqrt{\frac{\log|\mathcal{F}|}{nd}},\; 2\exp\Big(-\frac{dS^2}{8c}\Big)\right\}, \tag{13}$$

   for another absolute constant $K_2 > 0$.

Proof. : **1.** 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 $x_2 \in \mathcal{X}$

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

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 & Sellke in Bubeck & Sellke (2021) 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 (McShane (1934), Kirszbraun (1934)). $\Longrightarrow$ *This allows us to apply isoperimetry and thereby the result in (Bubeck & Sellke, 2021, 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 codomain 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 margin $h_f(x) : \mathcal{X} \to \mathbb{R}$, is 1-Lipschitz continuous with respect to the ℓ_2-norm ((Liu & Hansen, 2024, Prop. 7.5). $\Longrightarrow$ *This allows us to control* $\mathbb{P}(A_t(f)^c)$ *via isoperimetry:*

$$\mathbb{P}(A_t(f)^c) = \mathbb{P}(\overbrace{S(f)}^{=\mathbb{E}[h_f]} - h_f(x) \geq t) \leq \exp\left(-\frac{dt^2}{2cL(h_f)^2}\right) = \exp\left(-\frac{dt^2}{2c}\right). \tag{14}$$

Via Fact 1, we can bound the Rademacher complexity by

$$\begin{aligned}\mathcal{R}_{n,\mu}(\mathcal{F}) &= \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] \\ &\leq \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i F_f(x_i)\right|\right] + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right|\right] \\ &\leq C_1\frac{1}{\sqrt{n}} + C_2\frac{1}{S-t}\sqrt{\frac{c\,\log|\mathcal{F}|}{nd}} + \frac{1}{n}\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\right|\right]. \end{aligned} \tag{15}$$

To control the last term, we subdivide $\mathcal{X}^n$ into subsets on which specific samples achieve a minimum margin. 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 $r > 0$

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right| > r\right) \leq \sum_{f\in\mathcal{F}}\sum_{I\in\mathcal{P}([n])}\mathbb{P}\left(\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right| > r \ \wedge\ x \in A^I(f)\right) \\ &= \sum_{f\in\mathcal{F}}\sum_{I\in\mathcal{P}([n])}\mathbb{P}\left(\left|\sum_{i=1}^n \sigma_i(f-F_f)(x_i)\right| > r \,\middle|\, x\in A^I(f)\right)\mathbb{P}(A^I(f)). \end{aligned} \tag{16}$$

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 fact that the sum of k independent subgaussian random variables with variance proxy σ^2 is itself subgaussian with variance proxy $k\sigma^2$ (Rigollet & Hütter, 2023), we obtain for every $I \subsetneq [n]$ (the case $I = [n]$ being trivial) that

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

On the other hand, we get for $I \subset [n]$ via Equation 14 that

$$\mathbb{P}\left(A^I(f)\right) \leq \mathbb{P}\left(\forall j \in I^c :\ x_j \in A_{\frac{S}{2}}(f)^c\right) = \mathbb{P}\left(x \in A_{\frac{S}{2}}(f)^c\right)^{n-|I|} \leq \exp\left(-\frac{dS^2}{2^3 c}\right)^{n-|I|}.$$

Inserting in Equation 16 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

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

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

$$g(x) = \frac{r^2c_1}{n-x} + \frac{1}{c}(n-x)S^2dc_2,$$

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

$$g'(x) = \frac{r^2 c_1}{(n-x)^2} - \frac{1}{c} S^2 d c_2 \stackrel{!}{=} 0$$
$$\implies x_{+/-} = n \pm \frac{r}{S}\sqrt{\frac{c_1 c}{c_2 d}} =: n \pm \alpha(r) \tag{17}$$

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

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

- **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

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

Using Equation 17, condition $\alpha(r) > n$ is equivalent to $r > nS\sqrt{\frac{c_2 d}{c_1 c}}$. In this range, we have $\mathbb{P}_{(II)}(r) \leq \mathbb{P}_{(I)}(r)$ since

$$\mathbb{P}_{(II)}\left(nS\sqrt{\frac{c_2 d}{c_1 c}}\right) = 2|\mathcal{F}|2^n \exp\left(-2nS^2 d c^{-1} c_2\right) = \mathbb{P}_{(I)}\left(nS\sqrt{\frac{c_2 d}{c_1 c}}\right)$$

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 $r > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f-F)(x_i)\right| > r\right) \leq \mathbb{P}_{(I)}(r) = 2|\mathcal{F}|2^n \exp\left(-2rS\sqrt{\frac{dc_2c_1}{c}}\right). \tag{18}$$

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 Equation 18 that

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

- **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 Vershynin (2018), 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 Equation 15:

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

which, with the additional assumption $\log|\mathcal{F}|\geq n$, gives the result in 1.

2. By Lemma 18, every f admits the representation $f=\operatorname{sgn}\circ d_f$. This lets us follow the infinite-class analysis (presented in detail in the proof of Theorem 13), without the ε-net step in Equation 22. From Lemma 17, d_f is 1-Lipschitz, i.e., $L(d_f)=1$ under the given conditions. Furthermore, recalling the co-stability definition we get

$$S^*(d_f)=\mathbb{E}[|d_f|]=\mathbb{E}[h_f]=S(f).$$

Plugging this into the general bound in Equation 8 gives the result. □

C.1 Comparison to standard bound without accounting for stability

Note that the crucial expectation in the derivation, i.e., the last term in Equation 15, 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

$$\mathbb{E}\left[\max_{1\le i\le N} |X_i|\right] \le \sigma\sqrt{2\log(2N)}, \tag{19}$$

see for instance Rigollet & Hütter (2023). 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 Equation 5:

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

Thus, under the isoperimetry condition, our bound improves on the standard Rademacher complexity estimate whenever the class stability S exceeds $\sqrt{c/d}$, a mild requirement in high-dimensional settings.

D Proof of the Law of Robustness (Corollary 6)

Next, we provide the proof of Corollary 6, which we repeat for convenience.

Theorem (Law of Robustness for Discontinuous Functions)**.** *Assume (H1), (H2), and the additional conditions in 2. of Theorem 4 hold. Let $p := \log|\mathcal{F}| \ge n$. Fix $\varepsilon, \delta \in (0,1)$ and consider the 0–1 loss $\ell_{0\text{–}1}$. There exists an absolute constant $K > 0$ such that, if*

1. *the minimal risk $\sigma^2 := \min_{f\in\mathcal{F}} R_{0\text{–}1}(f)$ satisfies $\sigma^2 \ge \varepsilon$, and*
2. *the sample size n is large enough to ensure (i) $\frac{K}{\sqrt{n}} < \frac{\varepsilon}{3}$ and (ii) $\sqrt{\frac{2\log(2/\delta)}{n}} < \frac{\varepsilon}{2}$,*

then with probability at least $1 - \delta$ (over the sample), the following holds uniformly for all $f \in \mathcal{F}$:

$$\hat{R}_{0\text{–}1}(f) \le \sigma^2 - \varepsilon \quad \Longrightarrow \quad S(f) < \max\left\{\frac{3K}{\varepsilon}\sqrt{\frac{c\log|\mathcal{F}|}{nd}},\ \sqrt{\frac{8c}{d}\log\left(\frac{6K}{\varepsilon}\right)}\right\}.$$

Proof. Let $K > 0$ be an absolute constant such that Equation 5 holds, and define the threshold stability

$$S_* = S_*(p, n, d, \varepsilon) := \max\left\{\frac{3K}{\varepsilon}\sqrt{\frac{c\log|\mathcal{F}|}{nd}},\ \sqrt{\frac{8c}{d}\log\left(\frac{6K}{\varepsilon}\right)}\right\}.$$

Then, Theorem 4, together with condition $2(i)$, implies that

$$\mathcal{R}_{n,\mu}(\mathcal{F}_{S_*}) \le K\max\left\{\frac{1}{\sqrt{n}},\ \frac{\sqrt{c}}{S_*}\sqrt{\frac{\log|\mathcal{F}|}{nd}},\ 2\exp\Big(-\frac{dS_*^2}{8c}\Big)\right\} \le \varepsilon/3,$$

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

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

where we additionally used Equation 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

$$\begin{aligned} 1 - \delta \leq \mathbb{P}(\forall f \in \mathcal{F}_{S_*} : \hat{R}_{0\text{-}1}(f) > \sigma^2 - \varepsilon) = \mathbb{P}(\forall f \in \mathcal{F} : \hat{R}_{0\text{-}1}(f) > \sigma^2 - \varepsilon) \\ + \mathbb{P}(\exists f \in \mathcal{F}_{S_*}^c : \hat{R}_{0\text{-}1}(f) \leq \sigma^2 - \varepsilon), \end{aligned} \tag{20}$$

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 Equation 20. □

E Proof of Rademacher Bound for infinite function classes (Theorem 13)

Here we show how to extend the result for finite function classes to infinite function classes by a covering argument, where the Lipschitz continuity of the parameterization turns out to be crucial. Please find the exact statement about the Rademacher complexity of infinite function classes (of a certain form) below, after restating our new regularity hypothesis replacing (H2).

(H3) The hypothesis class $\mathcal{F}$ is of the form $\mathcal{F} = \operatorname{sgn} \circ \mathcal{G}$, where $\mathcal{G} = \{g_w : \mathcal{X} \to [-1, 1] : w \in \mathcal{W}\}$ is a parameterized class of Lipschitz continuous functions. The parameter space $\mathcal{W} \subset \mathbb{R}^p$ is bounded with $\operatorname{diam}(\mathcal{W}) \leq W$, and the parameterization is Lipschitz continuous, i.e.,

$$\|g_{w_1} - g_{w_2}\|_\infty \leq J \|w_1 - w_2\|.$$

Theorem. *Under assumptions (H1) and (H3), suppose that $S^*(g) > S^* > 0$ and $L(g) \leq L$ for all $g \in \mathcal{G}$. Furthermore, assume that $p \geq n$. Then, for any covering precision $\tilde{\varepsilon} > 0$,*

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \leq K \max\left\{ \sqrt{\tfrac{1}{n}},\ \tfrac{L}{S^*}\sqrt{\tfrac{p}{nd}}\sqrt{c\log(1 + 60WJ\tilde{\varepsilon}^{-1})},\ 2\exp\left(-\tfrac{dS^{*2}}{8cL^2}\right),\ \tfrac{J}{S^*}\tilde{\varepsilon} \right\},$$

where $K > 0$ is an absolute constant independent of $p, n, d, S^*, c, L, J, \tilde{\varepsilon}, W$.

Proof. Given any discontinuous classifier $f_w = \operatorname{sgn} \circ g_w$ for $g_w \in \mathcal{G}$, define its Lipschitz continuous approximation for $\gamma > 0$ as

$$F_{f_w} = \operatorname{sgn}_\gamma \circ g_w,$$

where

$$\operatorname{sgn}_\gamma(t) := \begin{cases} -1, & t \leq -\gamma, \\ \frac{t}{\gamma}, & t \in [-\gamma, \gamma], \\ 1, & t \geq \gamma. \end{cases}$$

This approximation satisfies the useful property that both F_{f_w} and the absolute difference $|f_w - F_{f_w}|$ are Lipschitz continuous in both the input space $\mathcal{X}$ and the weight space $\mathcal{W}$, with

$$L(|\operatorname{sgn}_\gamma \circ g_w - \operatorname{sgn} \circ g_w|) = L(\operatorname{sgn}_\gamma \circ g_w) = \tfrac{L(g_w)}{\gamma}. \tag{21}$$

Following the same strategy as in the proof of Theorem 4 with Lipschitz continuous approximations introduced above (see Equation 15), coupled with a covering argument as in Bubeck & Sellke (2021), we obtain

$$\begin{aligned}\mathcal{R}_{n,\mu}(\mathcal{F}) &= \frac{1}{n}\,\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \sigma_i f(x_i)\Big|\right] \\ &\le \frac{1}{n}\,\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \sigma_i F_f(x_i)\Big|\right] + \frac{1}{n}\,\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\Big|\right] \\ &\le C_1\tfrac{1}{\sqrt{n}} + C_2\tfrac{L}{\gamma}\sqrt{\tfrac{c}{nd}}\underbrace{\sqrt{p\log(1+60WJ\tilde{\varepsilon}^{-1})}}_{\ge\sqrt{\log|\mathcal{F}|}} + \frac{1}{n}\,\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\Big|\right].\end{aligned}$$

Here the parameter $\tilde{\varepsilon}>0$ is related to a $\tilde{\varepsilon}$-net of $\mathcal{W}$, which we denote by $\mathcal{W}_{\tilde{\varepsilon}}$. Note, that $|\mathcal{W}_{\tilde{\varepsilon}}| \le (1+60WJ\tilde{\varepsilon}^{-1})^p$ (see e.g. Vershynin (2018) Corollary 4.2.13) so the same holds true for the induced net $\mathcal{F}_{\tilde{\varepsilon}} = \{\operatorname{sgn}\circ g_w : w\in\mathcal{W}_{\tilde{\varepsilon}}\}$, which also allows us to treat the remaining expectation by subdividing the supremum:

$$\begin{aligned}&\frac{1}{n}\,\mathbb{E}^{\sigma_i,x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \sigma_i (f-F_f)(x_i)\Big|\right] = \frac{1}{n}\,\mathbb{E}^{\sigma_i,x_i}\left[\sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\sup_{w\in B_{\tilde{\varepsilon}}(w_{\tilde{\varepsilon}})}\Big|\sum_{i=1}^n \sigma_i (f_w-F_{f_w})(x_i)\Big|\right] \\ &\le \frac{1}{n}\,\mathbb{E}^{x_i}\left[\sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\sum_{i=1}^n |f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|(x_i)\right] \\ &\qquad + \frac{1}{n}\,\mathbb{E}^{x_i}\left[\sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\sup_{w\in B_{\tilde{\varepsilon}}(w_{\tilde{\varepsilon}})}\sum_{i=1}^n\Big||f_w-F_{f_w}|(x_i)-|f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|(x_i)\Big|\right]. \qquad (22)\end{aligned}$$

By Lipschitz continuity of the parameterization and of $|f-F_f|$ as derived in Equation 21, we obtain

$$\|\,|f_w-F_{f_w}|-|f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|\,\|_\infty \le \tfrac{J}{\gamma}\tilde{\varepsilon} \quad \text{for any } w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}} \text{ and } w\in B_{\tilde{\varepsilon}}(w_{\tilde{\varepsilon}})$$

so that

$$\frac{1}{n}\,\mathbb{E}^{x_i}\left[\sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\sup_{w\in B_{\tilde{\varepsilon}}(w_{\tilde{\varepsilon}})}\sum_{i=1}^n\Big||f_w-F_{f_w}|(x_i)-|f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|(x_i)\Big|\right] \le \frac{J}{\gamma}\tilde{\varepsilon}.$$

Via isoperimetry and using the same bound on the cardinality of $\mathcal{F}_{\tilde{\varepsilon}}$ as before, one concludes that the first expectation in Equation 22 is of the same form as Equation 19 with subgaussian variance proxy $\sigma^2 = \frac{L^2}{\gamma^2}\frac{cn}{d}$ so that

$$\begin{aligned}\frac{1}{n}\,\mathbb{E}^{x_i}\Big[\sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\sum_{i=1}^n |f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|(x_i)\Big] &= \frac{1}{n}\,\mathbb{E}^{x_i}\Big[\sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\sum_{i=1}^n |f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|(x_i) - \mathbb{E}[|f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|]\Big] \\ &\qquad + \sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\mathbb{E}[|f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|] \\ &\le C_3\frac{L}{\gamma}\sqrt{\frac{c}{nd}}\sqrt{p\log(1+60WJ\tilde{\varepsilon}^{-1})} + \sup_{w_{\tilde{\varepsilon}}\in\mathcal{W}_{\tilde{\varepsilon}}}\mathbb{E}[|f_{w_{\tilde{\varepsilon}}}-F_{f_{w_{\tilde{\varepsilon}}}}|].\end{aligned}$$

Finally, for every $f\in\mathcal{F}$,

$$\mathbb{E}[|f-F_f|] = \int_{\mathcal{X}} |f(x)-F_f(x)|\,d\mu(x) \le \mathbb{P}(g(x)\in[-\gamma,\gamma]). \qquad (23)$$

Choosing $\gamma = \frac{S^*(g)}{2}$, we obtain by the definitions of co-margin, and once again isoperimetry (since the co-margin inherits the Lipschitzness from g by design)

$$\begin{aligned}\mathbb{P}\left(g(x) \in [-\gamma, \gamma]\right) &= \mathbb{P}\left(|g(x)| \leq \frac{S^*(g)}{2}\right) \\ &\leq \mathbb{P}\left(|h_g^*(x) - S^*(g)| \geq \frac{S^*(g)}{2}\right) \\ &\leq 2\exp\left(-\frac{d\, S^*(g)^2}{8cL(g)^2}\right) \;\leq\; 2\exp\left(-\frac{d\, S^{*2}}{8cL^2}\right) = 2\exp\left(-\frac{d\, \bar{S}^{*2}}{8c}\right).\end{aligned}$$

Putting it all together, we have

$$\mathcal{R}_{n,\mu}(\mathcal{F}) \leq C_1 \tfrac{1}{\sqrt{n}} + C_2' \tfrac{L}{S^*}\sqrt{\tfrac{c}{nd}}\sqrt{p\log(1 + 60WJ\tilde{\varepsilon}^{-1})} + \frac{2J}{S^*}\tilde{\varepsilon} + 2\exp\left(-\frac{d\, S^{*2}}{8cL^2}\right).$$

□

F Multi-Class Classification

In this section, we briefly outline how our results extend to categorical distributions with $\mathcal{C} \in \mathbb{N}$ classes. We assume that a classifier is given by

$$f : \mathcal{X} \to \{0,1\}^{\mathcal{C}},$$

with exactly one non-zero entry for each $x \in \mathcal{X}$. The additional regularity assumption $(\mathrm{H3})'$, the adaptations of the conditions in $(\mathrm{H3})$ to the multi-class setting can be formalized as follows.

(H3)' The hypothesis class has the form $\mathcal{F} = \operatorname{argmax} \circ \mathcal{G}$, where $\mathcal{G} = \{g_w : \mathcal{X} \to [0,1]^{\mathcal{C}} : w \in \mathcal{W}\}$ is a parameterized family of Lipschitz functions. The parameter space $\mathcal{W} \subset \mathbb{R}^p$ is bounded with $\operatorname{diam}(\mathcal{W}) \leq W$, and the parameterization is Lipschitz:

$$\|g_{w_1} - g_{w_2}\|_\infty \;\leq\; J\,\|w_1 - w_2\|.$$

Thus, we can interpret $g \in \mathcal{G}$ as representing the class probabilities.

Remark 20. *For binary classification, i.e. $\mathcal{C} = 2$, the classifiers are of the form $f : \mathcal{X} \to \{0,1\}^2$, instead of $f : \mathcal{X} \to \{-1,1\}$, as considered earlier. However, one can translate between these representations by post-composing with either*

$$\alpha(x_1, x_2) := x_1 - x_2 \quad \textit{or} \quad \beta(x) := \left(\tfrac{x+1}{2}, \tfrac{1-x}{2}\right).$$

By the contraction principle for Rademacher complexity, it is therefore sufficient to compute the complexity for one of these models.

As in the binary case, our proofs start by considering the Rademacher complexity of the function class $\mathcal{F}$:

$$\mathcal{R}_{n,\mu}(\mathcal{F}) = \frac{1}{n}\,\mathbb{E}^{\sigma_{ij},x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^{n}\sum_{j=1}^{\mathcal{C}}\sigma_{ij}f_j(x_i)\Big|\right] \leq \sum_{j=1}^{\mathcal{C}}\frac{1}{n}\,\mathbb{E}^{\sigma_{ij},x_i}\left[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^{n}\sigma_{ij}f_j(x_i)\Big|\right].$$

Each summand corresponds to a binary classification problem with a one-vs-all classifier f_j. Indeed, f_j is $\frac{2}{S(f)-t}$-Lipschitz on $A_t(f)$. Transforming via

$$f_j \mapsto 2f_j - 1 : \mathcal{X} \to \{-1,1\},$$

we can follow the same reasoning as in Appendix C, obtaining, up to a linear factor of $\mathcal{C}$, the same result as the first part of Theorem 4, generalized to the multi-class setting.

Similarly, under assumption $(\mathrm{H3})$, we can write

$$2f_j - 1 = \operatorname{sgn}\big(g_j - \max_{i\neq j} g_i(x)\big),$$

Table 1: Multi-class definitions.

Concept	Definition
Isoperimetry	$\mathbb{P}(\|f(x) - \mathbb{E}[f]\|_\infty \geq t) \leq 2\exp\Big(-\frac{dt^2}{2cL^2}\Big)$
Rademacher complexity	$\mathcal{R}_{n,\mu}(\mathcal{F}) = \frac{1}{n}\mathbb{E}^{\sigma_{i,j},x_i}\left[\sup_{f\in\mathcal{F}}\left\|\sum_{i=1}^n\sum_{j=1}^{\mathcal{C}}\sigma_{ij}f_j(x_i)\right\|\right]$
Margin	$h_f(x) = \sum_{j=1}^{\mathcal{C}} h_f^j(x), \quad h_f^j(x) := \inf\{\|x-z\|_2 : f(z) \neq j,\, z\in\mathbb{R}^d\}$
Class stability	$S(f) = \sum_{j=1}^{\mathcal{C}} S(f)^j, \quad S(f)^j := \mathbb{E}[h_f^j]$
Co-margin	$h_g^*(x) = \sum_{j=1}^{\mathcal{C}} h_g^{*j}(x), \quad h_g^{*j}(x) := \max\big(0, g_j(x) - \max_{i\neq j} g_i(x)\big)$
Co-stability	$S^*(g) = \sum_{j=1}^{\mathcal{C}} S^{*j}(g), \quad S^{*j}(g) := \mathbb{E}[h_g^{*j}]$

which allows us to proceed as in Appendix E to obtain a multi-class generalization of the second part of Theorem 4 and Theorem 13. The only minor difference lies in bounding the term in Equation 23:

$$\mathbb{E}[|f_j - F_{f_j}|] \leq \mathbb{P}\big[|g_j(x) - \max_{i\neq j} g_i(x)| \leq \gamma\big].$$

Choosing $\gamma = \frac{S^*(g)}{2}$, we use that for all j, $|g_j(x) - \max_{i\neq j} g_i(x)| > h_g^*(x)$, which yields

$$\begin{aligned}\mathbb{P}\big[|g_j(x) - \max_{i\neq j} g_i(x)| \leq \tfrac{S^*(g)}{2}\big] &\leq \mathbb{P}\big[|h_g^*(x) - S^*(f)| \geq \tfrac{S^*(g)}{2}\big] \\ &\leq 2\exp\Big(-\tfrac{d\,S^*(g)^2}{8cL(g)^2}\Big) \\ &\leq 2\exp\Big(-\tfrac{d\,S^{*2}}{8cL^2}\Big) = 2\exp\left(-\frac{d\,\bar{S}^{*2}}{8c}\right).\end{aligned}$$

We conclude that all of our results extend to the multi-class case. Moreover, the measure used in our MNIST and CIFAR-10 experiments (Section 6) is the correct generalization.

G 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 $w \in \{128, 256, 512, 1024, 2048\}$ for MNIST and up to $w = 1024$ for CIFAR10. All models use ReLU activations, batch normalization, and were initialized with standard parametrization. Training was conducted using the Adam optimizer (Kingma & Ba, 2015) for the embedding and output layers, and the Muon optimizer (Jordan et al., 2024) for the hidden layers. Models were trained with a batch size of 256 and learning rate 10^{-3}, until at least 99% training accuracy was achieved, ensuring (near) interpolation. We further used sharpness-aware optimization based on (Foret et al., 2021; Kwon et al., 2021) to reduce variance of the normalized co-stability on MNIST.

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.

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 (Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Madry et al., 2018)) using the Foolbox library (Rauber et al., 2017). For each input x, we recorded the minimum perturbation norm required to change the classifier's prediction, over a grid of radii $\mathbf{r} = (0.002, 0.01, 0.05, 0.1, 1, 2)$. The final stability score $S(f)$ was taken as the average ℓ_2 distance across the dataset.

Normalized Co-Stability estimation. The empirical co-stability $S^*(g)$ is computed via the multi-class margin

$$g_j(x) - \max_{i \neq j} g_i(x), \qquad j = \arg\max_i g_i(x),$$

averaged over the dataset. We estimate the Lipschitz constant $L(g)$ using the efficient ECLIPSE method (Xu & Sivaranjani, 2024), and report the normalized ratio $S^*(g)/L(g)$ as a function of model size.

Implementation. Training and evaluation code is implemented in PyTorch (Paszke et al., 2019). 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.

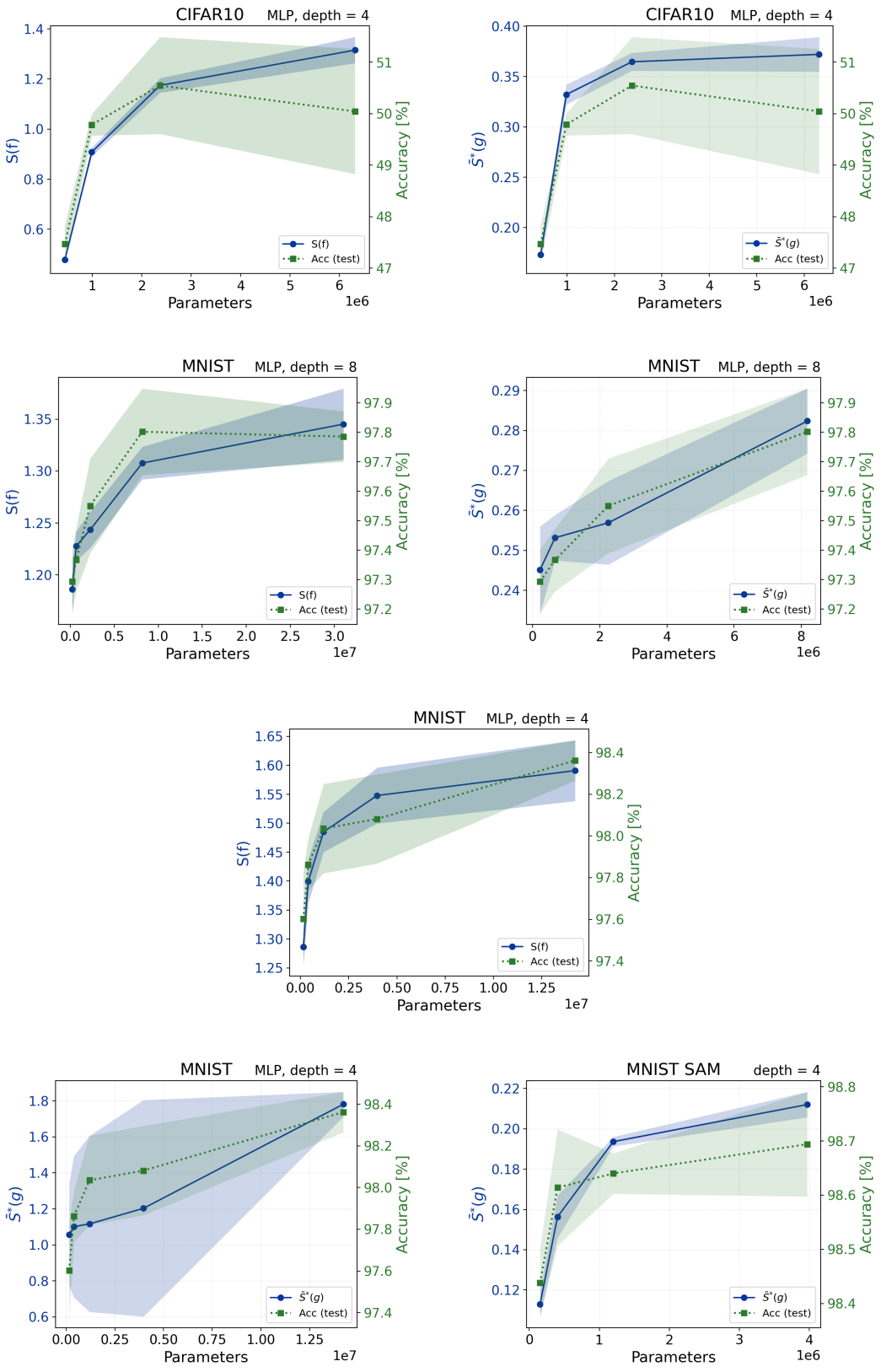

Figure 2: Stability measures for 4- and 8-layer MLPs trained on MNIST and CIFAR-10. For comparison, we also include a 4-layer MLP on MNIST trained with a sharpness-aware optimizer.

H RESPONSE TO REVIEWERS

H.1 EXPERIMENTS ON DIFFERENT ARCHITECTURES

Experiments using Vision Transformers are currently still running, but will be added as soon as possible.

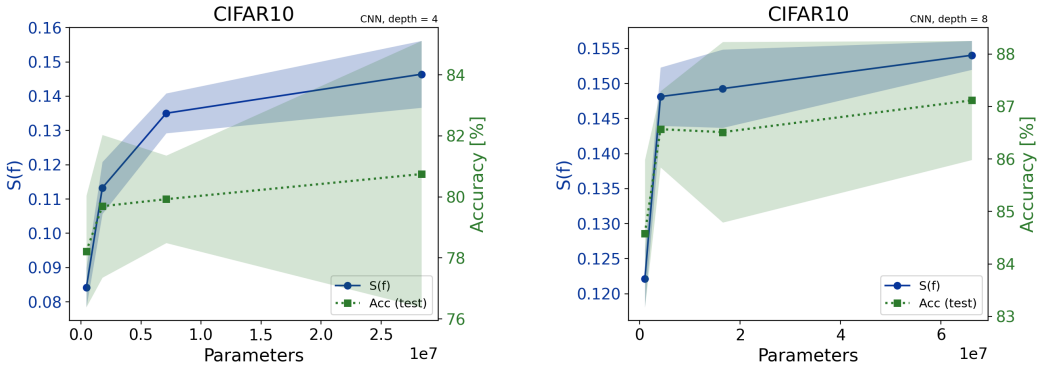

Figure 3: Stability measure for 4- and 8-layer CNNs trained on CIFAR-10.

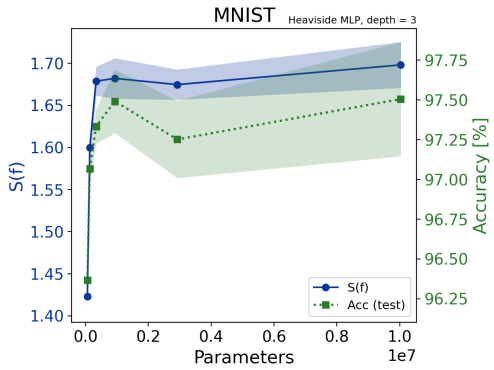

Figure 4: Stability measure for 4-layer heavyside-activation MLP trained on MNIST

H.2 ISOPERIMETRY TEST OF GAUSSIAN TOY-DATA, MNIST AND CIFAR10

Table 2: Isoperimetry test for Gaussian data with variance = 0.0001, $c_{\text{true}} = 0.0784$.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	0.01462	0.01443	0.01653	0.01892
1lip_mlp	R^2	0.9983	0.9985	0.9990	0.9992
linear	$\hat{c}$	0.06701	0.06697	0.06785	0.06836
linear	R^2	0.9983	0.9985	0.9990	0.9993
mlp	$\hat{c}$	2.296e+13	2.071e+10	9.311e+11	1.931e+13
mlp	R^2	0.9963	0.9965	0.9975	0.9982
trained_margin	$\hat{c}$	3.054	2.957	3.626	4.903
trained_margin	R^2	0.9518	0.9497	0.9577	0.9682

Table 3: Isoperimetry test for Gaussian data with variance = 0.001, $c_{\text{true}} = 0.784$.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	0.1113	0.1108	0.1196	0.1303
1lip_mlp	R^2	0.9980	0.9982	0.9988	0.9991
linear	$\hat{c}$	0.6701	0.6699	0.6782	0.6839
linear	R^2	0.9983	0.9985	0.9989	0.9993
mlp	$\hat{c}$	1.216e+13	2.985e+10	3.281e+11	3.432e+12
mlp	R^2	0.9965	0.9965	0.9977	0.9984
trained_margin	$\hat{c}$	15.67	14.41	16.52	25.23
trained_margin	R^2	0.9473	0.9477	0.9542	0.9575

Table 4: Isoperimetry test for Gaussian data with variance = 0.005, $c_{\text{true}} = 3.92$.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	0.4963	0.5011	0.5330	0.5556
1lip_mlp	R^2	0.9974	0.9977	0.9984	0.9989
linear	$\hat{c}$	3.351	3.350	3.391	3.421
linear	R^2	0.9983	0.9984	0.9990	0.9994
mlp	$\hat{c}$	2.826e+12	4.662e+10	8.582e+11	3.753e+12
mlp	R^2	0.9968	0.9968	0.9983	0.9988
trained_margin	$\hat{c}$	42.26	35.82	53.31	59.41
trained_margin	R^2	0.9548	0.9554	0.9601	0.9628

Table 5: Isoperimetry test for Gaussian data with variance = 0.01, $c_{\text{true}} = 7.84$.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	0.9652	0.9676	1.015	1.065
1lip_mlp	R^2	0.9972	0.9974	0.9981	0.9984
linear	$\hat{c}$	6.702	6.712	6.773	6.846
linear	R^2	0.9983	0.9985	0.9990	0.9993
mlp	$\hat{c}$	7.927e+16	4.487e+11	2.637e+13	3.904e+14
mlp	R^2	0.9968	0.9969	0.9979	0.9984
trained_margin	$\hat{c}$	61.88	56.82	78.33	81.59
trained_margin	R^2	0.9598	0.9603	0.9631	0.9677

Table 6: Isoperimetry test for Gaussian data with variance = 0.1, $c_{\text{true}} = 78.4$.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	8.917	8.949	9.408	9.747
1lip_mlp	R^2	0.9968	0.9971	0.9980	0.9983
linear	$\hat{c}$	67.12	67.17	67.89	68.62
linear	R^2	0.9983	0.9986	0.9990	0.9993
mlp	$\hat{c}$	1.606e+16	6.945e+12	5.642e+13	6.913e+14
mlp	R^2	0.9969	0.9970	0.9980	0.9987
trained_margin	$\hat{c}$	240.6	246.0	270.1	276.4
trained_margin	R^2	0.9404	0.9434	0.9524	0.9583

Table 7: Isoperimetry results on MNIST: mean and upper quantiles of $\hat{c}$ and R^2 across models.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	6.68	6.43	7.20	9.04
1lip_mlp	R^2	0.9923	0.9951	0.9970	0.9981
linear	$\hat{c}$	48.02	45.87	56.01	67.69
linear	R^2	0.9957	0.9977	0.9991	0.9996
mlp	$\hat{c}$	5.94e+17	2.24e+12	9.37e+13	1.49e+15
mlp	R^2	0.9879	0.9896	0.9946	0.9968
trained_margin	$\hat{c}$	973.87	552.88	821.24	2626.08
trained_margin	R^2	0.9471	0.9480	0.9563	0.9603

Table 8: Isoperimetry results on CIFAR-10: mean and upper quantiles of $\hat{c}$ and R^2 across models.

kind	metric	mean	q50	q75	q90
1lip_mlp	$\hat{c}$	28.76	26.08	31.46	39.22
1lip_mlp	R^2	0.9817	0.9835	0.9902	0.9943
linear	$\hat{c}$	180.29	165.71	212.00	263.15
linear	R^2	0.9931	0.9940	0.9969	0.9989
mlp	$\hat{c}$	3.32e17	1.38e14	2.62e15	5.65e16
mlp	R^2	0.9776	0.9798	0.9869	0.9892
trained_margin	$\hat{c}$	4845.16	2247.31	8964.91	11386.31
trained_margin	R^2	0.9931	0.9947	0.9979	0.9995

H.3 STABILITY TEST ON GAUSSIAN DATA

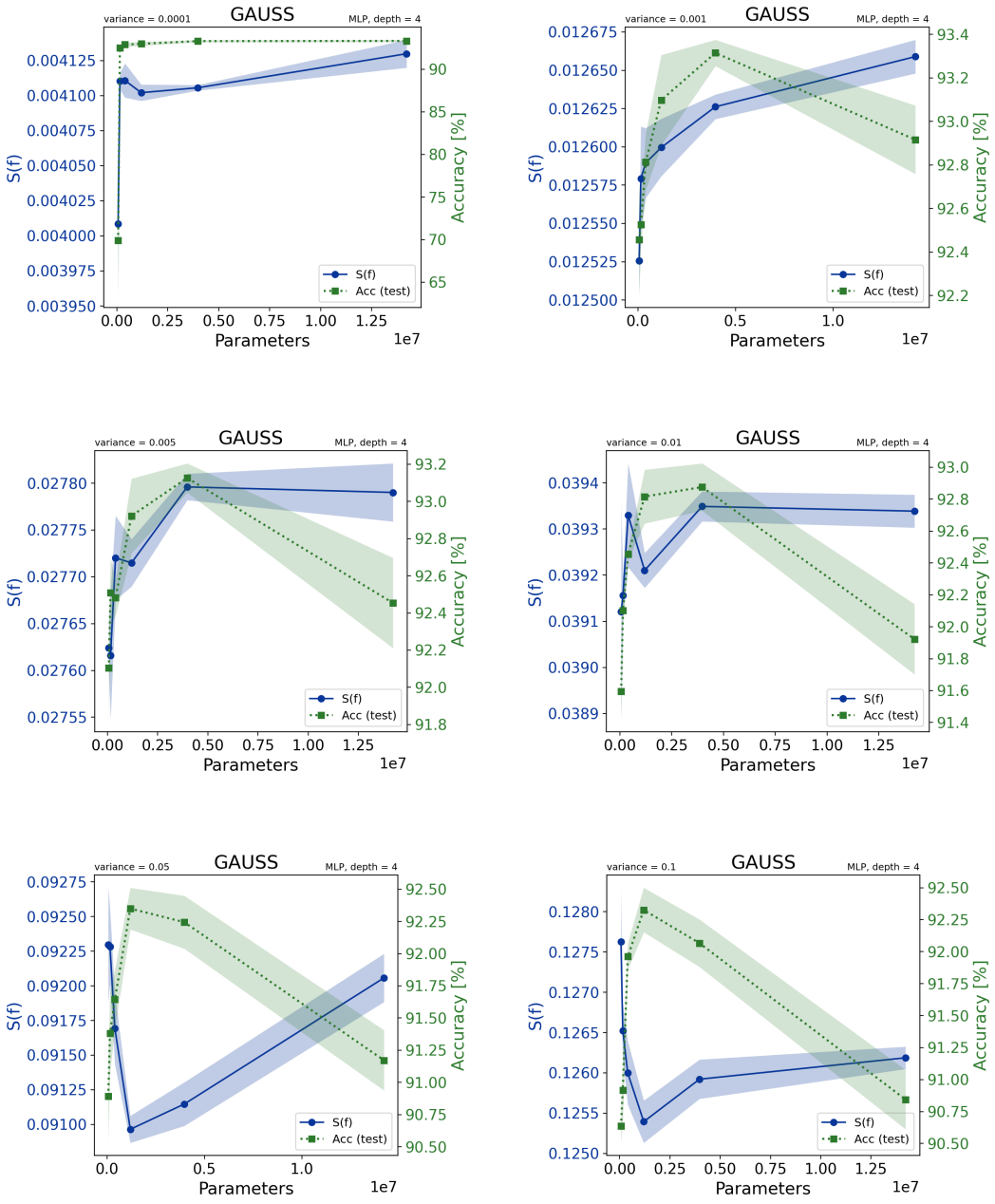

Figure 5: Stability measure for 4-layer MLPs trained on Gaussian toy-data with different variances.