Does Order Matter: Connecting the Law of Robustness to Robust Generalization

Mihir More
Ashoka University, Sonepat, Haryana

Aritra Das
Ashoka University, Sonepat, Haryana
Truth Audit Labs

Jaee Ponde
Ashoka University, Sonepat, Haryana

Himadri Mandal
Indian Statistical Institute, Kolkata

Vishnu Varadarajan
Ashoka University, Sonepat, Haryana

Debayan Gupta
Ashoka University, Sonepat, Haryana
Truth Audit Labs

Abstract

Bubek and Selke [2] propose the connection between the law of robustness and robust generalization error as a *fantastic* open problem. The Law of Robustness states that overparameterization is necessary for models to interpolate robustly, i.e., the function is required to be Lipschitz. Wu et al. [30] extend this law for arbitrary data distributions, proving L = $\Omega(n^{1/d})$. Robust generalization, on the other hand, asks if small robust training loss implies small robust test loss. This can be studied using statistical learning techniques like Rademacher complexities. Accordingly, a bound on the Rademacher complexitiy of the robust loss class implies a bound on the Lipschitzness of the function class. We use this to explicitly connect the two for arbitrary data distributions. i) We prove that the order of the Lipschitz bound remains the same when considering global Rademacher complexity of robust loss classes. ii) At the local scale, that is, for subsets of functions with small empirical error, the order of the Lipschitz bound changes with the perturbation radius ρ and localized concentration term $\sqrt{r/n}$.

1 Introduction

Deep neural networks have demonstrated exceptional performance across different machine learning applications [16]. One of the fundamental methodologies of deep learning involves overparameterized models. These networks are highly expressive and are often easy to optimize to (near) zero training error with SGD. [21] In many settings, making the model even larger can improve test performance beyond the interpolation threshold, a phenomenon also known as double descent [22]. Bubek et al. [3] provide a theoretical explanation for why overparameterization is necessary: they introduce a sharp law of robustness for two layer neural networks showing that achieving robust interpolation characterized by a small Lipschitz constant requires more parameters than interpolation alone. Bubek and Selke [2] then prove a universal version under isoperimetric covariate distributions, showing that smooth interpolation for $O(1)$ Lipschitz requires $p \gg nd$ parameters. Furthermore, Wu et al. [30] extend the theory beyond isoperimetry, providing robust-interpolation lower bounds for arbitrary bounded-support distribution. Despite their remarkable success, deep networks remain vulnerable to subtle perturbations, as revealed through the existence of adversarial samples [12]. The most common mitigation strategy is adversarial training, where models are trained on adversarially perturbed examples to improve robustness [26]. Other methods such as margin-based measures,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

flatness-based measures, and gradient-based measures have also been utilized to improve adverserial robustness [24]. Another line of work pursues certified robustness by constraining the Lipschitz constant of the model, which provides formal guarantees on the maximum possible perturbation impact [14].

However, none of these methods explicitly connect these two different notions of robustness, i.e., the Law of Robustness to robust generalization. Bubek and Selke [2] refer to this connection as a "fantastic open problem." We study this problem through the lens of Rademacher complexity, which provides distribution-dependent uniform bounds on the deviations of the induced loss class.

We first show that, under the assumption of overfitting, we can bound the robust generalization error. We then apply global Rademacher bounds to obtain upper and lower bounds on the complexity of the loss class. This allows us to derive a bound on the Lipschitz constant of the robust loss, which we observe does not change its order and remains $\Omega(n^{1/d})$, consistent with the distribution-free setting in [30].

However, global Rademacher complexity is too coarse for interpolation and low-error regimes, because it measures the complexity over the entire loss class rather than the small-risk region selected by the learning process [1, 15]. Local Rademacher complexity allows us to compute the complexity of subclasses with small empirical risk, thus providing tighter bounds. Covering numbers can be used to relate metric entropy estimates to local Rademacher complexity bounds [17]. Motivated by this, we polynomially bound the metric entropy of Lipschitz function classes using covering numbers and graph theory. We use this result to bound the metric entropy of the robust square loss, and then prove that polynomial entropy implies a local bound on the Rademacher complexity.

We extend our previously obtained lower bounds on the Robust Generalization gap in the global setting (where we showed that the order of the Lipschitz constant does *not* change), to the local setting. We now find that the bound on the Lipschitz constant of the robust loss *is dependent* upon the perturbation radius ρ and localized concentration term $\sqrt{r/n}$. So we answer the titular question, *does order matter*? It turns out this depends on how closely one looks!

2 Related Work

2.1 Adversarial robustness, adversarial training, and Lipschitz

Modern neural networks are vulnerable to *adversarial examples*, a phenomenon first identified by [27, 11]. The most widely used defense is *adversarial training*, which minimizes a worst-case objective over bounded perturbations [20], with principled variants such as TRADES making explicit the trade-off between standard and robust accuracy [35]. Furthermore, parameter-free evaluation suites such as AutoAttack are commonly used for reliable comparison due to robustness being highly sensitive to the threat model [8]. Another line of work deals *certified robustness* by controlling (global or local) Lipschitz. Bounding a model's Lipschitz constant yields worst-case guarantees on output variation under input perturbations, motivating Lipschitz regularization and Lipschitz-constrained architectures. Examples such as Parseval networks [7], Lipschitz-margin training for scalable certification [28], and more recent work emphasizing expressive Lipschitz networks for improved certified robustness [34]. However, computing tight global Lipschitz constants for deep neural networks is difficult, so much work has focused on computable bounds and approximations. CLEVER relates robustness evaluation to local Lipschitz estimation [29], while optimization-based methods provide tighter upper bounds for neural networks [10]. Zuhlke and Kudenko [38] provide a comprehensive survey unifying Lipschitz constants, adversarial attacks, robustness guarantees, and estimation methods.

2.2 Overparameterization and laws of robustness

Recent theory connects *model size* and *smooth interpolation* through lower bounds on the Lipschitz constant of interpolating predictors. Bubek et al. [3] formulate and prove a sharp trade-off for two-layer networks, suggesting that overparameterization is necessary to achieve small Lipschitz constant while fitting the data. Bubek and Selke [2] establish a *universal* version under isoperimetric-type covariate distributions, showing that smooth interpolation can require substantially more parameters than interpolation alone. Wu et al. [30] extend this direction beyond isoperimetry and derive

distribution-free robust-interpolation lower bounds, including an $\Omega(n^{1/d})$ regime for general bounded-support distributions. More recently, Das et al. [9] generalize universal-law-type results beyond squared loss to a broad family of Bregman divergence losses (covering common classification losses), and related work also studies robustness laws under weight-bounded classes [13].

2.3 Robust generalization and robust overfitting

A persistent challenge in adversarial robustness is robust generalization, where robust training error can be low while test error remains high. It is known that robust learning can require significantly more labeled data than standard learning, even in simple data models [25, 4, 33]. It has also been shown by [31] that robust generalization bounds exhibit unavoidable dimension dependence unless additional structure is imposed. A closely related phenomenon is *robust overfitting*: continued adversarial training can improve robust training loss while degrading robust test performance [23]. Subsequent work proposes mechanisms to mitigate robust overfitting via learned smoothing regularization [5] or by analyzing which subsets of adversarial examples drive overfitting and modifying training to counteract it [32]. Moreover, it has recently been shown why why robust generalization is intrinsically challenging from an expressive power perspective [19]. Finally, several recent papers develop stability-based and training-dynamics-based explanations for robust generalization behavior [36, 6, 37]. Additionally, recent work further formalizes the clean-generalization vs. robust-overfitting dichotomy through representation complexity and phase-transition analyses [18].

3 Law of Robustness and Robust Generalization

3.1 Notation

All random variables are defined on a common probability space. The data-generating distribution is denoted by P, and a generic sample is written as $Z = (X, Y) \sim P$, where (X, Y) takes values in $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times [-1, 1]$. Let $\sigma^2 = \mathbb{E}[\mathrm{Var}(Y \mid X)]$ be the irreducible noise level. The training sample is written as $S = \{Z_i = (X_i, Y_i)\}_{i=1}^n \sim P^n$. P_n denotes the empirical measure associated with S. Thus, for any measurable function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we write

$$Pg = \mathbb{E}[g(X, Y)] \qquad \text{and} \qquad P_n g = \frac{1}{n} \sum_{i=1}^{n} g(X_i, Y_i).$$

For $\rho \geq 0$, let $B_\rho(x) =$ denote the closed Euclidean ball of radius ρ centered at x. We also define the ρ-enlargement of the input space by $\mathcal{X}_\rho = \{x + u : x \in \mathcal{X},\ \|u\|_2 \leq \rho\}$. For $L \geq 0$, let

$$\mathcal{F}_L = \{f : \mathcal{X}_\rho \to [-1, 1] : |f(x) - f(x')| \leq L\|x - x'\|_2 \text{ for all } x, x' \in \mathcal{X}_\rho\}$$

be the class of bounded L-Lipschitz functions. When the perturbation radius is clear from context, we suppress the dependence of $\mathcal{F}_L$ on $\mathcal{X}_\rho$.

For $f \in \mathcal{F}_L$, the clean squared loss and the adversarial squared loss at radius ρ are $\ell_{0,f}(x, y) = (f(x) - y)^2$ and $\ell_{\rho,f}(x, y) = \sup_{\tilde{x} \in B_\rho(x)} (f(\tilde{x}) - y)^2$. When $\rho = 0$, this reduces to the clean loss. The corresponding clean population and empirical risks are

$R_0(f) = P\ell_{0,f}$ and $\widehat{R}_0(f) = P_n\ell_{0,f}$, while the robust population risks and empirical risks are $R_\rho(f) = P\ell_{\rho,f}$ and $\widehat{R}_\rho(f) = P_n\ell_{\rho,f}$. The robust generalization gap is denoted by $\mathrm{Gap}_\rho(f) = R_\rho(f) - \widehat{R}_\rho(f)$.

The clean and robust loss classes induced by $\mathcal{F}_L$ are

$$\mathcal{L}_0(\mathcal{F}_L) = \{\ell_{0,f} : f \in \mathcal{F}_L\}, \qquad \mathcal{L}_\rho(\mathcal{F}_L) = \{\ell_{\rho,f} : f \in \mathcal{F}_L\}.$$

When the underlying predictor class is clear, we abbreviate these by $\mathcal{L}_0$ and $\mathcal{L}_\rho$.

Finally, when (T, d_T) is a metric space and $\eta > 0$, the covering number $\mathcal{N}(\eta, T, d_T)$ is the smallest integer N for which there exist $t_1, \ldots, t_N \in T$ such that $T \subseteq \bigcup_{j=1}^N \{t \in T : d_T(t, t_j) \leq \eta\}$. If no such finite cover exists, we set $\mathcal{N}(\eta, T, d_T) = \infty$.

For a class $\mathcal{G}$ of functions and a probability measure Q, we write $\|g - h\|_{L_2(Q)} = \big(Q(g - h)^2\big)^{1/2}$. Thus $\mathcal{N}(\eta, \mathcal{G}, L_2(Q))$ denotes the covering number of $\mathcal{G}$ under the $L_2(Q)$-metric.

We first show that clean overfitting creates a robust train–test gap. Furthermore, we use Rademacher generalization bounds to obtain a lower bound on the complexity of the robust loss class.

Lemma 1 (Robust generalization gap and overfitting). *For any* $f \in \mathcal{F}_L$, $\widehat{R}_\rho(f) \le \left(\sqrt{\widehat{R}_0(f)} + L\rho\right)^2$, *and hence* $\mathrm{Gap}_\rho(f) \ge \sigma^2 - \left(\sqrt{\widehat{R}_0(f)} + L\rho\right)^2$.

Proof. For each $(X_i, Y_i) \in S$ and every $\tilde{X}_i \in B_\rho(X_i)$, $|f(\tilde{X}_i) - Y_i| \le |f(X_i) - Y_i| + |f(\tilde{X}_i) - f(X_i)| \le |f(X_i) - Y_i| + L\rho$. Taking the supremum over $\tilde{X}_i \in B_\rho(X_i)$, squaring, and averaging gives $\widehat{R}_\rho(f) \le \frac{1}{n}\sum_{i=1}^n \left(|f(X_i) - Y_i| + L\rho\right)^2$. By Cauchy–Schwarz, $\frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i| \le \sqrt{\widehat{R}_0(f)}$. Therefore

$$\widehat{R}_\rho(f) \le \widehat{R}_0(f) + 2L\rho\sqrt{\widehat{R}_0(f)} + L^2\rho^2 = \left(\sqrt{\widehat{R}_0(f)} + L\rho\right)^2.$$

Finally, $R_\rho(f) \ge R_0(f) \ge \sigma^2$, where the last inequality follows from the squared-loss variance decomposition. This proves the claim. □

Corollary 1 (Overfitting implies a robust gap). *If* $0 < \varepsilon < \sigma^2$ *and* $\widehat{R}_0(f) \le \sigma^2 - \varepsilon$, *then*

$$\mathrm{Gap}_\rho(f) \ge \Gamma_\rho(\varepsilon, L)$$

where $\Gamma_\rho(\epsilon, L) := \sigma^2 - \left(\sqrt{\sigma^2 - \varepsilon} + L\rho\right)^2 = \varepsilon - 2L\rho\sqrt{\sigma^2 - \varepsilon} - L^2\rho^2$. *In particular, the lower bound is positive whenever* $L\rho < \sigma - \sqrt{\sigma^2 - \varepsilon}$.

Proof. The assumption gives $\sqrt{\widehat{R}_0(f)} \le \sqrt{\sigma^2 - \varepsilon}$; substituting in Theorem 1 proves the result. □

We now convert the robust gap into a lower bound on the complexity of the robust loss class. Since $Y \in [-1, 1]$, every robust squared loss satisfies $0 \le \ell_{\rho,f} \le 4$.

Definition 1 (Rademacher complexities). Let $S = \{Z_i\}_{i=1}^n$ be a sample, and let $\xi_1, \ldots, \xi_n$ be independent Rademacher random variables, independent of S, with $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$.

For a class $\mathcal{G}$ of real-valued functions on $\mathcal{X} \times \mathcal{Y}$, the empirical and expected Rademacher complexity is

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\xi\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \xi_i g(Z_i)\right] \qquad \text{and} \qquad \mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_S\widehat{\mathfrak{R}}_S(\mathcal{G}).$$

Lemma 2 (Complexity forced by a robust gap). *With probability at least* $1 - \delta$ *over* S, *simultaneously for all* $f \in \mathcal{F}_L$,

$$\mathfrak{R}_n(\mathcal{L}_\rho) \ge \frac{1}{2}\left[\sigma^2 - \left(\sqrt{\widehat{R}_0(f)} + L\rho\right)^2 - 4\sqrt{\frac{\log(1/\delta)}{2n}}\right].$$

Consequently, if $\widehat{R}_0(f) \le \sigma^2 - \varepsilon$, *then*

$$\mathfrak{R}_n(\mathcal{L}_\rho) \ge \frac{1}{2}\left[\Gamma_\rho(\varepsilon, L) - 4\sqrt{\frac{\log(1/\delta)}{2n}}\right].$$

Proof. By the standard Rademacher generalization bound for classes bounded in $[0, 4]$, with probability at least $1 - \delta$, every $g \in \mathcal{L}_\rho$ satisfies

$$Pg - P_n g \le 2\mathfrak{R}_n(\mathcal{L}_\rho) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Applying this to $g = \ell_{\rho,f}$ gives, uniformly over $f \in \mathcal{F}_L$,

$$\mathrm{Gap}_\rho(f) \le 2\mathfrak{R}_n(\mathcal{L}_\rho) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Combining this upper bound with Lemma 1 yields

$$\sigma^2 - \left(\sqrt{\widehat{R}_0(f)} + L\rho\right)^2 \le 2\mathfrak{R}_n(\mathcal{L}_\rho) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Rearranging proves the first inequality. The below-noise version follows from Corollary 1. □

We now compare robust and clean squared-loss classes.

Lemma 3 (Extremal envelopes). *For $f \in \mathcal{F}_L$, define $f_\rho^+(x) = \sup_{\tilde{x}\in B_\rho(x)} f(\tilde{x})$ and $f_\rho^-(x) = \inf_{\tilde{x}\in B_\rho(x)} f(\tilde{x})$. Both f_ρ^+ and f_ρ^- are bounded in $[-1,1]$ and L-Lipschitz on $\mathcal{X}$. Moreover, for every $(x,y) \in \mathcal{X}\times[-1,1]$, $\ell_{\rho,f}(x,y) = \max\left\{(f_\rho^+(x)-y)^2, (f_\rho^-(x)-y)^2\right\}$. See Appendix A-8 for the proof.*

Corollary 2 (Squared-loss contraction). *Let $y_1,\dots,y_n \in [-1,1]$, and let $A \subseteq [-1,1]^n$. Then*

$$\widehat{\mathfrak{R}}\left(\left\{\left((a_i-y_i)^2\right)_{i=1}^n : a \in A\right\}\right) \le 4\,\widehat{\mathfrak{R}}(A).$$

Proof. Set $\phi_i(t) = (t-y_i)^2 - y_i^2$. Then $\phi_i(0) = 0$, and ϕ_i is 4-Lipschitz on $[-1,1]$. Extend ϕ_i to a 4-Lipschitz function on $\mathbb{R}$. We use the local contraction Lemma 9 (Appendix A) gives

$$\widehat{\mathfrak{R}}\left(\left\{\left((a_i-y_i)^2 - y_i^2\right)_{i=1}^n : a \in A\right\}\right) \le 4\,\widehat{\mathfrak{R}}(A).$$

Adding the fixed vector $(y_i^2)_{i=1}^n$ does not change Rademacher complexity. □

Lemma 4 (Coordinatewise maxima). *Let $A, B \subseteq \mathbb{R}^n$, and define $A \vee B = \{(a_1 \vee b_1, \dots, a_n \vee b_n) : a \in A,\ b \in B\}$. Then $\widehat{\mathfrak{R}}(A \vee B) \le \widehat{\mathfrak{R}}(A) + \widehat{\mathfrak{R}}(B)$.*

Proof. Using $u \vee v = (u+v+|u-v|)/2$,

$$\widehat{\mathfrak{R}}(A\vee B) \le \frac{1}{2}\widehat{\mathfrak{R}}(A) + \frac{1}{2}\widehat{\mathfrak{R}}(B) + \frac{1}{2}\mathbb{E}_\xi\left[\sup_{a\in A, b\in B}\frac{1}{n}\sum_{i=1}^n \xi_i|a_i-b_i|\right].$$

Apply Lemma 9 to the class $\{a-b : a\in A,\ b\in B\}$ and to $t\mapsto|t|$. The last expectation is at most

$$\mathbb{E}_\xi\left[\sup_{a\in A, b\in B}\frac{1}{n}\sum_{i=1}^n \xi_i(a_i-b_i)\right] \le \widehat{\mathfrak{R}}(A) + \widehat{\mathfrak{R}}(-B).$$

Since $\widehat{\mathfrak{R}}(-B) = \widehat{\mathfrak{R}}(B)$, the claim follows. □

Theorem 1 (Local robust-to-clean reduction). *Fix a sample $S = \{(X_i,Y_i)\}_{i=1}^n$. Let*

$$\mathcal{F}_L(\mathcal{X}) = \{h : \mathcal{X}\to[-1,1] : |h(x)-h(x')| \le L\|x-x'\|_2 \text{ for all } x,x' \in \mathcal{X}\}.$$

Then $\widehat{\mathfrak{R}}_S\big(\mathcal{L}_\rho(\mathcal{F}_L)\big) \le 2\,\widehat{\mathfrak{R}}_S\big(\mathcal{L}_0(\mathcal{F}_L(\mathcal{X}))\big) \le 8\,\widehat{\mathfrak{R}}_{X_1^n}\big(\mathcal{F}_L(\mathcal{X})\big)$.

Proof. By Lemma 3, every robust-loss vector $\big(\ell_{\rho,f}(X_i,Y_i)\big)_{i=1}^n$ is the coordinatewise maximum of two clean squared-loss vectors induced by f_ρ^+ and f_ρ^-. Since both envelopes belong to $\mathcal{F}_L(\mathcal{X})$, the set of robust-loss vectors is contained in $A \vee A$, where

$$A = \left\{\left((h(X_i)-Y_i)^2\right)_{i=1}^n : h \in \mathcal{F}_L(\mathcal{X})\right\}.$$

By monotonicity of Rademacher complexity and Lemma 4,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_\rho(\mathcal{F}_L)\big) \le \widehat{\mathfrak{R}}(A\vee A) \le 2\,\widehat{\mathfrak{R}}(A) = 2\,\widehat{\mathfrak{R}}_S\big(\mathcal{L}_0(\mathcal{F}_L(\mathcal{X}))\big).$$

The second inequality follows from Corollary 2, because $h(X_i)\in[-1,1]$ and $Y_i\in[-1,1]$:

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_0(\mathcal{F}_L(\mathcal{X}))\big) \le 4\,\widehat{\mathfrak{R}}_{X_1^n}\big(\mathcal{F}_L(\mathcal{X})\big).$$

Combining the two bounds proves the claim. Using Lemma 3.8 in [30], L = $\Omega(n^{1/d})$ □

4 Local Rademacher complexity and Robust Generalization

We now localize the robust loss class by its true $L_2(P)$-radius. For $r \geq 0$, define $\mathcal{L}_\rho(r) = \left\{\ell_{\rho,f} : f \in \mathcal{F}_L,\ P\ell_{\rho,f}^2 \leq r\right\}$.

Throughout this section, let $\mathcal{X} \subset \mathbb{R}^d$ be nonempty, compact, and connected, with $D := \operatorname{diam}(\mathcal{X}) < \infty$. Because the robust loss evaluates f on the enlarged set $\mathcal{X}_\rho$, while samples lie in $\mathcal{X}$, we first record a simple extension lemma. It shows that any Lipschitz function defined on $\mathcal{X}$ can be extended to $\mathcal{X}_\rho$ without increasing its Lipschitz constant or leaving the range $[-1, 1]$. This allows us to treat covers of restrictions to $\mathcal{X}$ as covers induced by functions in the original class on $\mathcal{X}_\rho$.

For purposes of bounding the empirical metric entropy on samples from $\mathcal{X}$, it is enough to construct covers of Lipschitz functions on $\mathcal{X}$. Any such approximant may be extended back to $\mathcal{X}_\rho$ while remaining in $\mathcal{B}_L$. This can be proved using McShane's Theorem (Appendix A).We next prove a covering bound for bounded Lipschitz functions on $\mathcal{X}$.

Theorem 2 (Metric entropy of bounded Lipschitz predictors)**.** *Let $\mathcal{F}_L(\mathcal{X})$ be as defined above. There is a constant $c_d > 0$, depending only on d, such that for every $0 < \eta \leq 1$,*

$$\log \mathcal{N}\left(\eta, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right) \leq c_d \left(\frac{1 + LD}{\eta}\right)^d.$$

Proof. The case $L = 0$ is immediate. Assume $L > 0$ and set $\delta = \eta/(8L)$. Let $x_1, \ldots, x_m$ be a maximal δ-separated subset of $\mathcal{X}$. Then it is a δ-net, and the standard packing bound gives

$$m \leq \left(1 + \frac{2D}{\delta}\right)^d = \left(1 + \frac{16LD}{\eta}\right)^d.$$

Connect i and j if $\overline{B}(x_i, \delta) \cap \overline{B}(x_j, \delta) \neq \varnothing$. This graph is connected; otherwise the corresponding unions of δ-balls over two graph components would separate the connected set $\mathcal{X}$. Fix a spanning tree rooted at an arbitrary vertex.

Let $h = \eta/4$ and $G_\eta = h\mathbb{Z} \cap [-1 - h/2, 1 + h/2]$. Choose $\pi_\eta(t) \in G_\eta$ with $|t - \pi_\eta(t)| \leq h/2 = \eta/8$, $t \in [-1, 1]$. Then $|G_\eta| \leq 2 + 8/\eta$. For $f \in \mathcal{F}_L(\mathcal{X})$, define its quantized net-value vector by

$$q(f) = (q_1(f), \ldots, q_m(f)), \qquad q_j(f) = \pi_\eta(f(x_j)).$$

The root coordinate has at most $2 + 8/\eta$ choices. If $p(j)$ is the parent of j in the spanning tree, then $\|x_j - x_{p(j)}\|_2 \leq 2\delta$, and hence $|f(x_j) - f(x_{p(j)})| \leq 2L\delta = \eta/4$. Therefore $|q_j(f) - q_{p(j)}(f)| \leq \eta/8 + \eta/4 + \eta/8 = \eta/2$. Since the grid spacing is $\eta/4$, each non-root coordinate has at most five possible values once its parent value is fixed. Thus the number of admissible quantized net-value vectors is at most $\left(2 + \frac{8}{\eta}\right) 5^{m-1}$.

If $f, g \in \mathcal{F}_L(\mathcal{X})$ have the same quantized net-value vector, then for any $x \in \mathcal{X}$, choosing x_j with $\|x - x_j\|_2 \leq \delta$ gives $|f(x) - g(x)| \leq L\delta + \eta/8 + \eta/8 + L\delta = \eta/2$. Hence each nonempty fiber of the quantization map has $\|\cdot\|_\infty$-diameter at most η. Taking one representative from each nonempty fiber yields an η-cover, so

$$\log \mathcal{N}\left(\eta, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right) \leq \log\left(2 + \frac{8}{\eta}\right) + (\log 5)\left(1 + \frac{16LD}{\eta}\right)^d.$$

Let $A = (1 + LD)/\eta$. Since $0 < \eta \leq 1$, $A \geq 1$, and

$$\log\left(2 + \frac{8}{\eta}\right) \leq \frac{10}{\eta} \leq 10A^d.$$

Also, $1 + \frac{16LD}{\eta} \leq 17A$ Hence $\log \mathcal{N}\left(\eta, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right) \leq \left(10 + 17^d \log 5\right) A^d$. Absorbing the constant into c_d gives

$$\log \mathcal{N}\left(\eta, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right) \leq c_d \left(\frac{1 + LD}{\eta}\right)^d.$$

□

Lemma 5 (Metric entropy of robust squared losses)**.** *There is a constant $C_d > 0$, depending only on d, such that for every $0 < \eta \le 4$,*

$$\log \mathcal{N}\left(\eta, \mathcal{L}_\rho(\mathcal{F}_L), \|\cdot\|_\infty\right) \le C_d \left(\frac{1+LD}{\eta}\right)^d.$$

See Appendix A-12 for the proofs.

Lemma 6 (Polynomial entropy implies a local bound)**.** *Let $\mathcal{G}$ be a measurable class of functions bounded in $[0, b]$, where $b > 0$. Assume that for some $p > 2$ and $A > 0$,*

$$\sup_Q \log \mathcal{N}\left(\eta, \mathcal{G}, L_2(Q)\right) \le A\eta^{-p} \qquad \text{for all } 0 < \eta \le b,$$

where the supremum is over all finitely supported probability measures Q. For $r \ge 0$, define

$$\mathcal{G}(r) = \{g \in \mathcal{G} : Pg^2 \le r\}.$$

Let $\bar{A} := 1 \vee A$. Then there is a constant $C_{p,b} > 0$, depending only on p and b, such that

$$\mathfrak{R}_n(\mathcal{G}(r)) \le C_{p,b}\left(\bar{A}^{1/p} n^{-1/p} + \sqrt{\frac{r}{n}}\right).$$

In particular, if $A \ge 1$, then

$$\mathfrak{R}_n(\mathcal{G}(r)) \le C_{p,b}\left(A^{1/p} n^{-1/p} + \sqrt{\frac{r}{n}}\right).$$

See Appendix A-13 for the proofs.

Theorem 3 (Localized robust Rademacher upper bound)**.** *Assume $d \ge 3$. There is a constant $C_d > 0$, depending only on d, such that for every $r \in [0, 16]$,*

$$\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big) \le C_d\left((1+LD)n^{-1/d} + \sqrt{\frac{r}{n}}\right).$$

Proof. Since $f : \mathcal{X}_\rho \to [-1, 1]$ and $Y \in [-1, 1]$, every robust squared loss satisfies

$$0 \le \ell_{\rho,f} \le 4.$$

By Lemma 5, and since $\|\cdot\|_{L_2(Q)} \le \|\cdot\|_\infty$,

$$\sup_Q \log \mathcal{N}\left(\eta, \mathcal{L}_\rho(\mathcal{F}_L), L_2(Q)\right) \le C_d\left(\frac{1+LD}{\eta}\right)^d.$$

Apply Lemma 6 with $\mathcal{G} = \mathcal{L}_\rho(\mathcal{F}_L)$, $p = d$, $b = 4$, and $A = C_d(1+LD)^d$. This gives

$$\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big) \le C_d\left((1+LD)n^{-1/d} + \sqrt{\frac{r}{n}}\right),$$

after absorbing constants depending only on d. □

Lemma 7 (Localized robust gap forces localized complexity)**.** *Fix $r \in [0, 16]$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$ over S, every $f \in \mathcal{F}_L$ satisfying*

$$P\ell_{\rho,f}^2 \le r \qquad \text{and} \qquad \widehat{R}_0(f) \le \sigma^2 - \varepsilon$$

also satisfies

$$\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big) \ge \frac{1}{2}\left[\Gamma_\rho(\varepsilon, L) - 4\sqrt{\frac{\log(1/\delta)}{2n}}\right].$$

This can be proved using standard McDiarmid's inequality. The detailed proof is in Appendix A.

Corollary 3 (Localized compatibility condition)**.** *Under the assumptions of Theorem 3, fix $r \in [0, 16]$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$, every $f \in \mathcal{F}_L$ satisfying*

$$P\ell_{\rho,f}^2 \le r \qquad \text{and} \qquad \widehat{R}_0(f) \le \sigma^2 - \varepsilon$$

must obey

$$\Gamma_\rho(\varepsilon, L) \le 2C_d \left((1 + LD)n^{-1/d} + \sqrt{\frac{r}{n}} \right) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Equivalently,

$$\varepsilon - 2L\rho\sqrt{\sigma^2 - \varepsilon} - L^2\rho^2 \le 2C_d \left((1 + LD)n^{-1/d} + \sqrt{\frac{r}{n}} \right) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Combine Corollary 11 with Theorem 3. □

5 Local Lipschitz Scaling from Localized Robust Rademacher Bounds

We now rewrite the localized compatibility condition as a lower bound on the Lipschitz scale L. Throughout this section, assume $d \ge 3$, $D = \operatorname{diam}(\mathcal{X}) < \infty$, $0 < \varepsilon < \sigma^2$, $r \in [0, 16]$, $\delta \in (0, 1)$, and $\rho \ge 0$. Define $s_\varepsilon := \sqrt{\sigma^2 - \varepsilon}$. By Corollary 3, with probability at least $1 - \delta$, every $f \in \mathcal{F}_L$ satisfying $P\ell_{\rho,f}^2 \le r$ and $\widehat{R}_0(f) \le \sigma^2 - \varepsilon$ must obey

$$\varepsilon - 2L\rho s_\varepsilon - L^2\rho^2 \le 2C_d \left((1 + LD)n^{-1/d} + \sqrt{\frac{r}{n}} \right) + 4\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{1}$$

Define the localized excess margin

$$\Delta_{n,r,\delta} := \varepsilon - 2C_d n^{-1/d} - 2C_d\sqrt{\frac{r}{n}} - 4\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{2}$$

Then (1) implies

$$\Delta_{n,r,\delta} \le \rho^2 L^2 + \left(2\rho s_\varepsilon + 2C_d D n^{-1/d} \right) L. \tag{3}$$

Thus a nontrivial lower bound on L is obtained precisely in the regime

$$\Delta_{n,r,\delta} > 0.$$

Corollary 4 (Local Lipschitz scaling)**.** *On the event of probability at least $1 - \delta$ from Corollary 3, suppose that $\Delta_{n,r,\delta} > 0$. Then every $f \in \mathcal{F}_L$ satisfying $P\ell_{\rho,f}^2 \le r$ and $\widehat{R}_0(f) \le \sigma^2 - \varepsilon$ must satisfy the following lower bounds.*

If $\rho > 0$, then

$$L \ge \frac{-B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2\Delta_{n,r,\delta}}}{2\rho^2}, \tag{4}$$

where

$$B_{n,\rho,\varepsilon} := 2\rho\sqrt{\sigma^2 - \varepsilon} + 2C_d D n^{-1/d}. \tag{5}$$

Equivalently,

$$L \ge \frac{2\Delta_{n,r,\delta}}{B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2\Delta_{n,r,\delta}}}. \tag{6}$$

In particular,

$$L \ge \frac{\Delta_{n,r,\delta}}{2\rho\sqrt{\sigma^2 - \varepsilon} + 2C_d D n^{-1/d} + \rho\sqrt{\Delta_{n,r,\delta}}}. \tag{7}$$

Consequently, up to constants depending only on d,

$$L = \Omega\left(\frac{\Delta_{n,r,\delta}}{\rho\sqrt{\sigma^2 - \varepsilon} + Dn^{-1/d} + \rho\sqrt{\Delta_{n,r,\delta}}} \right). \tag{8}$$

See Appendix A-5 for the proofs.

Remark 1 (A weaker form using σ). Since $\sqrt{\sigma^2 - \varepsilon} \leq \sigma$, one may replace $B_{n,\rho,\varepsilon}$ by the larger quantity $\overline{B}_{n,\rho} := 2\rho\sigma + 2C_d D n^{-1/d}$. This gives the weaker but sometimes simpler valid lower bound, for $\rho > 0$,

$$L \geq \frac{-\overline{B}_{n,\rho} + \sqrt{\overline{B}_{n,\rho}^2 + 4\rho^2 \Delta_{n,r,\delta}}}{2\rho^2},$$

and hence

$$L = \Omega\left(\frac{\Delta_{n,r,\delta}}{\rho\sigma + Dn^{-1/d} + \rho\sqrt{\Delta_{n,r,\delta}}}\right),$$

whenever $\Delta_{n,r,\delta} > 0$. This is a relaxation of the exact bound, not the exact compatibility condition.

Ignoring only constants depending on d, the confidence term, and the displayed lower-order localization terms, the exact nonvacuous regime can be summarized as

$$L = \Omega\left(\frac{\varepsilon - n^{-1/d} - \sqrt{r/n}}{\rho\sqrt{\sigma^2 - \varepsilon} + Dn^{-1/d} + \rho\sqrt{\varepsilon - n^{-1/d} - \sqrt{r/n}}}\right),$$

provided the numerator is positive.

Dependence on the number of parameters. The localized Rademacher argument above is geometric: it applies to the bounded Lipschitz class on a d-dimensional domain and therefore depends on $n, d, D, r, \rho, \sigma, \varepsilon$, and δ, but not directly on the number of network parameters p. A parameter-count dependence enters only after imposing an additional assumptions on the distribuion like the Isoperimteric distribution [2].

6 Discussion

This work takes a step towards unifying worst-case robustness and robust generalization, as posed by [2]. In the global setting, our analysis connects robust generalization gaps to the global Rademacher complexity of the induced robust loss class, finding that the Lipschitz constant remains of the order $\Omega(n^{1/d})$. Interestingly, when using local notions of Rademacher complexity, we find a contrary outcome: the order of the Lipschitz constant changes with ρ and the concentration term $\sqrt{r/n}$.

We surmise that this is because, in the global setting, the number of functions is huge: in this massive ocean of functions, perturbations are imperceptible in terms of order. However, when we zoom in and look at the functions which might actually be reached by the optimizing algorithm, the changes suddenly become visible: the order of the Lipschitz constant is now directly affected!

There are several natural directions for future work. First, there needs to be experimental exploration of the significance of this concentration term. Next, our theoretical results are stated for square-loss, whereas Robust Generalization is usually in the classification setting using cross-entropy loss; extending the analysis to Bregmann divergence losses would close this gap. Further, distributional robustness may provide another route to connect robust generalization with worst-case robustness.

References

[1] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282.

[2] Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2105.12806.

[3] Sébastien Bubeck, Yuanzhi Li, and Dheeraj M. Nagaraj. A law of robustness for two-layer neural networks. In *Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 804–820, 2021. URL https://proceedings.mlr.press/v134/bubeck21a.html.

[4] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.13736.

[5] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qZzy5urZw9.

[6] Xiwei Cheng, Kexin Fu, and Farzan Farnia. Stability and generalization in free adversarial training. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/pdf?id=62BN6qLQzf.

[7] Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 854–863, 2017. URL https://proceedings.mlr.press/v70/cisse17a.html.

[8] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216, 2020. URL https://proceedings.mlr.press/v119/croce20b.html.

[9] Santanu Das, Jatin Batra, and Piyush Srivastava. A direct proof of a unified law of robustness for bregman divergence losses. *IEEE Transactions on Information Theory*, 71(8):6340–6352, 2025. URL https://arxiv.org/abs/2405.16639.

[10] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1906.04893.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014. URL https://arxiv.org/abs/1412.6572. ICLR 2015.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

[13] Hisham Husain and Borja Balle. A law of robustness for weight-bounded neural networks, 2021. URL https://arxiv.org/abs/2102.08093.

[14] Hoki Kim, Jinseong Park, Yujin Choi, and Jaewook Lee. Fantastic robustness measures: The secrets of robust generalization. In *Advances in Neural Information Processing Systems*, volume 36, pages 48793–48818, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/98a5c0470e57d518ade4e56c6ee0b363-Paper-Conference.pdf.

[15] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. doi: 10.1214/009053606000001019.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

[17] Yunwen Lei, Ke Ding, and Jinfeng Bi. Local rademacher complexity bounds based on covering numbers, 2015. URL https://arxiv.org/abs/1510.01463.

[18] Binghui Li and Yuanzhi Li. On the clean generalization and robust overfitting in adversarial training from two theoretical views: Representation complexity and training dynamics. In *International Conference on Machine Learning*, 2025. URL https://icml.cc/virtual/2025/poster/44184.

[19] Binghui Li, Jikai Jin, Han Zhong, John E. Hopcroft, and Liwei Wang. Why robust generalization in deep learning is difficult: Perspective of expressive power. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Z26xiZkbjgE.

[20] 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*, 2018. URL https://arxiv.org/abs/1706.06083.

[21] Chris Mingard, Henry Rees, Guillermo Valle-Pérez, and Ard A. Louis. Deep neural networks have an inbuilt occam's razor. *Nature Communications*, 16(1), 2025. doi: 10.1038/s41467-024-54813-x. URL https://doi.org/10.1038/s41467-024-54813-x.

[22] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt, 2019. URL https://arxiv.org/abs/1912.02292.

[23] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8093–8104, 2020. URL https://proceedings.mlr.press/v119/rice20a.html.

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

[25] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1804.11285.

[26] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free!, 2019. URL https://arxiv.org/abs/1904.12843.

[27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013. URL https://arxiv.org/abs/1312.6199.

[28] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/13d7a8f4c7e1d60d1b6c7f281f978f62-Abstract.html.

[29] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.10578.

[30] Yihan Wu, Heng Huang, and Hongyang Zhang. A law of robustness beyond isoperimetry. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37439–37455, 2023. URL https://proceedings.mlr.press/v202/wu23g.html.

[31] Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7085–7094, 2019. URL https://proceedings.mlr.press/v97/yin19b.html.

[32] Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v162/yu22b.html.

[33] Runtian Zhai, Tianle Cai, Di He, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gdAC4KDB.

[34] Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2210.01787.

[35] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1901.08573.

[36] Kaibo Zhang, Yunjuan Wang, and Raman Arora. Stability and generalization of adversarial training for shallow neural networks with smooth activation. In *Advances in Neural Information Processing Systems*, 2024. URL https://neurips.cc/virtual/2024/poster/97015.

[37] Zhiyu Zhang, Moritz Backes, and Yang Zhang. Generating less certain adversarial examples improves robust generalization. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=MFPv6d9BCx.

[38] Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial robustness of neural networks from the perspective of lipschitz calculus: A survey. *ACM Computing Surveys*, 2025. doi: 10.1145/3648351. URL https://dl.acm.org/doi/10.1145/3648351.

A Technical appendices and supplementary material

Lemma 8 (Extremal envelopes)**.** *For $f \in \mathcal{F}_L$, define*

$$f_\rho^+(x) = \sup_{\tilde{x} \in B_\rho(x)} f(\tilde{x}), \qquad f_\rho^-(x) = \inf_{\tilde{x} \in B_\rho(x)} f(\tilde{x}).$$

Both f_ρ^+ and f_ρ^- are bounded in $[-1,1]$ and L-Lipschitz on $\mathcal{X}$. Moreover, for every $(x,y) \in \mathcal{X} \times [-1,1]$,

$$\ell_{\rho,f}(x,y) = \max\left\{(f_\rho^+(x) - y)^2,\, (f_\rho^-(x) - y)^2\right\}.$$

Proof. Since $B_\rho(x) \subseteq \mathcal{X}_\rho$, the envelopes are well-defined and bounded in $[-1,1]$. For $x, x' \in \mathcal{X}$ and $\|u\|_2 \le \rho$,

$$f(x+u) \le f(x'+u) + L\|x - x'\|_2.$$

Taking the supremum over u gives

$$f_\rho^+(x) \le f_\rho^+(x') + L\|x - x'\|_2.$$

Exchanging x, x' proves that f_ρ^+ is L-Lipschitz. Since $f_\rho^- = -(-f)_\rho^+$, the same holds for f_ρ^-.

Fix (x,y). Let

$$A_x = \{f(\tilde{x}) : \tilde{x} \in B_\rho(x)\}, \qquad M = \sup A_x, \quad m = \inf A_x.$$

Since $A_x \subseteq [m, M]$,

$$\sup_{z \in A_x} (z-y)^2 \le \max\{(M-y)^2, (m-y)^2\}.$$

The reverse inequality follows by approximating M and m with sequences in A_x. Hence

$$\sup_{z \in A_x} (z-y)^2 = \max\{(M-y)^2, (m-y)^2\},$$

which gives the claimed identity. □

Lemma 9 (Local contraction)**.** *Let $A \subseteq \mathbb{R}^n$. For each i, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be C-Lipschitz and satisfy $\phi_i(0) = 0$. Then*

$$\widehat{\mathfrak{R}}\left(\{(\phi_1(a_1), \ldots, \phi_n(a_n)) : a \in A\}\right) \le C\,\widehat{\mathfrak{R}}(A).$$

Lemma 10 (Range-preserving Lipschitz extension)**.** *Let $g : \mathcal{X} \to [-1,1]$ be L-Lipschitz. Then there exists an L-Lipschitz function $\widetilde{g} : \mathcal{X}_\rho \to [-1,1]$ such that $\widetilde{g} = g$ on $\mathcal{X}$.*

Proof. By the McShane extension theorem, the function

$$\widetilde{g}_0(z) := \inf_{x \in \mathcal{X}} \{g(x) + L\,\|z - x\|\}, \qquad z \in \mathcal{X}_\rho,$$

is an L-Lipschitz extension of g from $\mathcal{X}$ to $\mathcal{X}_\rho$. The metric projection $\Pi_{[-1,1]}(t) := \max\{-1, \min\{t, 1\}\}$ is 1-Lipschitz on $\mathbb{R}$, so $\widetilde{g} := \Pi_{[-1,1]} \circ \widetilde{g}_0$ is still L-Lipschitz, takes values in $[-1,1]$, and agrees with g on $\mathcal{X}$ because $g(\mathcal{X}) \subset [-1,1]$. □

Lemma 11 (Localized robust gap forces localized complexity)**.** *Fix $r \in [0, 16]$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$ over S, every $f \in \mathcal{F}_L$ satisfying*

$$P\ell_{\rho,f}^2 \le r \qquad \text{and} \qquad \widehat{R}_0(f) \le \sigma^2 - \varepsilon$$

also satisfies

$$\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big) \ge \frac{1}{2}\left[\Gamma_\rho(\varepsilon, L) - 4\sqrt{\frac{\log(1/\delta)}{2n}}\right].$$

Proof. Let

$$\Phi(S) := \sup_{g \in \mathcal{L}_\rho(r)} (Pg - P_n g).$$

Since $0 \le g \le 4$ for all $g \in \mathcal{L}_\rho(r)$, McDiarmid's inequality gives, with probability at least $1 - \delta$,

$$\Phi(S) \le \mathbb{E}_S \Phi(S) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

By the usual ghost-sample symmetrization,

$$\mathbb{E}_S \Phi(S) \le \mathbb{E}_{S,S'} \sup_{g \in \mathcal{L}_\rho(r)} \frac{1}{n} \sum_{i=1}^{n} \big(g(Z_i') - g(Z_i)\big) \le 2\,\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big).$$

Hence, on the same event,

$$\sup_{g \in \mathcal{L}_\rho(r)} (Pg - P_n g) \le 2\,\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Now let f satisfy $P\ell_{\rho,f}^2 \le r$. Then $\ell_{\rho,f} \in \mathcal{L}_\rho(r)$, so

$$R_\rho(f) - \widehat{R}_\rho(f) \le 2\,\mathfrak{R}_n\big(\mathcal{L}_\rho(r)\big) + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

If also $\widehat{R}_0(f) \le \sigma^2 - \varepsilon$, then Corollary 1 gives

$$R_\rho(f) - \widehat{R}_\rho(f) \ge \Gamma_\rho(\varepsilon, L).$$

Combining the last two displays and rearranging proves the claim. □

Lemma 12 (Metric entropy of robust squared losses)**.** *There is a constant $C_d > 0$, depending only on d, such that for every $0 < \eta \le 4$,*

$$\log \mathcal{N}\left(\eta, \mathcal{L}_\rho(\mathcal{F}_L), \|\cdot\|_\infty\right) \le C_d \left(\frac{1 + LD}{\eta}\right)^d.$$

Proof. By Lemma 3, for every $f \in \mathcal{F}_L$,

$$\ell_{\rho,f}(x, y) = \max\left\{(f_\rho^+(x) - y)^2,\, (f_\rho^-(x) - y)^2\right\},$$

where $f_\rho^+, f_\rho^- \in \mathcal{F}_L(\mathcal{X})$. Define the auxiliary class

$$\mathcal{M}_\rho := \left\{m_{h_+,h_-} : (x, y) \mapsto \max\left\{(h_+(x) - y)^2,\, (h_-(x) - y)^2\right\} : h_+, h_- \in \mathcal{F}_L(\mathcal{X})\right\}.$$

Then

$$\mathcal{L}_\rho(\mathcal{F}_L) \subseteq \mathcal{M}_\rho.$$

We first convert covers of the larger class $\mathcal{M}_\rho$ into proper covers of the subclass $\mathcal{L}_\rho(\mathcal{F}_L)$. Let $\alpha > 0$, and let $m_1, \dots, m_N \in \mathcal{M}_\rho$ be an α-cover of $\mathcal{M}_\rho$ in $\|\cdot\|_\infty$, where

$$N = \mathcal{N}\left(\alpha, \mathcal{M}_\rho, \|\cdot\|_\infty\right).$$

For each index j such that

$$B_\infty(m_j, \alpha) \cap \mathcal{L}_\rho(\mathcal{F}_L) \ne \varnothing,$$

choose one element

$$g_j \in B_\infty(m_j, \alpha) \cap \mathcal{L}_\rho(\mathcal{F}_L).$$

Discard the remaining indices. We claim that the selected functions $\{g_j\} \subseteq \mathcal{L}_\rho(\mathcal{F}_L)$ form a 2α-cover of $\mathcal{L}_\rho(\mathcal{F}_L)$. Indeed, for any $g \in \mathcal{L}_\rho(\mathcal{F}_L)$, since $\mathcal{L}_\rho(\mathcal{F}_L) \subseteq \mathcal{M}_\rho$, there exists j such that

$$\|g - m_j\|_\infty \le \alpha.$$

For this index j, the ball $B_\infty(m_j, \alpha)$ intersects $\mathcal{L}_\rho(\mathcal{F}_L)$, so a corresponding g_j was selected and satisfies

$$\|g_j - m_j\|_\infty \le \alpha.$$

Therefore, by the triangle inequality,

$$\|g - g_j\|_\infty \le \|g - m_j\|_\infty + \|m_j - g_j\|_\infty \le 2\alpha.$$

Thus

$$\mathcal{N}\left(2\alpha, \mathcal{L}_\rho(\mathcal{F}_L), \|\cdot\|_\infty\right) \le \mathcal{N}\left(\alpha, \mathcal{M}_\rho, \|\cdot\|_\infty\right).$$

Taking $\alpha = \eta/2$, we obtain

$$\mathcal{N}\left(\eta, \mathcal{L}_\rho(\mathcal{F}_L), \|\cdot\|_\infty\right) \le \mathcal{N}\left(\frac{\eta}{2}, \mathcal{M}_\rho, \|\cdot\|_\infty\right).$$

It remains to bound the covering number of $\mathcal{M}_\rho$. Set $\beta := \frac{\eta}{8}$. Since $0 < \eta \le 4$, we have $0 < \beta \le 1/2$, so Lemma 2 applies at scale β. Let $h_1, \dots, h_M \in \mathcal{F}_L(\mathcal{X})$ be a proper β-cover of $\mathcal{F}_L(\mathcal{X})$ in $\|\cdot\|_\infty$, with

$$M = \mathcal{N}\left(\beta, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right).$$

For each pair $1 \le j, k \le M$, define

$$m_{jk}(x, y) := \max\left\{(h_j(x) - y)^2,\ (h_k(x) - y)^2\right\}.$$

By construction, $m_{jk} \in \mathcal{M}_\rho$.

We claim that the functions $\{m_{jk} : 1 \le j, k \le M\}$ form an $\eta/2$-cover of $\mathcal{M}_\rho$. Fix $m_{h_+,h_-} \in \mathcal{M}_\rho$. Choose j, k such that

$$\|h_+ - h_j\|_\infty \le \beta, \qquad \|h_- - h_k\|_\infty \le \beta.$$

For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}\left|(h_+(x) - y)^2 - (h_j(x) - y)^2\right| &= |h_+(x) - h_j(x)|\,|h_+(x) + h_j(x) - 2y| \\ &\le 4\|h_+ - h_j\|_\infty \le 4\beta.\end{aligned}$$

Here we used $h_+(x), h_j(x), y \in [-1, 1]$. Similarly,

$$\left|(h_-(x) - y)^2 - (h_k(x) - y)^2\right| \le 4\beta.$$

Since the maximum map is 1-Lipschitz with respect to the sup norm on $\mathbb{R}^2$, we have

$$\begin{aligned}\left|m_{h_+,h_-}(x, y) - m_{jk}(x, y)\right| &\le \max\left\{\left|(h_+(x) - y)^2 - (h_j(x) - y)^2\right|, \left|(h_-(x) - y)^2 - (h_k(x) - y)^2\right|\right\} \\ &\le 4\beta = \frac{\eta}{2}.\end{aligned}$$

Taking the supremum over (x, y) gives $\|m_{h_+,h_-} - m_{jk}\|_\infty \le \frac{\eta}{2}$. Therefore

$$\mathcal{N}\left(\frac{\eta}{2}, \mathcal{M}_\rho, \|\cdot\|_\infty\right) \le \mathcal{N}\left(\frac{\eta}{8}, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right)^2.$$

Combining the preceding bounds yields

$$\mathcal{N}\left(\eta, \mathcal{L}_\rho(\mathcal{F}_L), \|\cdot\|_\infty\right) \le \mathcal{N}\left(\frac{\eta}{8}, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right)^2.$$

Taking logarithms and applying Lemma 2,

$$\begin{aligned}\log \mathcal{N}\left(\eta, \mathcal{L}_\rho(\mathcal{F}_L), \|\cdot\|_\infty\right) &\le 2 \log \mathcal{N}\left(\frac{\eta}{8}, \mathcal{F}_L(\mathcal{X}), \|\cdot\|_\infty\right) \\ &\le 2c_d \left(\frac{1 + LD}{\eta/8}\right)^d \\ &= 2 \cdot 8^d c_d \left(\frac{1 + LD}{\eta}\right)^d.\end{aligned}$$

Thus the claim holds with

$$C_d := 2 \cdot 8^d c_d.$$

□

Lemma 13 (Polynomial entropy implies a local bound)**.** *Let $\mathcal{G}$ be a measurable class of functions bounded in $[0, b]$, where $b > 0$. Assume that for some $p > 2$ and $A > 0$,*

$$\sup_Q \log \mathcal{N}\left(\eta, \mathcal{G}, L_2(Q)\right) \le A\eta^{-p} \qquad \text{for all } 0 < \eta \le b,$$

where the supremum is over all finitely supported probability measures Q. For $r \ge 0$, define

$$\mathcal{G}(r) = \{g \in \mathcal{G} : Pg^2 \le r\}.$$

Let

$$\bar{A} := 1 \vee A.$$

Then there is a constant $C_{p,b} > 0$, depending only on p and b, such that

$$\mathfrak{R}_n(\mathcal{G}(r)) \le C_{p,b}\left(\bar{A}^{1/p} n^{-1/p} + \sqrt{\frac{r}{n}}\right).$$

In particular, if $A \ge 1$, then

$$\mathfrak{R}_n(\mathcal{G}(r)) \le C_{p,b}\left(A^{1/p} n^{-1/p} + \sqrt{\frac{r}{n}}\right).$$

Proof. If $\mathcal{G}(r) = \varnothing$, the claim is trivial. We therefore assume $\mathcal{G}(r) \ne \varnothing$. Define

$$H_{\mathcal{G}}(\eta) := \sup_Q \log \mathcal{N}\left(\eta, \mathcal{G}, L_2(Q)\right),$$

where the supremum is over finitely supported probability measures Q.

Since every function in $\mathcal{G}$ takes values in $[0, b]$, the $L_2(Q)$-diameter of $\mathcal{G}$ is at most b. Hence, for $\eta \ge b$,

$$\mathcal{N}(\eta, \mathcal{G}, L_2(Q)) \le 1 \qquad \text{for every finitely supported } Q.$$

Consequently,

$$H_{\mathcal{G}}(\eta) \le A\eta^{-p} \qquad \text{for every } \eta > 0.$$

Indeed, for $0 < \eta \le b$ this is the entropy assumption, while for $\eta > b$ the left-hand side is zero.

Let

$$\widetilde{\mathcal{G}} := \{g - g' : g, g' \in \mathcal{G}\}.$$

We first bound the entropy of $\widetilde{\mathcal{G}}$. Fix a finitely supported probability measure Q. If $\{g_1, \dots, g_N\}$ is an $(\eta/2)$-cover of $\mathcal{G}$ in $L_2(Q)$, then the functions

$$g_j - g_k, \qquad 1 \le j, k \le N,$$

belong to $\widetilde{\mathcal{G}}$ and form an η-cover of $\widetilde{\mathcal{G}}$ in $L_2(Q)$. Therefore

$$\mathcal{N}\left(\eta, \widetilde{\mathcal{G}}, L_2(Q)\right) \le \mathcal{N}\left(\frac{\eta}{2}, \mathcal{G}, L_2(Q)\right)^2.$$

Taking logarithms and the supremum over Q gives

$$H_{\widetilde{\mathcal{G}}}(\eta) := \sup_Q \log \mathcal{N}\left(\eta, \widetilde{\mathcal{G}}, L_2(Q)\right) \le 2H_{\mathcal{G}}(\eta/2) \le 2^{p+1} A\eta^{-p} \qquad \text{for all } \eta > 0.$$

For a sample $S = (Z_1, \dots, Z_n)$, write P_n for the empirical measure and define

$$\Psi(\varepsilon) := \mathbb{E}_S \widehat{\mathfrak{R}}_S\left(\left\{h \in \widetilde{\mathcal{G}} : P_n h^2 \le \varepsilon^2\right\}\right).$$

We use Theorem 2 of Lei et al. [17], which gives

$$\mathfrak{R}_n(\mathcal{G}(r)) \le \inf_{\varepsilon > 0}\left[2\Psi(\varepsilon) + \frac{8b\, H_{\mathcal{G}}(\varepsilon/2)}{n} + \sqrt{\frac{2r\, H_{\mathcal{G}}(\varepsilon/2)}{n}}\right].$$

It remains to bound $\Psi(\varepsilon)$. Fix a sample S and set

$$d_n(h, h') := \left(P_n(h - h')^2\right)^{1/2}.$$

For a fixed $\varepsilon > 0$, define the empirical localized difference class

$$\widetilde{\mathcal{G}}_{n,\varepsilon} := \left\{ h \in \widetilde{\mathcal{G}} : P_n h^2 \le \varepsilon^2 \right\}.$$

We need to be careful because covering numbers are proper in our convention. We claim that for every $\alpha > 0$,

$$\mathcal{N}\left(2\alpha, \widetilde{\mathcal{G}}_{n,\varepsilon}, d_n\right) \le \mathcal{N}\left(\alpha, \widetilde{\mathcal{G}}, d_n\right).$$

Indeed, let $u_1, \dots, u_M \in \widetilde{\mathcal{G}}$ be an α-cover of $\widetilde{\mathcal{G}}$ in d_n. For each j such that

$$B_{d_n}(u_j, \alpha) \cap \widetilde{\mathcal{G}}_{n,\varepsilon} \neq \varnothing,$$

choose one element

$$v_j \in B_{d_n}(u_j, \alpha) \cap \widetilde{\mathcal{G}}_{n,\varepsilon}.$$

Discard the remaining indices. Then the selected v_j's lie inside $\widetilde{\mathcal{G}}_{n,\varepsilon}$. Moreover, for every $h \in \widetilde{\mathcal{G}}_{n,\varepsilon}$, there exists j such that $d_n(h, u_j) \le \alpha$, and for that same j, $d_n(v_j, u_j) \le \alpha$. Hence

$$d_n(h, v_j) \le d_n(h, u_j) + d_n(u_j, v_j) \le 2\alpha.$$

Thus the selected elements form a proper 2α-cover of $\widetilde{\mathcal{G}}_{n,\varepsilon}$.

Now apply Lemma A.5 of Lei et al. [17] to $\widetilde{\mathcal{G}}_{n,\varepsilon}$. Let $\varepsilon_k := 2^{-k}\varepsilon$, $k \ge 0$. For every integer $N \ge 1$, $\widehat{\mathfrak{R}}_S\left(\widetilde{\mathcal{G}}_{n,\varepsilon}\right) \le 4\sum_{k=1}^{N} \varepsilon_{k-1}\sqrt{\frac{\log \mathcal{N}\left(\varepsilon_k, \widetilde{\mathcal{G}}_{n,\varepsilon}, d_n\right)}{n}} + \varepsilon_N$. By the proper-subclass covering argument above, $\log \mathcal{N}\left(\varepsilon_k, \widetilde{\mathcal{G}}_{n,\varepsilon}, d_n\right) \le \log \mathcal{N}\left(\frac{\varepsilon_k}{2}, \widetilde{\mathcal{G}}, d_n\right)$. Since P_n is finitely supported, the entropy bound for $\widetilde{\mathcal{G}}$ applies with $Q = P_n$. Therefore $\log \mathcal{N}\left(\frac{\varepsilon_k}{2}, \widetilde{\mathcal{G}}, d_n\right) \le H_{\widetilde{\mathcal{G}}}(\varepsilon_k/2) \le 2^{2p+1} A \varepsilon_k^{-p}$. Thus, for a constant $C_p > 0$ depending only on p,

$$\widehat{\mathfrak{R}}_S\left(\widetilde{\mathcal{G}}_{n,\varepsilon}\right) \le C_p \sqrt{\frac{A}{n}} \sum_{k=1}^{N} \varepsilon_{k-1} \varepsilon_k^{-p/2} + \varepsilon_N.$$

Since $\varepsilon_{k-1} = 2\varepsilon_k$, $\varepsilon_{k-1}\varepsilon_k^{-p/2} = 2\varepsilon_k^{1-p/2} = 2\varepsilon^{1-p/2} 2^{k(p/2-1)}$. Hence

$$\widehat{\mathfrak{R}}_S\left(\widetilde{\mathcal{G}}_{n,\varepsilon}\right) \le C_p \sqrt{\frac{A}{n}}\, \varepsilon^{1-p/2} \sum_{k=1}^{N} 2^{k(p/2-1)} + 2^{-N}\varepsilon.$$

Taking expectation over S, the same bound holds for $\Psi(\varepsilon)$:

$$\Psi(\varepsilon) \le C_p \sqrt{\frac{A}{n}}\, \varepsilon^{1-p/2} \sum_{k=1}^{N} 2^{k(p/2-1)} + 2^{-N}\varepsilon.$$

Now choose $\varepsilon_0 := \bar{A}^{1/p}$, $N := \max\left\{1, \left\lceil \frac{\log_2 n}{p} \right\rceil\right\}$. Since $p > 2$, $\sum_{k=1}^{N} 2^{k(p/2-1)} \le C_p n^{1/2-1/p}$, $2^{-N} \le n^{-1/p}$. Therefore

$$\Psi(\varepsilon_0) \le C_p \sqrt{A}\, \varepsilon_0^{1-p/2} n^{-1/p} + \varepsilon_0 n^{-1/p}.$$

Because $\varepsilon_0^p = \bar{A} = 1 \vee A$, we have $\sqrt{A}\, \varepsilon_0^{1-p/2} \le \varepsilon_0$. Indeed, if $A \ge 1$, then

$$\sqrt{A}\, \varepsilon_0^{1-p/2} = A^{1/2} A^{(1-p/2)/p} = A^{1/p} = \varepsilon_0,$$

while if $A < 1$, then $\varepsilon_0 = 1$ and

$$\sqrt{A}\, \varepsilon_0^{1-p/2} = \sqrt{A} \le 1 = \varepsilon_0.$$

Thus

$$\Psi(\varepsilon_0) \le C_p \bar{A}^{1/p} n^{-1/p}.$$

We next bound the entropy terms in Theorem 2 of Lei et al. [17] at ε_0. Since the entropy bound has been extended above to all positive scales,

$$H_{\mathcal{G}}(\varepsilon_0/2) \le A(\varepsilon_0/2)^{-p} = 2^p A \varepsilon_0^{-p} = 2^p \frac{A}{\bar{A}} \le 2^p.$$

Substituting $\varepsilon = \varepsilon_0$ into the local Rademacher bound gives

$$\mathfrak{R}_n(\mathcal{G}(r)) \le C_p \bar{A}^{1/p} n^{-1/p} + \frac{C_{p,b}}{n} + C_p \sqrt{\frac{r}{n}}.$$

Since $n \ge 1$, $p > 2$, and $\bar{A}^{1/p} \ge 1$,

$$\frac{1}{n} \le \bar{A}^{1/p} n^{-1/p}.$$

Absorbing constants depending only on p and b, we obtain

$$\mathfrak{R}_n(\mathcal{G}(r)) \le C_{p,b} \left(\bar{A}^{1/p} n^{-1/p} + \sqrt{\frac{r}{n}} \right).$$

This proves the claim. □

Corollary 5 (Local Lipschitz scaling)**.** *On the event of probability at least $1 - \delta$ from Corollary 3, suppose that*

$$\Delta_{n,r,\delta} > 0.$$

Then every $f \in \mathcal{F}_L$ satisfying

$$P\ell_{\rho,f}^2 \le r \qquad \text{and} \qquad \widehat{R}_0(f) \le \sigma^2 - \varepsilon$$

must satisfy the following lower bounds.

If $\rho > 0$, then

$$L \ge \frac{-B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2 \Delta_{n,r,\delta}}}{2\rho^2}, \tag{9}$$

where

$$B_{n,\rho,\varepsilon} := 2\rho\sqrt{\sigma^2 - \varepsilon} + 2C_d D n^{-1/d}. \tag{10}$$

Equivalently,

$$L \ge \frac{2\Delta_{n,r,\delta}}{B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2 \Delta_{n,r,\delta}}}. \tag{11}$$

In particular,

$$L \ge \frac{\Delta_{n,r,\delta}}{2\rho\sqrt{\sigma^2 - \varepsilon} + 2C_d D n^{-1/d} + \rho\sqrt{\Delta_{n,r,\delta}}}. \tag{12}$$

Consequently, up to constants depending only on d,

$$L = \Omega\left(\frac{\Delta_{n,r,\delta}}{\rho\sqrt{\sigma^2 - \varepsilon} + D n^{-1/d} + \rho\sqrt{\Delta_{n,r,\delta}}} \right). \tag{13}$$

If $\rho = 0$ and $D > 0$, then

$$L \ge \frac{\Delta_{n,r,\delta}}{2C_d D n^{-1/d}} = \Omega\left(\frac{\Delta_{n,r,\delta} n^{1/d}}{D} \right). \tag{14}$$

If $\rho = 0$, $D = 0$, and $\Delta_{n,r,\delta} > 0$, then no such f can satisfy the two assumptions above on the event of Corollary 3.

Proof. Starting from (3), set

$$B_{n,\rho,\varepsilon} = 2\rho\sqrt{\sigma^2 - \varepsilon} + 2C_d D n^{-1/d}.$$

Then every admissible f must satisfy

$$\rho^2 L^2 + B_{n,\rho,\varepsilon} L - \Delta_{n,r,\delta} \ge 0.$$

First suppose $\rho > 0$. The quadratic

$$q(L) := \rho^2 L^2 + B_{n,\rho,\varepsilon} L - \Delta_{n,r,\delta}$$

has one negative root and one positive root because $\Delta_{n,r,\delta} > 0$ and $B_{n,\rho,\varepsilon} \geq 0$. Since $L \geq 0$, the condition $q(L) \geq 0$ forces L to be at least the positive root. Therefore

$$L \geq \frac{-B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2 \Delta_{n,r,\delta}}}{2\rho^2},$$

which proves (4). Rationalizing the numerator gives

$$\frac{-B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2 \Delta_{n,r,\delta}}}{2\rho^2} = \frac{2\Delta_{n,r,\delta}}{B_{n,\rho,\varepsilon} + \sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2 \Delta_{n,r,\delta}}},$$

which proves (6).

Moreover,

$$\sqrt{B_{n,\rho,\varepsilon}^2 + 4\rho^2 \Delta_{n,r,\delta}} \leq B_{n,\rho,\varepsilon} + 2\rho\sqrt{\Delta_{n,r,\delta}}.$$

Hence

$$L \geq \frac{2\Delta_{n,r,\delta}}{2B_{n,\rho,\varepsilon} + 2\rho\sqrt{\Delta_{n,r,\delta}}} = \frac{\Delta_{n,r,\delta}}{B_{n,\rho,\varepsilon} + \rho\sqrt{\Delta_{n,r,\delta}}}.$$

Substituting the definition of $B_{n,\rho,\varepsilon}$ gives (7), and the order form (8) follows immediately.

Now suppose $\rho = 0$. Then (3) reduces to

$$\Delta_{n,r,\delta} \leq 2C_d D L n^{-1/d}.$$

If $D > 0$, rearranging gives

$$L \geq \frac{\Delta_{n,r,\delta}}{2C_d D n^{-1/d}},$$

which proves (14). If $D = 0$, the right-hand side of the reduced compatibility condition is zero, contradicting $\Delta_{n,r,\delta} > 0$. Hence no admissible f can exist in that case. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the theoretical goal of connecting law-of-robustness bounds to robust generalization.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Discussion section states that the current theory is for squared loss while the experiments use cross-entropy, and it identifies extensions to Bregmann losses, local Lipschitz notions, distributional robustness, etc.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [Yes]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper, regardless of whether the code and data are provided?

   Answer: [N/A]

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details, e.g., data splits, hyperparameters, how they were chosen, type of optimizer, necessary to understand the results?

   Answer: [N/A]

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources, type of compute workers, memory, and time of execution needed to reproduce the experiments?

   Answer: [N/A]

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse, e.g., pre-trained language models, image generators, or scraped datasets?

    Answer: [N/A]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets, e.g., code, data, models, used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?

    Answer: [N/A]

15. **Institutional review board approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board approvals or an equivalent approval/review based on the requirements of the authors' country or institution were obtained?

    Answer: [N/A]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]