STATISTICAL GUARANTEES FOR ROBUST DEEP LEARNING UNDER HEAVY-TAILED NOISE AND COVARIATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust loss functions are widely used in deep-learning pipelines to reduce sensitivity to outliers, heavy tails, and other forms of data corruption. In this paper, we support these empirical findings by providing statistical guarantees. We derive oracle inequalities for empirical risk minimization with Lipschitz losses over nonconvex function classes, including feedforward neural networks. Our results only require weak moment assumptions on the covariates, while remaining noise-agnostic in the oracle comparison. We also study non-Lipschitz target losses through training with Lipschitz surrogates. The resulting bounds provide guarantees for the original target risk and make explicit the estimation–approximation tradeoff, in accordance with the classical statistical robustness–efficiency perspective. The framework recovers Huber-type procedures and extends to more general losses. Numerical experiments under heavy-tailed corruption illustrate the benefits of Lipschitz surrogate training.

1 INTRODUCTION

Deep-learning methods typically rely on large data sets, whose quality is often difficult to assess in advance. Outliers, heavy tails, recording errors, and other atypical observations may have a substantial impact on learning procedures, while remaining difficult to detect or remove systematically. It is therefore crucial to understand when statistical guarantees remain valid under such weak data conditions. Robustness in deep learning has also been studied through adversarial training against input perturbations Madry et al. (2017); Salman et al. (2019). Here, we focus on statistical robustness under heavy-tailed noise and covariates, rather than guarantees against worst-case input perturbations. This perspective is rooted in the classical field of robust statistics (see, for instance, Huber & Ronchetti (2009); Hampel et al. (2011); Stigler (2010)). A widely used approach is to replace outlier-sensitive losses, such as the squared loss, by Lipschitz-continuous alternatives, with controlled growth in the tails, such as the absolute-deviation Huber and Charbonnier losses. More recently, related robust losses have also been used in deep-learning applications, including deep regression and adaptive robust-loss constructions Belagiannis et al. (2015); Barron (2019).

We develop nonasymptotic statistical guarantees for robust learning from i.i.d. observations. Our main contributions are as follows.

1. *Oracle inequalities under weak moments.* We control the oracle-centered excess risk of ordinary empirical risk minimization with Lipschitz losses through the Rademacher complexity of $\mathcal{F}$, without requiring convexity or moment assumptions on the response noise. For norm-constrained feedforward networks, a finite p-th moment of the covariates, $p \in (1, 2]$, yields the rate $n^{-(p-1)/p}$. Finite input variance gives the $n^{-1/2}$ rate, while bounded inputs yield exponential confidence.
2. *Guarantees for Lipschitz surrogate training.* We obtain guarantees for the original target risk when training uses a Lipschitz surrogate, with an explicit estimation–approximation tradeoff, depending on the expected gap between them. For power losses, the surrogate is a generalized Huber loss, and the truncation level in the bound can be optimized explicitly.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

3. *Sharpness of empirical-risk estimation.* We establish that the dependence on sample size and confidence level in empirical-risk-level bound in Theorem 2 is sharp under our weak moment assumptions; see Section 5.

Literature comparison. For Lipschitz convex losses, (Chinot et al., 2020, Theorem 1) derive excess-risk and L_2 estimation bounds under convexity of the candidate class, a sub-Gaussian condition on predictor differences, and a local Bernstein condition, without imposing moment assumptions on the responses. Our Theorems 1 and 3 address a complementary regime: we obtain global oracle inequalities without convexity or Bernstein assumptions. In particular, for norm-constrained feedforward neural networks, only weak moment assumptions on the covariates are required, while the oracle comparison remains agnostic to the response-noise distribution.

A complementary approach to heavy-tailed learning is to robustify empirical averages themselves. For example, Brownlees et al. (2015) replace empirical risks by Catoni-type risk estimates. Our results instead concern ordinary empirical risk minimization with Lipschitz losses: robustness is introduced through the loss rather than through a separate robust estimator of its mean.

More broadly, statistical guarantees for neural networks have been developed in a variety of more structured settings, including classification sample-complexity bounds Bartlett (1998), nonparametric regression with sparse ReLU networks Schmidt-Hieber (2020), and regularized least-squares estimation under light-tailed noise Taheri et al. (2021). Our focus is different: we study how Lipschitz losses permit finite-sample oracle guarantees for nonconvex neural-network classes under weak moment assumptions, and how such guarantees can be transferred to non-Lipschitz target risks through surrogate training.

Robust estimation with quadratic evaluation is classical in robust statistics Huber & Ronchetti (2009) and is also common in image restoration Zhao et al. (2016); Jeevan et al. (2024). Our surrogate results provide oracle inequalities for the original target risk when training uses a Lipschitz surrogate. Under suitable moment assumptions, these bounds make explicit the resulting estimation–approximation tradeoff. The construction recovers Huberization of the squared loss and also applies to more general non-Lipschitz target losses, including members of the Barron family Barron (2019). These connections are developed in Section 4.

2 Setup

Data, loss and empirical risk minimization. The data, also called *data process*, are the i.i.d. pairs $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ and $(y, \boldsymbol{x})$ is a *fresh copy*: a pair distributed as $(y_1, \boldsymbol{x}_1)$ and independent of the data; $E_{y,\boldsymbol{x}}$ denotes expectation with respect to its law only, while $\mathbb{E}$ denotes the expectation with respect to all sources of randomness in the context.

The candidate class is denoted by $\mathcal{F}$, and the oracle $\mathfrak{f}^*$ is a fixed element of $\mathcal{F}$. As for the losses themselves, $\mathfrak{h} : \mathbb{R} \to \mathbb{R}$ is Lipschitz with zero bias at zero:

$$\mathfrak{h}[0] = 0 \quad \text{and} \quad |\mathfrak{h}[a] - \mathfrak{h}[b]| \leq c_{\mathfrak{h}}|a - b|, \tag{LIP}$$

for some positive constant $c_{\mathfrak{h}}$. The normalization $\mathfrak{h}[0] = 0$ is a convention, since one can replace $\mathfrak{h}$ by $\mathfrak{h}' = \mathfrak{h} - \mathfrak{h}[0]$, which has the same Lipschitz constant. The empirical-risk minimizer is

$$\widehat{\mathfrak{f}} \in \arg\min_{\mathfrak{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathfrak{h}[y_i - \mathfrak{f}[\boldsymbol{x}_i]],$$

where we assume the minimum is attained and choose $\widehat{\mathfrak{f}}$ to be a measurable empirical-risk minimizer.

Sizes and complexity measures.

Definition 1 (Complexity measures)**.** *Given a function $\mathfrak{f}^* \in \mathcal{F}$ and $p \geq 1$, we call,*

$$s_{\boldsymbol{x}}^{(p)} := \left(E_{y,\boldsymbol{x}}\left[\|\boldsymbol{x}\|_2^p\right]\right)^{\frac{1}{p}} \quad \text{and} \quad s_{y|\boldsymbol{x}}^{(p)} := \left(E_{y,\boldsymbol{x}}\left[|y - \mathfrak{f}^*[\boldsymbol{x}]|^p\right]\right)^{\frac{1}{p}} \tag{1}$$

the expected size of the input and expected size of the noise, respectively,

$$w_{\mathcal{F}}^{(p)} := \left(E_{y,\boldsymbol{x}}\left[\sup_{\mathfrak{f}\in\mathcal{F}} |\mathfrak{f}[\boldsymbol{x}] - \mathfrak{f}^*[\boldsymbol{x}]|^p\right]\right)^{\frac{1}{p}} \tag{2}$$

the expected size of an envelope of $\mathcal{F}$, and, for an individual $\mathfrak{f} \in \mathcal{F}$,

$$v_{\mathfrak{f}}^{(p)} := \left(E_{y,\boldsymbol{x}}\left[|\mathfrak{f}[\boldsymbol{x}] - \mathfrak{f}^*[\boldsymbol{x}]|^p\right]\right)^{\frac{1}{p}} \tag{3}$$

the expected size of $\mathfrak{f}$ relative to the oracle. Note that $v_{\mathfrak{f}}^{(p)} \leq w_{\mathcal{F}}^{(p)}$ for every $\mathfrak{f} \in \mathcal{F}$.

Uniformity over the whole class $\mathcal{F}$ sometimes requires another complexity measure beyond the envelope $w_{\mathcal{F}}^{(p)}$. More specifically, we use the Rademacher complexity (centered at the oracle).

Definition 2 (Rademacher complexity). *Let $r_1, r_2, \ldots, r_n$ be i.i.d Rademacher random variables (that is, $\mathbb{P}(r_i = -1) = \mathbb{P}(r_i = 1) = 1/2$) independent of the data process and the fresh copy $(y, \boldsymbol{x})$. We call the following quantity*

$$\mathfrak{R}_n := \mathbb{E}\left[\sup_{\mathfrak{f}\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} r_i(\mathfrak{f}[\boldsymbol{x}_i] - \mathfrak{f}^*[\boldsymbol{x}_i])\right|\right] \tag{4}$$

the Rademacher complexity of $\mathcal{F}$ centered at the oracle.

We will assume that $\mathfrak{R}_n < \infty$, since all the bounds hold trivially otherwise. This assumption already provides all the integrability needed in the proofs: conditioning on r_1 and applying Jensen's inequality to the remaining signs gives

$$\mathfrak{R}_n \;\geq\; \frac{1}{n}\mathbb{E}\left|\mathfrak{f}[\boldsymbol{x}_1] - \mathfrak{f}^*[\boldsymbol{x}_1]\right| \;=\; \frac{v_{\mathfrak{f}}^{(1)}}{n} \qquad \text{for every } \mathfrak{f} \in \mathcal{F},$$

so that, by equation LIP, $E_{y,\boldsymbol{x}}\left|\mathfrak{h}[y - \mathfrak{f}[x]] - \mathfrak{h}[y - \mathfrak{f}^*[x]]\right| \leq c_{\mathfrak{h}} v_{\mathfrak{f}}^{(1)} \leq c_{\mathfrak{h}} n \mathfrak{R}_n < \infty$ for every $\mathfrak{f} \in \mathcal{F}$.

Remark 1 (Measurability). *We adopt here an extra assumption on the family $\mathcal{F}$: that it is* pointwise measurable *(Section 2.3 of van der Vaart & Wellner (1996)), or separable for short. That is, there exists a countable subclass $\mathcal{F}_0 \subset \mathcal{F}$ such that every function $\mathfrak{f} \in \mathcal{F}$ is the pointwise limit of functions in $\mathcal{F}_0$. This way, suprema over $\mathcal{F}$ that depend on $\mathfrak{f}$ only through its values, like the complexity measures defined above, coincide with the suprema over $\mathcal{F}_0$, and therefore are measurable. For other functionals like $\mathfrak{f} \mapsto E_{y,\boldsymbol{x}}[\mathfrak{h}[y - \mathfrak{f}[x]]]$, a dominated convergence argument suffices (under $w_{\mathcal{F}}^{(1)} < \infty$, the relevant differences being dominated by $2c_{\mathfrak{h}} \sup_{\mathfrak{f}\in\mathcal{F}} |\mathfrak{f}[x] - \mathfrak{f}^*[x]|$). Such assumptions are mild. For instance, they hold for any class parameterized by a separable set. In particular, neural network classes with Θ living in some subset of some $\mathbb{R}^d$.*

Forward Neural Networks. We consider layered, feedforward neural networks, that is,

$$\mathcal{F} := \{\mathfrak{f}_\Theta : \mathbb{R}^d \to \mathbb{R} : \Theta \in \mathcal{M}\}$$

with $\mathcal{M}$ a nonempty subset of $\bar{\mathcal{M}} := \{\Theta = (\Theta^\ell, \ldots, \Theta^0) : \Theta^j \in \mathbb{R}^{p_{j+1}\times p_j}\}$ and

$$\mathfrak{f}_\Theta[\boldsymbol{x}] := \Theta^\ell \mathfrak{a}^\ell\left[\Theta^{\ell-1} \ldots \mathfrak{a}^1[\Theta^0 \boldsymbol{x}]\right]$$

for $\boldsymbol{x} \in \mathbb{R}^d$. The functions $\mathfrak{a}^j : \mathbb{R}^{p_j} \to \mathbb{R}^{p_j}$ are called the *activation functions*, ℓ the depth of the network, $p_0 := d$ and $p_{\ell+1} := 1$ the input and output dimensions, respectively, and $w := \max\{p_1, \ldots, p_\ell\}$ the width of the network. The empirical-risk minimizers are then the functions

$$\hat{\mathfrak{f}} := \mathfrak{f}_{\widehat{\Theta}} \;\text{ with }\; \widehat{\Theta} \in \arg\min_{\Theta\in\mathcal{M}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}[y_i - \mathfrak{f}_\Theta[\boldsymbol{x}_i]]\right\}.$$

The parameter set is assumed to satisfy

$$\mathcal{M} \subset \left\{\Theta \in \bar{\mathcal{M}} : \max_{j\in\{0,\ldots,\ell\}} |||\Theta^j|||_F \leq b_{\mathcal{M}}\right\} \tag{$\mathcal{M}$}$$

for a fixed $b_{\mathcal{M}} \in [0, \infty)$ and the Frobenius norm

$$|||\Theta^j|||_F := \sqrt{\sum_{i=1}^{p_{j+1}}\sum_{k=1}^{p_j}|(\Theta^j)_{ik}|^2}$$

for $j \in \{0, \dots, \ell\}, \Theta^j \in \mathbb{R}^{p_{j+1} \times p_j}$. Such choices of $\mathcal{M}$ have been popular for more than three decades already and are known under the name "weight decay". We require the activation functions to be a_{Lip}-Lipschitz in the ℓ_2-norm and to have zero bias. More precisely, we assume that for all $j \in \{1, \dots, \ell\}$

$$\mathfrak{a}^j[0] = 0 \quad \text{and} \quad \|\mathfrak{a}^j[\boldsymbol{x}] - \mathfrak{a}^j[\boldsymbol{x}']\|_2 \le a_{\mathrm{Lip}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2. \tag{AL}$$

Some results will also require an additional hypothesis on the activation functions. In particular, that they act coordinatewise through positively homogeneous scalar functions: for all $j \in \{1, \dots, \ell\}$,

$$\begin{aligned} \mathfrak{a}^j[(u_1, \dots, u_{p_j})] &= \big(\sigma_j(u_1), \dots, \sigma_j(u_{p_j})\big), \quad \sigma_j \ a_{\mathrm{Lip}}\text{-Lipschitz}, \\ \sigma_j(\lambda u) &= \lambda\, \sigma_j(u) \text{ for all } \lambda \ge 0,\ u \in \mathbb{R}. \end{aligned} \tag{AH}$$

Observe that the standard ReLU satisfies equation AH with $a_{\mathrm{Lip}} = 1$. Note that equation AH implies equation AL: positive homogeneity forces $\sigma_j(0) = 0$, and coordinatewise Lipschitz continuity gives the ℓ_2-Lipschitz property with the same constant.

3 Main Results

3.1 Robust deep learning with Lipschitz losses

We begin with the deep-learning consequence of our oracle inequalities. The following result compares the risk of the fitted neural network predictor $\widehat{\mathfrak{f}}$ with any fixed $\mathfrak{f}^* \in \mathcal{F}$ under weak moment assumptions on the covariates.

Theorem 1 (Robust deep learning). *Let the oracle $\mathfrak{f}^* = \mathfrak{f}_{\Theta^*}$ for some $\Theta^* \in \mathcal{M}$. Under equation LIP, equation $\mathcal{M}$ and equation AH, for every $p \in (1, 2]$ and every $t \in (0, 1)$, it holds that*

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\Big[\mathfrak{h}[y - \widehat{\mathfrak{f}}[\boldsymbol{x}]] - \mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\Big] \le \frac{4\kappa_{\ell,p} c_{\mathfrak{h}} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)}}{t\, n^{(p-1)/p}} \right) \ge 1 - t,$$

where

$$\kappa_{\ell,p} := \sqrt{2\log(2)(\ell+1)} + 1 + 2^{(2-p)/p}.$$

For fixed architecture, norm constraint, and confidence level, the input moment order determines the guaranteed rate. The theorem allows infinite-variance covariates when $1 < p < 2$, while finite input variance yields the rate $n^{-1/2}$. The bound decays exponentially fast in ℓ if $a_{\mathrm{Lip}} b_{\mathcal{M}} < 1$. At the same time, no moment assumption on the response noise is required. The comparison remains meaningful even when the individual risks are infinite. Indeed, by equation 18,

$$\Big|\mathfrak{h}[y - \widehat{\mathfrak{f}}[\boldsymbol{x}]] - \mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\Big| \le 2 c_{\mathfrak{h}} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} \|\boldsymbol{x}\|_2,$$

so $s_{\boldsymbol{x}}^{(p)} < \infty$ ensures integrability of the loss difference, regardless of the distribution of the noise.

For bounded covariates, the rate remains $n^{-1/2}$, but the dependence on the confidence level improves, without additional assumptions on the response noise.

Corollary 1 (Bounded inputs). *Under the assumptions of Theorem 1, suppose that $\|\boldsymbol{x}\|_2 \le B$ almost surely for some $B \in (0, \infty)$. Then, for every $t \in (0, 1)$, with probability at least $1 - t$,*

$$E_{y,\boldsymbol{x}}\Big[\mathfrak{h}[y - \widehat{\mathfrak{f}}[\boldsymbol{x}]] - \mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\Big] \le \frac{c_{\mathfrak{h}} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} B\big(4\kappa_{\ell,2} + 2\sqrt{2\log(1/t)}\big)}{\sqrt{n}},$$

where $\kappa_{\ell,2} := \sqrt{2\log(2)(\ell+1)} + 2$.

3.2 Oracle Inequalities for Lipschitz losses

We first consider uniform estimation of population risk over $\mathcal{F}$. The following bound separates the complexity of the candidate class from fluctuations of the loss at $\mathfrak{f}^*$.

Theorem 2 (Uniform risk bound). *Let $\mathfrak{h}$ satisfy equation LIP. Then, for every $p \in (1, 2]$ with $s^{(p)}_{y|\boldsymbol{x}} < \infty$ and $t \in (0, 1)$, it holds that*

$$\mathbb{P}\left(\sup_{\mathfrak{f}\in\mathcal{F}}\left|\mathbb{E}\left[\mathfrak{h}[y-\mathfrak{f}[\boldsymbol{x}]]\right]-\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}[y_i-\mathfrak{f}[\boldsymbol{x}_i]]\right| \le \frac{8c_{\mathfrak{h}}\mathfrak{R}_n}{t}+2^{\frac{3}{p}}c_{\mathfrak{h}}\frac{s^{(p)}_{y|\boldsymbol{x}}}{\sqrt[p]{n^{p-1}t}}\right)\ge 1-t.$$

The bound involves two terms. The Rademacher complexity measures the richness of the class $\mathcal{F}$ around the oracle, and does not depend on the noise. The response-moment term above controls fluctuations of the absolute risk level.

For oracle comparisons, we instead work directly with loss differences relative to $\mathfrak{f}^*$, obtaining a bound without response-moment assumptions. Combined with the complexity bound of Proposition 5, the following inequality yields Theorem 1.

Theorem 3 (Oracle-ERM inequality). *Let $\mathfrak{h}$ satisfy equation LIP, and let $\mathfrak{f}^* \in \mathcal{F}$. Then, for every $t \in (0, 1)$, it holds that*

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[\mathfrak{h}[y-\widehat{\mathfrak{f}}[\boldsymbol{x}]]-\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]\right]\le\frac{4c_{\mathfrak{h}}\mathfrak{R}_n}{t}\right)\ge 1-t. \tag{5}$$

The excess risk inside the probability is finite for every realization of the data whenever $\mathfrak{R}_n < \infty$, whatever the distribution of the noise: its integrand is bounded by $c_{\mathfrak{h}}|\widehat{\mathfrak{f}}[\boldsymbol{x}]-\mathfrak{f}^*[\boldsymbol{x}]|$, and $E_{y,\boldsymbol{x}}|\mathfrak{f}[\boldsymbol{x}]-\mathfrak{f}^*[\boldsymbol{x}]|\le n\mathfrak{R}_n$ for every $\mathfrak{f}\in\mathcal{F}$ (Section 2).

We stress that the bound is noise-agnostic: it only appears through the benchmark $\mathbb{E}\left[\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]\right]$. Therefore, no direct assumption on the noise is needed in comparisons between the ERM and the oracle. This is a feature of Lipschitz losses. For a least-squares loss, on the other hand, one has

$$\mathfrak{h}[y-\widehat{\mathfrak{f}}[\boldsymbol{x}]]-\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]=(\mathfrak{f}^*[\boldsymbol{x}]-\widehat{\mathfrak{f}}[\boldsymbol{x}])(2\varepsilon+\mathfrak{f}^*[\boldsymbol{x}]-\widehat{\mathfrak{f}}[\boldsymbol{x}])$$

resulting in a coupling between the noise and the complexity, represented by the distance from the oracle.

When the excess losses are uniformly bounded, the confidence dependence can be strengthened without imposing response-moment assumptions. The following result yields the bounded-input neural-network guarantee of Corollary 1.

Corollary 2 (Exponential confidence for bounded excess losses). *Let $\mathfrak{h}$ satisfy equation LIP, let $\mathfrak{f}^* \in \mathcal{F}$, and suppose that for some $D_g \in (0, \infty)$,*

$$\sup_{\mathfrak{f}\in\mathcal{F}}\left|\mathfrak{h}[y-\mathfrak{f}[\boldsymbol{x}]]-\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]\right|\le D_g \qquad \textit{almost surely.}$$

Then, for every $t \in (0, 1)$, it holds that

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[\mathfrak{h}[y-\widehat{\mathfrak{f}}[\boldsymbol{x}]]-\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]\right]\le 4c_{\mathfrak{h}}\mathfrak{R}_n+D_g\sqrt{\frac{2\log(1/t)}{n}}\right)\ge 1-t.$$

The hypothesis holds with $D_g = c_{\mathfrak{h}}D$ whenever $\sup_{\mathfrak{f}\in\mathcal{F}}|\mathfrak{f}[\boldsymbol{x}]-\mathfrak{f}^*[\boldsymbol{x}]|\le D$ almost surely, and with $D_g = \sup\mathfrak{h}-\inf\mathfrak{h}$ whenever $\mathfrak{h}$ is bounded.

3.3 The Lipschitz truncation principle

The above results require the loss $\mathfrak{h}$ to be Lipschitz. However, one might want to evaluate under non-Lipschitz losses like the squared loss or any other power loss. We show that these cases can still be handled by training with a Lipschitz surrogate, at a price of the gap between them. More formally, we measure performance under a loss $\mathfrak{h}$, but select the candidate $\widehat{\mathfrak{f}}$ using another loss $\widetilde{\mathfrak{h}}$:

$$\widehat{\mathfrak{f}}\in\arg\min_{\mathfrak{f}\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathfrak{h}}[y_i-\mathfrak{f}[\boldsymbol{x}_i]]\right\}.$$

The following principle formalizes the above stating that, whenever the surrogate $\widetilde{\mathfrak{h}}$ is Lipschitz and never exceeds $\mathfrak{h}$, the price of selecting under $\widetilde{\mathfrak{h}}$ is the largest expected gap between the two losses over the class.

Proposition 1 (Surrogate principle). *Let $\mathfrak{f}^* \in \mathcal{F}$, $\mathfrak{h} : \mathbb{R} \to [0, \infty)$ be measurable, and let $\widetilde{\mathfrak{h}} : \mathbb{R} \to [0, \infty)$ satisfy equation LIP, and $\widetilde{\mathfrak{h}} \le \mathfrak{h}$ pointwise. Set $R := \mathfrak{h} - \widetilde{\mathfrak{h}} \ge 0$ and*

$$\Phi(\mathfrak{h}, \widetilde{\mathfrak{h}}) := \sup_{\mathfrak{f} \in \mathcal{F}} E_{y,\boldsymbol{x}} \left[R(y - \mathfrak{f}[\boldsymbol{x}]) \right] \in [0, \infty].$$

Then the empirical-risk minimizer $\widehat{\mathfrak{f}}$ under the surrogate loss $\widetilde{\mathfrak{h}}$ (minimum assumed attained) satisfies, for every $t \in (0, 1)$,

$$\mathbb{P}\left(E_{y,\boldsymbol{x}} \left[\mathfrak{h}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}]) \right] \le E_{y,\boldsymbol{x}} \left[\mathfrak{h}(y - \mathfrak{f}^*[\boldsymbol{x}]) \right] + \frac{4c_{\widetilde{\mathfrak{h}}} \mathfrak{R}_n}{t} + \Phi(\mathfrak{h}, \widetilde{\mathfrak{h}}) \right) \ge 1 - t,$$

all expectations being those of nonnegative functions, hence elements of $[0, \infty]$.

The natural choices for surrogates are truncations at a certain level τ. That is, given $\mathfrak{h}$, we let $\mathfrak{h}_\tau$ be

$$\mathfrak{h}_\tau(x) = \begin{cases} \mathfrak{h}(x), & \text{if } |x| \le \tau \\ \mathfrak{h}(\tau) + \mathfrak{h}'(\tau)(|x| - \tau), & \text{otherwise} \end{cases}, \tag{6}$$

which is a Lipschitz function with Lipschitz constant $\mathfrak{h}'(\tau)$ whenever $\mathfrak{h} : \mathbb{R} \to [0, \infty)$ is an even convex function in $C^1(\mathbb{R})$ by virtue of Lemma 2.

We also let

$$R_\tau(x) := \mathfrak{h}(x) - \mathfrak{h}_\tau(x) \qquad \text{and} \qquad \Phi_{\mathfrak{h}}(\tau) := \sup_{\mathfrak{f} \in \mathcal{F}} E_{y,\boldsymbol{x}} \left[R_\tau(y - \mathfrak{f}[\boldsymbol{x}]) \right]. \tag{7}$$

In the notation of Proposition 1, R_τ is the gap of the surrogate $\mathfrak{h}_\tau$ and $\Phi_{\mathfrak{h}}(\tau) = \Phi(\mathfrak{h}, \mathfrak{h}_\tau)$.

Proposition 2 (Lipschitz selection principle). *Let $\mathfrak{h} : \mathbb{R} \to [0, \infty)$ be even, convex and C^1 with $\mathfrak{h}[0] = 0$, and let $\mathfrak{f}^* \in \mathcal{F}$. Then, for every $\tau > 0$ and every $t \in (0, 1)$, the empirical-risk minimizer $\widehat{\mathfrak{f}}$ under the loss $\mathfrak{h}_\tau$ satisfies*

$$\mathbb{P}\left(E_{y,\boldsymbol{x}} \left[\mathfrak{h}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}]) \right] \le E_{y,\boldsymbol{x}} \left[\mathfrak{h}(y - \mathfrak{f}^*[\boldsymbol{x}]) \right] + \frac{4\mathfrak{h}'(\tau) \mathfrak{R}_n}{t} + \Phi_{\mathfrak{h}}(\tau) \right) \ge 1 - t.$$

The bound makes explicit the *estimation–approximation* tradeoff induced by truncation. Increasing τ reduces the gap $\Phi_{\mathfrak{h}}(\tau)$ between the training and target losses, but can increase the Lipschitz constant $\mathfrak{h}'(\tau)$ and hence the estimation term. The estimation term retains the absence of direct response-noise dependence from the oracle inequality, whereas the approximation term depends on the residual distribution. The following results control this term under higher-moment assumptions.

Explicit bounds on $\Phi_{\mathfrak{h}}(\tau)$. To make $\Phi_{\mathfrak{h}}(\tau)$ explicit, we impose a growth condition on $\mathfrak{h}$ and a higher-moment assumption on the data. We say that $\mathfrak{h}$ has *polynomial growth of order r* if there exists a constant $C_{\mathfrak{h}} > 0$ such that

$$\mathfrak{h}[\boldsymbol{x}] \le C_{\mathfrak{h}} |x|^r \quad \text{for all } x \in \mathbb{R}, \tag{PG}$$

and we assume that, for some $q > r$,

$$s^{(q)}_{y|\boldsymbol{x}} < \infty \quad \text{and} \quad w^{(q)}_{\mathcal{F}} < \infty. \tag{HM}$$

Under equation HM, the quantity

$$M_q := \left(2^{q-1} \left({s^{(q)}_{y|\boldsymbol{x}}}^q + {w^{(q)}_{\mathcal{F}}}^q \right) \right)^{1/q} \tag{8}$$

is finite.

The quantity M_q^q bounds $E_{y,\boldsymbol{x}} |y - \mathfrak{f}[\boldsymbol{x}]|^q$ uniformly over $\mathfrak{f} \in \mathcal{F}$. Lemma 3 in Appendix A.4 uses this moment bound to control the approximation term, yielding oracle inequalities for the original target loss.

The resulting oracle inequalities for general candidate classes are given in Corollaries 3 and 4 in the supplement. For power losses, the truncation level in the bound can be optimized explicitly. We now specialize these results to the neural-network class of Section 2.

3.4 Robust deep learning under generalized Huber loss

Definition 3 (Generalized Huber loss of order r). *For $r > 1$ and $\tau > 0$, the* generalized Huber loss of order r at level τ *is defined as*

$$\mathfrak{h}_\tau(a) = \begin{cases} |a|^r & |a| \le \tau, \\ \tau^r + r\tau^{r-1}(|a| - \tau) & |a| > \tau, \end{cases} \tag{9}$$

that is, the power loss $|a|^r$ on $[-\tau, \tau]$, continued linearly outside with the slope $r\tau^{r-1}$ of the power loss at τ. For $r = 2$ this equals $2\ell_\tau$, twice the classical Huber loss Huber (1964).

Theorem 4 (Robust deep learning with generalized Huber losses). *Let $\mathcal{F}$ be the network class of Section 2 under equation $\mathcal{M}$ and equation AH, and let the oracle be $\mathfrak{f}^* = \mathfrak{f}_{\Theta^*}$ for some $\Theta^* \in \mathcal{M}$. Let $r > 1$ and $q > r$ with $s_{\boldsymbol{x}}^{(q)} < \infty$ and $s_{y|\boldsymbol{x}}^{(q)} < \infty$, let $p \in (1, 2]$ with $s_{\boldsymbol{x}}^{(p)} < \infty$ (for instance $p = \min\{q, 2\}$), and set*

$$\bar{B}_n(t) := \frac{4\,\kappa_{\ell,p}\, a_{\mathrm{Lip}}^{\ell}\, b_{\mathcal{M}}^{\ell+1}\, s_{\boldsymbol{x}}^{(p)}}{t\, n^{\frac{p-1}{p}}}, \qquad t \in (0, 1).$$

For $\tau > 0$, let $\widehat{\mathfrak{f}}_\tau$ be the empirical-risk minimizer over $\mathcal{F}$ under the generalized Huber loss of order r at level τ. Then, for every $\tau > 0$ and $t \in (0, 1)$, it holds with probability at least $1 - t$ that

$$\mathbb{E}_{y,\boldsymbol{x}}\big[|y - \widehat{\mathfrak{f}}_\tau[\boldsymbol{x}]|^r\big] \le \mathbb{E}_{y,\boldsymbol{x}}\big[|y - \mathfrak{f}^*[\boldsymbol{x}]|^r\big] + r\tau^{r-1}\,\bar{B}_n(t) + \frac{M_q^q}{\tau^{q-r}}. \tag{10}$$

If moreover $\bar{B}_n(t) > 0$ and $M_q > 0$, then for the level

$$\bar{\tau}^* := \left(\frac{(q-r)M_q^q}{r(r-1)\bar{B}_n(t)}\right)^{\frac{1}{q-1}}$$

it holds with probability at least $1 - t$ that

$$\mathbb{E}_{y,\boldsymbol{x}}\big[|y - \widehat{\mathfrak{f}}_{\bar{\tau}^*}[\boldsymbol{x}]|^r\big] \le \mathbb{E}_{y,\boldsymbol{x}}\big[|y - \mathfrak{f}^*[\boldsymbol{x}]|^r\big] + C(r,q)\big(\bar{B}_n(t)\big)^{\frac{q-r}{q-1}}\big(M_q^q\big)^{\frac{r-1}{q-1}}, \tag{11}$$

where $C(r,q) := (q-1)\, r^{\frac{q-r}{q-1}}\, (r-1)^{-\frac{r-1}{q-1}}\, (q-r)^{-\frac{q-r}{q-1}}$.

4 Robust Losses in Imaging-Motivated Applications

Our results directly cover robust losses commonly used in imaging. We first specialize our bounds to the Barron loss family, and then consider a setting inspired by image-restoration problems in which robust training losses are combined with L_2-based evaluation criteria.

4.1 Barron losses and Lipschitz truncation

The Barron family Barron (2019),

$$\mathfrak{h}_{\alpha,\beta}(u) = \frac{|\alpha - 2|}{\alpha}\left[\left(1 + \frac{(u/\beta)^2}{|\alpha - 2|}\right)^{\alpha/2} - 1\right], \qquad \beta > 0,$$

with continuous extensions at $\alpha = 0$ and $\alpha = 2$, includes the Charbonnier loss at $\alpha = 1$ and a scaled squared loss at $\alpha = 2$. For $\alpha \le 1$, the loss is globally Lipschitz and falls directly within Theorem 1. For $1 < \alpha \le 2$, it is non-Lipschitz, even, convex, and of polynomial growth. Applying the tangent-line truncation equation 20 gives a Lipschitz training loss while preserving the original Barron loss below the cutoff. Under suitable higher-moment assumptions, Proposition 2 then provides guarantees for the original Barron risk. Explicit bounds are given in Section A.6 of the supplement.

Robust training with squared-error evaluation. In image restoration, robust losses are often used during training, while performance is assessed through squared-error criteria such as MSE and PSNR Lai et al. (2017). For example, Zhao et al. (2016) report that L_1 training can outperform L_2 training in PSNR, while Jeevan et al. (2024) obtain their highest PSNR with Huber training. Our truncation results address this distinction between training and evaluation: for $\mathfrak{h}(u) = u^2$, the Lipschitz truncation coincides with the classical Huber loss up to rescaling, and Theorem 4 provides a guarantee for the original squared-error risk under its moment assumptions.

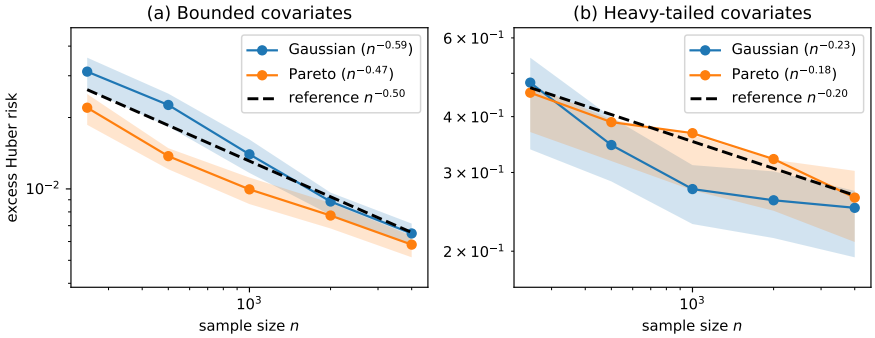


Figure 1: Test excess Huber risk versus sample size for bounded (left) and Student-$t_{1.25}$ (right) covariates. Curves show averages over repetitions and shading indicates the interquartile range. Dashed lines show $n^{-1/2}$ and the limiting-exponent reference $n^{-0.20}$ as $p \uparrow 1.25$, respectively. Results show agreement with the theoretical exponents, and tighter bands for bounded covariates.

4.2 Numerical simulation

We examine the oracle excess risk in Theorem 1 using norm-constrained neural networks trained with Huber loss. We generate observations according to $y = \mathfrak{f}^*[\boldsymbol{x}] + \varepsilon$, $\mathfrak{f}^* \in \mathcal{F}$, and compare Gaussian and symmetric Pareto response noise under bounded and Student-$t_{1.25}$ covariates. For increasing sample sizes, we estimate the oracle-centered loss difference on independent test samples. The detailed setup is given in Section A.7 of the supplement.

Figure 1 reports the average excess risk over repetitions. The dashed lines indicate the reference rates $n^{-1/2}$ in the bounded-input setting and $n^{-(p-1)/p}$ in the finite-p-moment setting. The empirical decay is very similar under Gaussian and Pareto response noise within each input regime. This is consistent with our results, in which the rate of the oracle-centered excess risk is noise-agnostic. The bounded-input experiments exhibit smaller variability across repetitions, consistent with the stronger concentration guarantees available under bounded covariates.

4.3 L_2 evaluation of Lipschitz surrogates under heavy-tailed corruption

Motivated by Proposition 1 and Theorem 4, we compare robust training losses using squared prediction error as a common evaluation criterion. We generate $n = 1000$ observations with bounded covariates according to

$$y = \mathfrak{f}^*[\boldsymbol{x}] + \sigma Z + B\xi, \qquad Z \sim N(0,1), \quad B \sim \text{Bernoulli}(\epsilon),$$

where ξ is Student distributed and the noise components are mutually independent and independent of $\boldsymbol{x}$. We vary ϵ from 0 to 0.2 and compare L_2, L_1, Huber, Barron, and truncated Barron training, with cutoff $\tau = 3\sigma$. All methods use the same samples within each repetition. Figure 2 reports the results; experimental details are given in Section A.8 of the supplement.

Figure 2 illustrates the classical robustness–efficiency tradeoff Huber & Ronchetti (2009). Huber and truncated Barron preserve their respective central loss shapes while treating residuals beyond the cutoff linearly. Least-squares training deteriorates as corruption increases, while L_1 can become preferable to fixed-cutoff Huber. We include ordinary Barron as an ablation to distinguish the benefit of choosing the Barron family from that of truncation. It performs comparably to Huber, whereas its truncated counterpart achieves lower average squared prediction error in the tested settings.

5 Optimality and limits

We establish sharpness of the empirical-risk-level term in Theorem 2, using a scaled-Bernoulli construction following the approach of Devroye et al. (2016).

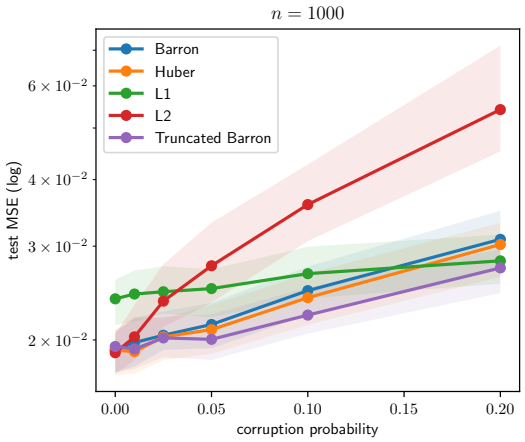


Figure 2: L_2 test error under heavy-tailed response corruption, when training with L_2 vs. robust and truncated losses. The cutoff is fixed at $\tau = 3\sigma$. Solid lines show averages over repetitions and shaded regions the interquartile range. The Huber and truncated Barron loss improve over the untruncated counterparts, with truncated Barron performing best, illustrating the benefit of Lipschitz truncation.

Proposition 3 (Lower bound for empirical mean under weak moments)**.** *Let $p \in (1, 2]$, $n \in \mathbb{N}$, $t \in (0, 1/4]$ and $m > 0$. Also, let X be a random variable taking value $a := m(n/2t)^{1/p}$ with probability $q := 2t/n$ and value 0 otherwise. Then, n independent copies of X satisfy*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \geq \frac{m}{2^{1+1/p}\sqrt[p]{n^{p-1}t}}\right) > t.$$

To relate the proposition to Theorem 2, take $\mathfrak{h}(u) = u$, $\mathcal{F} = \{\mathfrak{f}^*\}$ with $\mathfrak{f}^* \equiv 0$, and $y = X$ independently of $\boldsymbol{x}$. Then $\mathfrak{R}_n = 0$ and $s^{(p)}_{y|\boldsymbol{x}} = m$, so the theorem gives an empirical-mean deviation bound of $2^{3/p} m/(n^{p-1}t)^{1/p}$ with probability at least $1 - t$. Proposition 3 matches its sample-size and confidence dependence up to constants. Moreover, in Devroye et al. (2016) (Theorem 3.1), the authors show that for any estimator of the mean $\widehat{E}$, and $t \leq 1/2$,

$$\sup_{P:\mathbb{E}_P|X-\mu_P|^p=M} \mathbb{P}\left(|\widehat{E}_n - \mu_P| > M^{1/p}(\log(2/t)/n)^{(p-1)/p}\right) \geq t,$$

At fixed confidence, this gives the same sample-size exponent as our empirical-risk bound. Thus, the exponent $(p-1)/p$ appearing in Theorem 2 is the minimax one, and only the confidence-level dependence separates empirical means from the best possible ones. Robust mean estimators can attain the corresponding logarithmic confidence dependence Devroye et al. (2016), whereas Proposition 3 shows that the empirical mean necessarily retains polynomial dependence. Thus, the $t^{-1/p}$ is not an artifact of the proof techniques, and improving it would require changing the estimator of the risk level.

6 Conclusion

We established nonasymptotic oracle guarantees for ordinary empirical risk minimization with Lipschitz losses under weak moment assumptions, including for norm-constrained feedforward neural networks. The oracle-centered bounds require no moment assumptions on the response noise, while Lipschitz surrogate training extends the framework to non-Lipschitz target risks through an explicit estimation–approximation tradeoff. Together with the sharpness results, this provides a statistical basis for robust loss design under heavy-tailed data. Extending these guarantees to dependent data and incorporating optimization error are natural directions for further work.

AI USE STATEMENT

In this work, we have used AI tools for coding assistance, text polishing and navigating related literature.

We have not used generative AI tools for any of the following tasks: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

We have reviewed all AI-assisted work. LLM-assisted code was verified and tested for correctness, suggested references were manually checked and we proofread the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4331–4339, 2019. doi: 10.1109/CVPR.2019.00446.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, and Nassir Navab. Robust optimization for deep regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2830–2838, 2015. doi: 10.1109/ICCV.2015.324.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507 – 2536, 2015. doi: 10.1214/15-AOS1350. URL https://doi.org/10.1214/15-AOS1350.

Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with lipschitz and convex loss functions: G. chinot et al. *Probability Theory and related fields*, 176(3):897–940, 2020.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020. Theorem numbering as in arXiv:1712.06541.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 2011. doi: 10.1002/9781118186435.

Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, 2 edition, 2009. doi: 10.1002/9780470434697.

Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. Wavemixsr: Resource-efficient neural network for image super-resolution. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5872–5880. IEEE, 2024.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 2017.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 1991.

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

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989.

Iosif Pinelis. Best possible bounds of the von bahr–esseen type. *Annals of Functional Analysis*, 6 (4):1–29, 2015.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

Stephen M. Stigler. The changing history of robustness. *The American Statistician*, 64(4):277–281, 2010. doi: 10.1198/tast.2010.10159.

Mahsa Taheri, Fang Xie, and Johannes Lederer. Statistical guarantees for regularized neural networks. *Neural Networks*, 142:148–161, 2021.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

A Proofs

A.1 Proofs of Theorems 2 and 3

Throughout these sections, we make use of the notation $\mathbb{E}_r$ to denote the expectation with respect to the random signs only, the data being held fixed. More formally,

$$\mathbb{E}_r\left[\,\cdot\,\right] := \frac{1}{2^n}\sum_{s\in\{-1,1\}^n} (\,\cdot\,)\,\big|_{r=s}.$$

Since r is independent of the data, and since every integrand to which $\mathbb{E}_r$ is applied below is non-negative, Tonelli's theorem gives $\mathbb{E}[\,\cdot\,] = \mathbb{E}\big[\mathbb{E}_r[\,\cdot\,]\big]$.

We will need intermediate results.

For $\mathfrak{f} \in \mathcal{F}$, define the excess loss

$$g_{\mathfrak{f}}(y, \boldsymbol{x}) := \mathfrak{h}[y - \mathfrak{f}[\boldsymbol{x}]] - \mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]. \tag{12}$$

We write $\mathbb{E}[g_{\mathfrak{f}}] := E_{y,\boldsymbol{x}}[g_{\mathfrak{f}}(y, \boldsymbol{x})]$ for its fresh-copy expectation, which equals $\mathbb{E}[g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i)]$ for every i when the data are i.i.d.

Lemma 1 (Symmetrization and contraction)**.** *Let the data process $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be i.i.d and let $\mathfrak{h}$ satisfy equation LIP. Then,*

$$\mathbb{E}\left[\sup_{\mathfrak{f}\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i) - \mathbb{E}[g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i)]\right|\right] \le 4c_{\mathfrak{h}}\mathfrak{R}_n.$$

Proof. The proof follows from two steps: the first one is a symmetrization step, so we can replace the centering $\mathbb{E}[g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i)]$ by random signs; the second one is a contraction principle that leverages the fact that our loss function $\mathfrak{h}$ is Lipschitz.

As observed in Remark 1, the supremum over $\mathcal{F}$ can be replaced by the supremum over the countable class $\mathcal{F}_0$ without changing its value; for this reason, we will simply write $\sup_{\mathfrak{f}}$. Moreover, the excess losses are integrable, since $\mathbb{E}|g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i)| \leq c_{\mathfrak{h}} v_{\mathfrak{f}}^{(1)} \leq c_{\mathfrak{h}} n \mathfrak{R}_n < \infty$ (Section 2), and the data are i.i.d., so we can apply Lemma 4 to obtain

$$\mathbb{E}\left[\sup_{\mathfrak{f}} \left| \frac{1}{n} \sum_{i=1}^{n} g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i) - \mathbb{E}[g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i)] \right|\right] \leq 2\mathbb{E}\left[\sup_{\mathfrak{f}} \left| \frac{1}{n} \sum_{i=1}^{n} r_i g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i) \right|\right]. \tag{13}$$

Now, in order to use the contraction principle, the trick is to build a contraction from $g_{\mathfrak{f}}$. For each i and fixed $(y_i, \boldsymbol{x}_i)$, define

$$\psi_i(u) := \frac{1}{c_{\mathfrak{h}}} \big(\mathfrak{h}[y_i - \mathfrak{f}^*[\boldsymbol{x}_i] - u] - \mathfrak{h}[y_i - \mathfrak{f}^*[\boldsymbol{x}_i]] \big). \tag{14}$$

From the definition, we have $\psi_i(0) = 0$ and since $\mathfrak{h}$ is $c_{\mathfrak{h}}$-Lipschitz,

$$|\psi_i(u) - \psi_i(v)| = \frac{1}{c_{\mathfrak{h}}} |\mathfrak{h}[y_i - \mathfrak{f}^*[\boldsymbol{x}_i] - u] - \mathfrak{h}[y_i - \mathfrak{f}^*[\boldsymbol{x}_i] - v]| \leq |u - v|.$$

Moreover, we also have the following relation between ψ_i and $g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i)$

$$g_{\mathfrak{f}}(y_i, \boldsymbol{x}_i) = c_{\mathfrak{h}} \psi_i(\mathfrak{f}[\boldsymbol{x}_i] - \mathfrak{f}^*[\boldsymbol{x}_i]). \tag{15}$$

The next step is then an application of Theorem 5 that will allow us to bound the right hand side of equation 13 by the Rademacher complexity. In order to do that, for a fixed $\mathfrak{f} \in \mathcal{F}$ and a data point $(y_i, \boldsymbol{x}_i)$, let $t_{\mathfrak{f}}(i)$ be

$$t_{\mathfrak{f}}(i) := \mathfrak{f}[\boldsymbol{x}_i] - \mathfrak{f}^*[\boldsymbol{x}_i].$$

And for a fixed set of n data points, let $T \subset \mathbb{R}^n$ be the following set

$$T := \{ t_{\mathfrak{f}} = (t_{\mathfrak{f}}(1), \ldots, t_{\mathfrak{f}}(n)) \, : \, \mathfrak{f} \in \mathcal{F}_0 \}. \tag{16}$$

In order to apply Theorem 5, we need to show that T is bounded. This is the statement of the following

Claim. *Let T be as in equation 16. Then, T is bounded $\mathbb{P}$-almost surely.*

Proof of Claim. Fix $r \in \{-1, 1\}^n$, and let $r^{(i)}$ be the sign vector obtained from r by flipping the sign of the i-th coordinate of r. Then, $r - r^{(i)} = 2r_i e_i$, where e_i is the i-th canonical vector of the canonical base of $\mathbb{R}^n$. Then,

$$2|t_{\mathfrak{f}}(i)| = |\langle r - r^{(i)}, t_{\mathfrak{f}} \rangle| \leq \sup_{\mathfrak{f}} |\langle r, t_{\mathfrak{f}} \rangle| + \sup_{\mathfrak{f}} |\langle r^{(i)}, t_{\mathfrak{f}} \rangle|.$$

Now, taking the supremum over $\mathfrak{f}$ again, and then the expectation with respect to the random signs r, we obtain

$$\sup_{\mathfrak{f}} |t_{\mathfrak{f}}(i)| \leq \mathbb{E}_r \left[\sup_{\mathfrak{f}} |\langle r, t_{\mathfrak{f}} \rangle| \right] < \infty, \tag{17}$$

since the map $s \mapsto s^{(i)}$ is a bijection, so it preserves the uniform distribution over the signs. For the finiteness of the right hand side, notice that

$$\mathbb{E}_r \left[\sup_{\mathfrak{f}} |\langle r, t_{\mathfrak{f}} \rangle| \right] = \mathbb{E}_r \left[\sup_{\mathfrak{f}} \left| \sum_{i=1}^{n} r_i (\mathfrak{f}^*[\boldsymbol{x}_i] - \mathfrak{f}[\boldsymbol{x}_i]) \right| \right]$$

and $\mathfrak{R}_n < \infty$. Thus, when integrated with respect to the data points, the positive random variable $\mathbb{E}_r \left[\sup_{\mathfrak{f}} |\langle r, t_{\mathfrak{f}} \rangle| \right]$ has finite expectation, which implies that $\mathbb{E}_r \left[\sup_{\mathfrak{f}} |\langle r, t_{\mathfrak{f}} \rangle| \right]$ itself must be finite $\mathbb{P}$-a.s. This is enough to show that T is bounded almost surely, which proves the claim. ■

Now we know T is bounded, we can apply the contraction Theorem 5, which gives us

$$\mathbb{E}_r\left[\sup_{\mathfrak{f}}\left|\sum_{i=1}^{n} r_i\psi_i(t_{\mathfrak{f}}(i))\right|\right] \leq 2\mathbb{E}_r\left[\sup_{\mathfrak{f}}\left|\sum_{i=1}^{n} r_i(\mathfrak{f}^*[\boldsymbol{x}_i]-\mathfrak{f}[\boldsymbol{x}_i])\right|\right].$$

Taking the expectation with respect to the data points now (recall that $\mathbb{E}=\mathbb{E}\circ\mathbb{E}_r$, the integrands being nonnegative) leads us to

$$\mathbb{E}\left[\sup_{\mathfrak{f}}\left|\sum_{i=1}^{n} r_i\psi_i(t_{\mathfrak{f}}(i))\right|\right] \leq 2\mathbb{E}\left[\sup_{\mathfrak{f}}\left|\sum_{i=1}^{n} r_i(\mathfrak{f}^*[\boldsymbol{x}_i]-\mathfrak{f}[\boldsymbol{x}_i])\right|\right] = 2n\mathfrak{R}_n.$$

Finally, combining the above bound with equation 15 and equation 13 gives us

$$\mathbb{E}\left[\sup_{\mathfrak{f}}\left|\frac{1}{n}\sum_{i=1}^{n} g_{\mathfrak{f}}(y_i,\boldsymbol{x}_i)-\mathbb{E}[g_{\mathfrak{f}}(y_i,\boldsymbol{x}_i)]\right|\right] \leq 4c_{\mathfrak{h}}\mathfrak{R}_n,$$

as desired. □

We can now prove Theorem 3.

Proof of Theorem 3. The proof follows from Lemma 1, a Markov inequality and the fact that $\widehat{\mathfrak{f}}$ is the empirical-risk minimizer. We begin by defining the nonnegative random variable

$$\bar{G} := \sup_{\mathfrak{f}\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g_{\mathfrak{f}}(y_i,\boldsymbol{x}_i)-\mathbb{E}[g_{\mathfrak{f}}]\right|,$$

so that $\mathbb{E}[\bar{G}] \leq 4c_{\mathfrak{h}}\mathfrak{R}_n$ by Lemma 1. We may assume $\mathfrak{R}_n > 0$: if $\mathfrak{R}_n = 0$, then $v_{\mathfrak{f}}^{(1)} = 0$ for every $\mathfrak{f}\in\mathcal{F}$ (Section 2), so every $\mathfrak{f}\in\mathcal{F}_0$, hence, by pointwise limits, every $\mathfrak{f}\in\mathcal{F}$, coincides with $\mathfrak{f}^*$ almost surely under the input distribution; then all excess losses $g_{\mathfrak{f}}(y_i,\boldsymbol{x}_i)$ and all $\mathbb{E}[g_{\mathfrak{f}}]$ vanish almost surely, $\bar{G}=0$ almost surely, and the claim is trivial. By Markov's inequality,

$$\mathbb{P}\left(\bar{G} \geq \frac{4c_{\mathfrak{h}}\mathfrak{R}_n}{t}\right) \leq \frac{t\,\mathbb{E}[\bar{G}]}{4c_{\mathfrak{h}}\mathfrak{R}_n} \leq t.$$

On the complementary event that holds for all functions in $\mathcal{F}$ simultaneously, since $\widehat{\mathfrak{f}}$ minimizes the empirical risk and $\mathfrak{f}^*\in\mathcal{F}$,

$$\begin{aligned} E_{y,\boldsymbol{x}}\left[g_{\widehat{\mathfrak{f}}}(y,\boldsymbol{x})\right] &= \left(\mathbb{E}[g_{\mathfrak{f}}]-\frac{1}{n}\sum_{i=1}^{n} g_{\mathfrak{f}}(y_i,\boldsymbol{x}_i)\right)\Bigg|_{\mathfrak{f}=\widehat{\mathfrak{f}}}+\frac{1}{n}\sum_{i=1}^{n} g_{\widehat{\mathfrak{f}}}(y_i,\boldsymbol{x}_i) \\ &\leq \bar{G}+0 < \frac{4c_{\mathfrak{h}}\mathfrak{R}_n}{t}, \end{aligned}$$

which concludes the proof.

□

Finally, we finish this subsection with the proof of Theorem 2.

Proof of Theorem 2. For any $\mathfrak{f}\in\mathcal{F}$, decompose

$$\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}[y_i-\mathfrak{f}[\boldsymbol{x}_i]]-\mathbb{E}[\mathfrak{h}[y-\mathfrak{f}[\boldsymbol{x}]]] = \left(\frac{1}{n}\sum_{i=1}^{n} g_{\mathfrak{f}}(y_i,\boldsymbol{x}_i)-\mathbb{E}[g_{\mathfrak{f}}]\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}[y_i-\mathfrak{f}^*[\boldsymbol{x}_i]]-\mathbb{E}[\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]]\right),$$

so that, with $\bar{G}$ as in the previous proof,

$$\sup_{\mathfrak{f}\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}[y_i-\mathfrak{f}[\boldsymbol{x}_i]]-\mathbb{E}[\mathfrak{h}[y-\mathfrak{f}[\boldsymbol{x}]]]\right| \leq \bar{G}+\left|\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}[y_i-\mathfrak{f}^*[\boldsymbol{x}_i]]-\mathbb{E}[\mathfrak{h}[y-\mathfrak{f}^*[\boldsymbol{x}]]]\right| =: \bar{G}+D.$$

We may assume $\mathfrak{R}_n > 0$ and $s^{(p)}_{y|\boldsymbol{x}} > 0$: if $\mathfrak{R}_n = 0$, then $\bar{G} = 0$ almost surely (as shown in the previous proof) and the first term may be ignored; if $s^{(p)}_{y|\boldsymbol{x}} = 0$, then $y = \mathfrak{f}^*[\boldsymbol{x}]$ almost surely, so $\mathfrak{h}[y_i - \mathfrak{f}^*[\boldsymbol{x}_i]] = \mathfrak{h}[0] = 0$ almost surely and $D = 0$, and the second term may be ignored. By Markov's inequality and Lemma 1, $\mathbb{P}\big(\bar{G} \geq 8c_{\mathfrak{h}}\mathfrak{R}_n/t\big) \leq t/2$.

For the second term, let

$$Z_i := \mathfrak{h}[y_i - \mathfrak{f}^*[\boldsymbol{x}_i]] - E_{y,\boldsymbol{x}}[\mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]]\,.$$

The variables $Z_1, \ldots, Z_n$ are i.i.d. and centered. Moreover, since $\mathfrak{h}[0] = 0$ and $\mathfrak{h}$ is $c_{\mathfrak{h}}$-Lipschitz,

$$(\mathbb{E}|Z_i|^p)^{1/p} \leq 2\left(E_{y,\boldsymbol{x}}\big|\mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\big|^p\right)^{1/p} \leq 2c_{\mathfrak{h}}s^{(p)}_{y|\boldsymbol{x}}.$$

Hence, by the von Bahr–Esseen inequality,

$$\mathbb{E}\left|\sum_{i=1}^n Z_i\right|^p \leq 2^{2-p}\sum_{i=1}^n \mathbb{E}|Z_i|^p \leq 4nc_{\mathfrak{h}}^p {s^{(p)}_{y|\boldsymbol{x}}}^p.$$

Therefore, Markov's inequality gives

$$\mathbb{P}\left(D \geq 2^{\frac{2}{p}}c_{\mathfrak{h}}\frac{s^{(p)}_{y|\boldsymbol{x}}}{\sqrt[p]{n^{p-1}(t/2)}}\right) \leq \frac{t}{2}.$$

Since $2^{2/p}\,2^{1/p} = 2^{3/p}$, a union bound over the two events concludes the proof.

□

A.2 Proof of Corollary 2

Proof of Corollary 2. Let $\mathcal{F}_0 \subset \mathcal{F}$ be the countable subclass of Remark 1, let $g_{\mathfrak{f}}$ be the excess losses equation 12, and let

$$S := \Big\{(y', \boldsymbol{x}') \in \mathbb{R} \times \mathbb{R}^d : \sup_{\mathfrak{f}\in\mathcal{F}_0}\big|g_{\mathfrak{f}}(y', \boldsymbol{x}')\big| \leq D_g\Big\},$$

a measurable set, being a countable intersection of measurable sets. If $\mathfrak{f} \in \mathcal{F}$ is the pointwise limit of a sequence $(\mathfrak{f}_k)$ in $\mathcal{F}_0$, then $g_{\mathfrak{f}_k} \to g_{\mathfrak{f}}$ pointwise by continuity of $\mathfrak{h}$, so the supremum over $\mathcal{F}_0$ in the definition of S equals the supremum over $\mathcal{F}$, and $\mathbb{P}((y, \boldsymbol{x}) \in S) = 1$ by assumption; hence also $\mathbb{P}((y_i, \boldsymbol{x}_i) \in S) = 1$ for every i. Modifying the data on a null set affects none of the probabilities in the statement, so we assume that $(y_i, \boldsymbol{x}_i) \in S$ for every i. For $z = (z_1, \ldots, z_n) \in S^n$ define

$$Z(z) := \sup_{\mathfrak{f}\in\mathcal{F}_0}\left|\frac{1}{n}\sum_{i=1}^n g_{\mathfrak{f}}(z_i) - \mathbb{E}[g_{\mathfrak{f}}]\right|,$$

a measurable function of z, being a countable supremum. For $z' \in S$ and $\mathfrak{f} \in \mathcal{F}$ we have $|g_{\mathfrak{f}}(z')| \leq D_g$, and therefore also $|\mathbb{E}[g_{\mathfrak{f}}]| \leq D_g$, since $(y, \boldsymbol{x}) \in S$ almost surely. Moreover, the random variable $Z := Z\big((y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\big)$ coincides with the supremum $\bar{G}$ over $\mathcal{F}$ defined in the proof of Theorem 3: for $\mathfrak{f} \in \mathcal{F}$ and $(\mathfrak{f}_k) \subset \mathcal{F}_0$ with $\mathfrak{f}_k \to \mathfrak{f}$ pointwise, the empirical averages of $g_{\mathfrak{f}_k}$ converge to that of $g_{\mathfrak{f}}$, and $\mathbb{E}[g_{\mathfrak{f}_k}] \to \mathbb{E}[g_{\mathfrak{f}}]$ by dominated convergence, the $g_{\mathfrak{f}_k}$ being bounded by D_g on S. Hence $\mathbb{E}[Z] \leq 4c_{\mathfrak{h}}\mathfrak{R}_n$ by Lemma 1. If $z, z' \in S^n$ differ in the i-th coordinate only, then for every $\mathfrak{f} \in \mathcal{F}_0$,

$$\left|\frac{1}{n}\sum_{j=1}^n g_{\mathfrak{f}}(z_j) - \frac{1}{n}\sum_{j=1}^n g_{\mathfrak{f}}(z'_j)\right| = \frac{|g_{\mathfrak{f}}(z_i) - g_{\mathfrak{f}}(z'_i)|}{n} \leq \frac{2D_g}{n},$$

and since $\big|\sup_{\mathfrak{f}}|A_{\mathfrak{f}}| - \sup_{\mathfrak{f}}|A'_{\mathfrak{f}}|\big| \leq \sup_{\mathfrak{f}}|A_{\mathfrak{f}} - A'_{\mathfrak{f}}|$ for any two families of reals with finite suprema, $|Z(z) - Z(z')| \leq 2D_g/n$. The data being independent, Theorem 6 applies with $c_i = 2D_g/n$ for all i and gives, for every $\varepsilon > 0$,

$$\mathbb{P}\big(Z - \mathbb{E}[Z] \geq \varepsilon\big) \leq \exp\left(-\frac{2\varepsilon^2}{n\,(2D_g/n)^2}\right) = \exp\left(-\frac{n\varepsilon^2}{2D_g^2}\right).$$

The choice $\varepsilon := D_g\sqrt{2\log(1/t)/n}$ makes the right-hand side equal to t. Hence, with probability at least $1-t$,

$$\bar{G} = Z < \mathbb{E}[Z] + D_g\sqrt{\frac{2\log(1/t)}{n}} \le 4c_{\mathfrak{h}}\mathfrak{R}_n + D_g\sqrt{\frac{2\log(1/t)}{n}},$$

and on this event the identity displayed at the end of the proof of Theorem 3, valid for every realization, since $\widehat{\mathfrak{f}}$ minimizes the empirical risk and $\mathfrak{f}^* \in \mathcal{F}$, gives $E_{y,\boldsymbol{x}}[g_{\widehat{\mathfrak{f}}}(y,\boldsymbol{x})] \le \bar{G}$, which is the claim. The two particular cases follow from $|g_{\mathfrak{f}}(y,\boldsymbol{x})| \le c_{\mathfrak{h}}|\mathfrak{f}[\boldsymbol{x}] - \mathfrak{f}^*[\boldsymbol{x}]|$, by equation LIP, and from $|g_{\mathfrak{f}}| \le \sup\mathfrak{h} - \inf\mathfrak{h}$. □

A.3 Proofs of Theorem 1 and Corollary 1

The proof of Theorem 1 combines the oracle inequality of Theorem 3 with a bound on the Rademacher complexity $\mathfrak{R}_n$ of the network class. We first state and prove an envelope bound for the network class, which also feeds the tail bounds of Section 3.3.

Proposition 4 (Envelope bound)**.** *Let the oracle be $\mathfrak{f}^* = \mathfrak{f}_{\Theta^*}$ for some $\Theta^* \in \mathcal{M}$. Under assumptions equation AL and equation $\mathcal{M}$, it holds that*

$$w_{\mathcal{F}}^{(p)} \le 2a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)}$$

for all $p \ge 1$.

Proof of Proposition 4. The proof follows from the two general observations below

1. Composition of Lipschitz functions is Lipschitz with a constant that is at most the product of the Lipschitz constants;
2. Any linear map is Lipschitz with a constant that is bounded by its Frobenius norm.

Proof of Observation 1. Fix two Lipschitz maps $\phi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ and $\psi : \mathbb{R}^{d_3} \to \mathbb{R}^{d_1}$, with Lipschitz constants given by c_ϕ and c_ψ respectively. That is, ϕ and ψ satisfy

$$\|\phi(x) - \phi(y)\|_2 \le c_\phi\|x-y\|_2 \quad \text{and} \quad \|\psi(x) - \psi(y)\|_2 \le c_\psi\|x-y\|_2,$$

where the ℓ-2 norms are taken in the respective spaces. Now, observe that, for any $x, y \in \mathbb{R}^{d_3}$

$$\|\phi\circ\psi(x) - \phi\circ\psi(y)\|_2 \le c_\phi\|\psi(x)-\psi(y)\|_2 \le c_\phi c_\psi\|x-y\|_2$$

Proof of Observation 2. A shorter way to see this is to recall that $\|\Theta\|_2 \le \|\Theta\|_F$, and Θ is $\|\Theta\|_2$-Lipschitz. We opt here for a self-contained proof. Fixed any vector $x \in \mathbb{R}^d$ and a linear map $\Theta : \mathbb{R}^d \to \mathbb{R}^m$, the i-th entry of Θx is given by

$$\langle \Theta_i, x\rangle,$$

where Θ_i is the i-th row of Θ and $\langle\cdot,\cdot\rangle$ is the euclidean inner product. Then, by Cauchy-Schwarz's inequality,

$$|(\Theta x)_i|^2 = |\langle\Theta_i, x\rangle|^2 \le \|\Theta_i\|_2^2\|x\|_2^2.$$

Thus,

$$\|\Theta x\|_2^2 = \sum_{i=1}^m |(\Theta x)_i|^2 \le \|x\|_2^2\sum_{i=1}^m\|\Theta_i\|_2^2 = \|x\|_2^2\|\Theta\|_F^2,$$

which gives what we want since $\|\Theta x - \Theta y\|_2 = \|\Theta(x-y)\|_2$.

Finally, by Observation 1 and assumption equation AL, any feedforward neural network $\mathfrak{f}_\Theta$ is the composition of $2\ell+1$ Lipschitz maps. Thus, $\mathfrak{f}_\Theta$ itself is c_Θ-Lipschitz with constant bounded by

$$c_\Theta \le a_{\mathrm{Lip}}^{\ell}\prod_{j=0}^{\ell}\|\Theta^j\|_F \overset{\text{equation } \mathcal{M}}{\le} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1}.$$

On the other hand, since $\mathfrak{a}^j[\mathbf{0}] = \mathbf{0}$ for all activation functions, and Θ^j are linear maps, all our neural networks satisfy $\mathfrak{f}_\Theta[\mathbf{0}] = 0$, which implies that

$$|\mathfrak{f}_\Theta[\boldsymbol{x}]| = |\mathfrak{f}_\Theta[\boldsymbol{x}] - \mathfrak{f}_\Theta[\mathbf{0}]| \le a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} \|\boldsymbol{x}\|_2 \qquad \text{for all } \Theta \in \mathcal{M} \text{ and } \boldsymbol{x} \in \mathbb{R}^d. \tag{18}$$

Hence, by the triangle inequality and since $\Theta^* \in \mathcal{M}$,

$$\sup_{\Theta \in \mathcal{M}} |\mathfrak{f}_\Theta[\boldsymbol{x}] - \mathfrak{f}_{\Theta^*}[\boldsymbol{x}]| \le 2 a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} \|\boldsymbol{x}\|_2 \qquad \text{pointwise in } \boldsymbol{x}.$$

Raising this inequality to the power p, taking expectations, and recalling that $s_{\boldsymbol{x}}^{(p)} = \left(E_{y,\boldsymbol{x}}\left[\|\boldsymbol{x}\|_2^p\right]\right)^{\frac{1}{p}}$, we obtain that, for every $p \ge 1$,

$$w_{\mathcal{F}}^{(p)} = \left(E_{y,\boldsymbol{x}} \left[\sup_{\Theta \in \mathcal{M}} |\mathfrak{f}_\Theta[\boldsymbol{x}] - \mathfrak{f}_{\Theta^*}[\boldsymbol{x}]|^p \right] \right)^{\frac{1}{p}} \le 2 a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)},$$

as desired. □

Proposition 5 (Rademacher complexity of the network class)**.** *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ have the same distribution as $\boldsymbol{x}$, and let the oracle be $\mathfrak{f}^* = \mathfrak{f}_{\Theta^*}$ for some $\Theta^* \in \mathcal{M}$. Under assumptions equation $\mathcal{M}$ and equation AH, it holds for every $p \in (1, 2]$ that*

$$\mathfrak{R}_n \le \left(\sqrt{2\log(2)(\ell+1)} + 1 + 2^{\frac{2-p}{p}} \right) a_{\mathrm{Lip}}^{\ell}\, b_{\mathcal{M}}^{\ell+1}\, \frac{s_{\boldsymbol{x}}^{(p)}}{n^{\frac{p-1}{p}}} \;=\; \kappa_{\ell,p}\, a_{\mathrm{Lip}}^{\ell}\, b_{\mathcal{M}}^{\ell+1}\, \frac{s_{\boldsymbol{x}}^{(p)}}{n^{\frac{p-1}{p}}}.$$

Proof of Proposition 5. We may assume $a_{\mathrm{Lip}} > 0$ and $s_{\boldsymbol{x}}^{(p)} < \infty$: if $a_{\mathrm{Lip}} = 0$ (and $\ell \ge 1$), all activations vanish identically, so $\mathfrak{f}_\Theta \equiv 0$ for every Θ and $\mathfrak{R}_n = 0$; if $s_{\boldsymbol{x}}^{(p)} = \infty$, the bound is trivial. Write $\bar{\mathcal{M}}_b := \{\Theta = (\Theta^\ell, \dots, \Theta^0) : \|\|\Theta^j\|\|_F \le b_{\mathcal{M}} \text{ for all } j\}$ for the full Frobenius ball, so that $\mathcal{M} \subset \bar{\mathcal{M}}_b$ by equation $\mathcal{M}$. By the triangle inequality,

$$\mathfrak{R}_n \le \mathbb{E}\left[\sup_{\Theta \in \mathcal{M}} \left| \frac{1}{n} \sum_{i=1}^n r_i\, \mathfrak{f}_\Theta[\boldsymbol{x}_i] \right| \right] + \mathbb{E}\left[\left| \frac{1}{n} \sum_{i=1}^n r_i\, \mathfrak{f}^*[\boldsymbol{x}_i] \right| \right]. \tag{19}$$

First term. Since $\mathcal{M} \subset \bar{\mathcal{M}}_b$, and since $\bar{\mathcal{M}}_b$ is invariant under the map $\Theta^\ell \mapsto -\Theta^\ell$, which leaves all Frobenius norms unchanged and maps $\mathfrak{f}_\Theta$ to $-\mathfrak{f}_\Theta$, we have

$$\sup_{\Theta \in \mathcal{M}} \left| \sum_{i=1}^n r_i\, \mathfrak{f}_\Theta[\boldsymbol{x}_i] \right| \le \sup_{\Theta \in \bar{\mathcal{M}}_b} \left| \sum_{i=1}^n r_i\, \mathfrak{f}_\Theta[\boldsymbol{x}_i] \right| = \sup_{\Theta \in \bar{\mathcal{M}}_b} \sum_{i=1}^n r_i\, \mathfrak{f}_\Theta[\boldsymbol{x}_i].$$

By equation AH, each activation is σ_j applied coordinatewise. Writing $\mathfrak{a}^j = a_{\mathrm{Lip}}\, \tilde{\mathfrak{a}}^j$ with $\tilde{\mathfrak{a}}^j$ the coordinatewise application of $\sigma_j / a_{\mathrm{Lip}}$ (a 1-Lipschitz, positively homogeneous scalar function), positive homogeneity allows us to pull the accumulated scalar factors out through all layers, by induction over the layers, $\tilde{\mathfrak{a}}^j[\lambda u] = \lambda\, \tilde{\mathfrak{a}}^j[u]$ for $\lambda = a_{\mathrm{Lip}}^{j-1} \ge 0$ passes the factor collected so far through the j-th activation, giving $\mathfrak{f}_\Theta = a_{\mathrm{Lip}}^{\ell}\, \mathfrak{g}_\Theta$, where $\mathfrak{g}_\Theta$ is the network with the same weight matrices and activations $\tilde{\mathfrak{a}}^j$. The class $\{\mathfrak{g}_\Theta : \Theta \in \bar{\mathcal{M}}_b\}$ therefore consists of networks with $\ell + 1$ weight matrices, each of Frobenius norm at most $b_{\mathcal{M}}$, and 1-Lipschitz positively homogeneous element-wise activations. Hence, conditionally on the data, Theorem 1 of Golowich et al. (2020) gives

$$\mathbb{E}_r\left[\sup_{\Theta \in \bar{\mathcal{M}}_b} \frac{1}{n} \sum_{i=1}^n r_i\, \mathfrak{g}_\Theta[\boldsymbol{x}_i] \right] \le \frac{b_{\mathcal{M}}^{\ell+1} \left(\sqrt{2\log(2)(\ell+1)} + 1 \right)}{n} \sqrt{\sum_{i=1}^n \|\boldsymbol{x}_i\|_2^2}.$$

(Their Theorem 1 is stated over a bounded input domain, but the empirical inequality above does not involve the domain bound and may in any case be applied realization-wise with the bound $\max_i \|\boldsymbol{x}_i\|_2 < \infty$; their Lemma 1 is applied once per layer, so layer-dependent activations σ_j are covered; and their Rademacher complexity is the one-sided supremum, matching the display.) Since $\|v\|_2 \le \|v\|_p$ on $\mathbb{R}^n$ for $p \le 2$, we have $\sqrt{\sum_i \|\boldsymbol{x}_i\|_2^2} \le \left(\sum_i \|\boldsymbol{x}_i\|_2^p\right)^{1/p}$, and by Jensen's inequality (concavity of $u \mapsto u^{1/p}$) since $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are identically distributed,

$$\mathbb{E}\left[\left(\sum_{i=1}^n \|\boldsymbol{x}_i\|_2^p \right)^{\frac{1}{p}} \right] \le \left(\sum_{i=1}^n \mathbb{E}\left[\|\boldsymbol{x}_i\|_2^p\right] \right)^{\frac{1}{p}} = n^{\frac{1}{p}}\, s_{\boldsymbol{x}}^{(p)}.$$

Altogether, the first term in equation 19 is at most $\left(\sqrt{2\log(2)(\ell+1)}+1\right) a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)} n^{-\frac{p-1}{p}}$.

Second term. The variables $\xi_i := r_i \mathfrak{f}^*[\boldsymbol{x}_i]$ are centered with finite p-th moments (by equation 18 and $s_{\boldsymbol{x}}^{(p)} < \infty$). Since the Rademacher vector $(r_1, \ldots, r_n)$ is independent of the data process and has independent coordinates, r_{i+1} is independent of the sigma algebra $\mathcal{H}_i := \sigma\big(r_1, \ldots, r_i, \{(y_j, \boldsymbol{x}_j)\}_{j=1}^n\big)$, which contains $\sigma(\xi_1 + \cdots + \xi_i)$ and makes $\mathfrak{f}^*[\boldsymbol{x}_{i+1}]$ measurable; hence $\mathbb{E}[\xi_{i+1} \mid \mathcal{H}_i] = \mathfrak{f}^*[\boldsymbol{x}_{i+1}]\,\mathbb{E}[r_{i+1}] = 0$, and the tower property gives $\mathbb{E}[\xi_{i+1} \mid \xi_1 + \cdots + \xi_i] = 0$ for every i, regardless of any dependence within the data. The von Bahr–Esseen inequality (Theorem 7) together with Jensen's inequality thus gives

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right|\right] \le \left(\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right|^p\right]\right)^{\frac{1}{p}} \le \frac{1}{n}\left(2^{2-p}\sum_{i=1}^n \mathbb{E}\big[|\mathfrak{f}^*[\boldsymbol{x}_i]|^p\big]\right)^{\frac{1}{p}} \le 2^{\frac{2-p}{p}} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)} n^{-\frac{p-1}{p}},$$

where the last step uses equation 18 (recall $\Theta^* \in \mathcal{M}$) and and the fact that the $\boldsymbol{x}_i$ are identically distributed. Adding the two bounds concludes the proof. □

Proof of Theorem 1. Since equation AH implies equation AL, and since the data are i.i.d., Theorem 3 applies and gives

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[\mathfrak{h}[y - \widehat{\mathfrak{f}}[\boldsymbol{x}]] - \mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\right] \le \frac{4c_{\mathfrak{h}}\mathfrak{R}_n}{t}\right) \ge 1 - t,$$

and Proposition 5 gives $\mathfrak{R}_n \le \kappa_{\ell,p} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)} n^{-(p-1)/p}$; substituting this bound inside the event, which only enlarges it, proves the claim. Finally, the bound on the integrand displayed after the theorem follows from equation LIP, equation 18 and the triangle inequality, since $\widehat{\mathfrak{f}} = \mathfrak{f}_{\widehat{\Theta}}$ and $\mathfrak{f}^* = \mathfrak{f}_{\Theta^*}$ with $\widehat{\Theta}, \Theta^* \in \mathcal{M}$. □

Proof of Corollary 1. Recall that equation AH implies equation AL, so that equation 18 in the proof of Proposition 4 is available. If $a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} = 0$, then equation 18 gives $\mathfrak{f}_\Theta \equiv 0$ for every $\Theta \in \mathcal{M}$, so $\widehat{\mathfrak{f}} = \mathfrak{f}^*$ and the claim is trivial; assume henceforth $a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} > 0$ and set $D := 2a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} B \in (0, \infty)$. By equation 18 and the triangle inequality, since $\Theta, \Theta^* \in \mathcal{M}$,

$$\sup_{\Theta \in \mathcal{M}} \left|\mathfrak{f}_\Theta[\boldsymbol{x}] - \mathfrak{f}_{\Theta^*}[\boldsymbol{x}]\right| \le 2a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} \|\boldsymbol{x}\|_2 \le D \qquad \text{almost surely,}$$

so Corollary 2 applies with $D_g = c_{\mathfrak{h}} D$ and gives

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[\mathfrak{h}[y - \widehat{\mathfrak{f}}[\boldsymbol{x}]] - \mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\right] \le 4c_{\mathfrak{h}}\mathfrak{R}_n + 2c_{\mathfrak{h}} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} B\sqrt{\frac{2\log(1/t)}{n}}\right) \ge 1 - t.$$

Since $s_{\boldsymbol{x}}^{(2)} = (\mathbb{E}\|\boldsymbol{x}\|_2^2)^{1/2} \le B$, Proposition 5 with $p = 2$ gives $\mathfrak{R}_n \le \kappa_{\ell,2} a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} B/\sqrt{n}$; substituting this bound inside the event yields the claim. □

A.4 Proofs of the truncation results

Lemma 2 (Lipschitz truncation). *Let $\mathfrak{h} : \mathbb{R} \to [0, \infty)$ be an even convex function in $C^1(\mathbb{R})$, and $\tau > 0$. Then, the function $\mathfrak{h}_\tau$ defined as*

$$\mathfrak{h}_\tau(x) = \begin{cases} \mathfrak{h}(x), & \text{if } |x| \le \tau \\ \mathfrak{h}(\tau) + \mathfrak{h}'(\tau)(|x| - \tau), & \text{otherwise} \end{cases} \tag{20}$$

is a Lipschitz function with Lipschitz constant $\mathfrak{h}'(\tau)$.

Proof of Lemma 2. Being $\mathfrak{h}$ even, convex and C^1, its derivative $\mathfrak{h}'$ is odd and non-decreasing; in particular $\mathfrak{h}'(0) = 0$, $\mathfrak{h}'(\tau) \ge 0$, and $|\mathfrak{h}'(x)| \le \mathfrak{h}'(\tau)$ for all $x \in [-\tau, \tau]$. The function $\mathfrak{h}_\tau$ is continuous at $\pm\tau$ (both branches take the value $\mathfrak{h}(\tau)$ there, using $\mathfrak{h}(-\tau) = \mathfrak{h}(\tau)$) and differentiable everywhere: on $(-\tau, \tau)$ its derivative is $\mathfrak{h}'(x)$; on (τ, ∞) and on $(-\infty, -\tau)$ it is linear with slopes $\mathfrak{h}'(\tau)$ and $-\mathfrak{h}'(\tau)$, respectively; and at $x = \pm\tau$ the one-sided derivatives match ($\mathfrak{h}'(\tau)$ at τ; $-\mathfrak{h}'(\tau) = \mathfrak{h}'(-\tau)$ at $-\tau$, by oddness of $\mathfrak{h}'$). Hence $\mathfrak{h}_\tau \in C^1(\mathbb{R})$ with $|\mathfrak{h}_\tau'| \le \mathfrak{h}'(\tau)$ everywhere, and the Mean Value Theorem yields the Lipschitz property with constant $\mathfrak{h}'(\tau)$. □

Proof of Proposition 1. All expectations below are of nonnegative measurable functions, hence well defined in $[0,\infty]$, and every step of the argument is valid there. Suppose first that $c_{\widetilde{\mathfrak{h}}} = 0$. Then $\widetilde{\mathfrak{h}} \equiv \widetilde{\mathfrak{h}}[0] = 0$, so $R = \mathfrak{h}$, and the claim holds deterministically: $E_{y,\boldsymbol{x}}[\mathfrak{h}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])] = E_{y,\boldsymbol{x}}[R(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])] \le \Phi(\mathfrak{h}, \widetilde{\mathfrak{h}})$, since $\widehat{\mathfrak{f}} \in \mathcal{F}$, and this is at most the right-hand side. Assume henceforth $c_{\widetilde{\mathfrak{h}}} > 0$. Then $\widetilde{\mathfrak{h}}$ satisfies equation LIP with the positive constant $c_{\widetilde{\mathfrak{h}}}$, so Theorem 3 applies to the loss $\widetilde{\mathfrak{h}}$ and its empirical-risk minimizer $\widehat{\mathfrak{f}}$: on an event of probability at least $1 - t$, the excess loss $\widetilde{g}_{\widehat{\mathfrak{f}}}(y, \boldsymbol{x}) := \widetilde{\mathfrak{h}}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}]) - \widetilde{\mathfrak{h}}(y - \mathfrak{f}^*[\boldsymbol{x}])$, which is integrable (Section 2), satisfies $E_{y,\boldsymbol{x}}[\widetilde{g}_{\widehat{\mathfrak{f}}}(y, \boldsymbol{x})] \le 4c_{\widetilde{\mathfrak{h}}}\mathfrak{R}_n/t$. Since $\widetilde{\mathfrak{h}} \ge 0$, this gives, in $[0,\infty]$,

$$E_{y,\boldsymbol{x}}\left[\widetilde{\mathfrak{h}}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])\right] = E_{y,\boldsymbol{x}}\left[\widetilde{g}_{\widehat{\mathfrak{f}}}(y, \boldsymbol{x})\right] + E_{y,\boldsymbol{x}}\left[\widetilde{\mathfrak{h}}(y - \mathfrak{f}^*[\boldsymbol{x}])\right] \le E_{y,\boldsymbol{x}}\left[\widetilde{\mathfrak{h}}(y - \mathfrak{f}^*[\boldsymbol{x}])\right] + \frac{4c_{\widetilde{\mathfrak{h}}}\mathfrak{R}_n}{t}.$$

Using $\mathfrak{h} = \widetilde{\mathfrak{h}} + R$ with $R \ge 0$ and $\widetilde{\mathfrak{h}} \le \mathfrak{h}$, we obtain, on the same event,

$$\begin{aligned} E_{y,\boldsymbol{x}}\left[\mathfrak{h}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])\right] &= E_{y,\boldsymbol{x}}\left[\widetilde{\mathfrak{h}}(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])\right] + E_{y,\boldsymbol{x}}\left[R(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])\right] \\ &\le E_{y,\boldsymbol{x}}\left[\widetilde{\mathfrak{h}}(y - \mathfrak{f}^*[\boldsymbol{x}])\right] + \frac{4c_{\widetilde{\mathfrak{h}}}\mathfrak{R}_n}{t} + \Phi(\mathfrak{h}, \widetilde{\mathfrak{h}}) \\ &\le E_{y,\boldsymbol{x}}\left[\mathfrak{h}(y - \mathfrak{f}^*[\boldsymbol{x}])\right] + \frac{4c_{\widetilde{\mathfrak{h}}}\mathfrak{R}_n}{t} + \Phi(\mathfrak{h}, \widetilde{\mathfrak{h}}), \end{aligned}$$

where the middle inequality combines the bound just derived (for the first summand) with the pointwise-in-the-sample bound $E_{y,\boldsymbol{x}}[R(y - \widehat{\mathfrak{f}}[\boldsymbol{x}])] \le \Phi(\mathfrak{h}, \widetilde{\mathfrak{h}})$, valid since $\widehat{\mathfrak{f}} \in \mathcal{F}$, and the last inequality uses $\widetilde{\mathfrak{h}} \le \mathfrak{h}$. Note that the Rademacher complexity $\mathfrak{R}_n$ depends only on $\mathcal{F}$, $\mathfrak{f}^*$ and the inputs, not on the loss, so the *same* quantity appears for every surrogate. □

Proof of Proposition 2. The truncated loss $\mathfrak{h}_\tau$ is nonnegative with $\mathfrak{h}_\tau(0) = \mathfrak{h}(0) = 0$ and satisfies $\mathfrak{h}_\tau \le \mathfrak{h}$, as shown after equation 7, and it is Lipschitz with the constant $\mathfrak{h}'(\tau) \in [0,\infty)$ by Lemma 2. Proposition 1, applied with $\widetilde{\mathfrak{h}} = \mathfrak{h}_\tau$, $c_{\widetilde{\mathfrak{h}}} = \mathfrak{h}'(\tau)$ and $R = R_\tau$, for which $\Phi(\mathfrak{h}, \mathfrak{h}_\tau) = \Phi_{\mathfrak{h}}(\tau)$, is the claim. □

Lemma 3 (Tail bound)**.** *Let $\mathfrak{h}$ be as in Lemma 2. Under equation PG and equation HM with $q > r$, for every $\tau > 0$,*

$$\Phi_{\mathfrak{h}}(\tau) \le \frac{C_{\mathfrak{h}} M_q^q}{\tau^{q-r}}.$$

Proof of Lemma 3. Notice that by equation PG and the definition equation 7 of the remainder R_τ, $R_\tau(x) \le \mathfrak{h}[\boldsymbol{x}] \le C_{\mathfrak{h}}|x|^r$ for all x. Since $R_\tau(x) = 0$ for $|x| \le \tau$, we only need the bound for $|x| > \tau$. For $|x| > \tau$, since $\tau^{q-r} \le |x|^{q-r}$,

$$C_{\mathfrak{h}}|x|^r = C_{\mathfrak{h}}\frac{|x|^r \cdot \tau^{q-r}}{\tau^{q-r}} \le \frac{C_{\mathfrak{h}}|x|^q}{\tau^{q-r}}.$$

Thus $R_\tau(x) \le C_{\mathfrak{h}}|x|^q/\tau^{q-r}$ for all $x \in \mathbb{R}$.

Now, for any $\mathfrak{f} \in \mathcal{F}$, by the convexity of $u \mapsto |u|^q$, which gives $|a+b|^q \le 2^{q-1}(|a|^q + |b|^q)$,

$$|y - \mathfrak{f}[\boldsymbol{x}]|^q \le 2^{q-1}\left(|y - \mathfrak{f}^*[\boldsymbol{x}]|^q + |\mathfrak{f}^*[\boldsymbol{x}] - \mathfrak{f}[\boldsymbol{x}]|^q\right).$$

Taking expectations and using $E_{y,\boldsymbol{x}}[|\mathfrak{f}^*[\boldsymbol{x}] - \mathfrak{f}[\boldsymbol{x}]|^q] \le w_{\mathcal{F}}^{(q)q}$ gives $E_{y,\boldsymbol{x}}[|y - \mathfrak{f}[\boldsymbol{x}]|^q] \le M_q^q$, so

$$E_{y,\boldsymbol{x}}[R_\tau(y - \mathfrak{f}[\boldsymbol{x}])] \le \frac{C_{\mathfrak{h}} M_q^q}{\tau^{q-r}}.$$

Taking the supremum over $\mathfrak{f} \in \mathcal{F}$ yields the result. □

Corollary 3 (Selection under polynomial-growth losses)**.** *Let $\mathfrak{h}$ be as in Lemma 2 with $\mathfrak{h}[0] = 0$, and assume equation PG and equation HM with $q > r$; let $\mathfrak{f}^* \in \mathcal{F}$ and, for $\tau > 0$, let $\widehat{\mathfrak{f}}_\tau$ minimize the empirical $\mathfrak{h}_\tau$-risk. Then, for every $\tau > 0$ and every $t \in (0,1)$,*

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[\mathfrak{h}[y - \widehat{\mathfrak{f}}_\tau[\boldsymbol{x}]]\right] \le E_{y,\boldsymbol{x}}\left[\mathfrak{h}[y - \mathfrak{f}^*[\boldsymbol{x}]]\right] + \mathfrak{h}'(\tau)\, B_n(t) + \frac{C_{\mathfrak{h}} M_q^q}{\tau^{q-r}}\right) \ge 1 - t, \qquad B_n(t) := \frac{4\mathfrak{R}_n}{t}. \tag{21}$$

For the power losses $\mathfrak{h}(a) = |a|^r$, $r > 1$, which include the squared loss, the level τ can be optimized in closed form. Here $\mathfrak{h}'(\tau) = r\tau^{r-1}$, $C_{\mathfrak{h}} = 1$, and the truncation equation 20 is the generalized Huber loss of order r of Definition 3 below.

Corollary 4 (Power losses). *Let $\mathfrak{h}(a) = |a|^r$ with $r > 1$, assume equation HM with $q > r$, let $\mathfrak{f}^* \in \mathcal{F}$, and let $t \in (0,1)$. If $B_n(t) = 4\mathfrak{R}_n/t > 0$ and $M_q > 0$, then for the level*

$$\tau^* := \left(\frac{(q-r)M_q^q}{r(r-1)B_n(t)} \right)^{\frac{1}{q-1}}, \tag{22}$$

the empirical-risk minimizer $\widehat{\mathfrak{f}}_{\tau^*}$ under $\mathfrak{h}_{\tau^*}$ satisfies

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[|y - \widehat{\mathfrak{f}}_{\tau^*}[\boldsymbol{x}]|^r \right] \le E_{y,\boldsymbol{x}}\left[|y - \mathfrak{f}^*[\boldsymbol{x}]|^r \right] + C(r,q)\big(B_n(t)\big)^{\frac{q-r}{q-1}} \left(M_q^q\right)^{\frac{r-1}{q-1}} \right) \ge 1 - t, \tag{23}$$

where $C(r,q) := (q-1)\, r^{\frac{q-r}{q-1}}\, (r-1)^{-\frac{r-1}{q-1}}\, (q-r)^{-\frac{q-r}{q-1}}$. Moreover, τ^* is the unique minimizer over $\tau \in (0,\infty)$ of the τ-dependent part $r\tau^{r-1}B_n(t) + M_q^q/\tau^{q-r}$ of the bound in Corollary 3.

Proof of Corollary 3. On the event of Proposition 2, replace $\Phi_{\mathfrak{h}}(\tau)$ by its upper bound $C_{\mathfrak{h}} M_q^q/\tau^{q-r}$ from Lemma 3; this only enlarges the event. □

Proof of Corollary 4. The loss $\mathfrak{h}(a) = |a|^r$ with $r > 1$ is even, convex, C^1, nonnegative, and satisfies $\mathfrak{h}[0] = 0$, $\mathfrak{h}'(\tau) = r\tau^{r-1}$, and equation PG with $C_{\mathfrak{h}} = 1$; its Lipschitz truncation equation 20 at level τ is the generalized Huber loss of order r of Definition 3. By Corollary 3, for every $\tau > 0$,

$$\mathbb{P}\left(E_{y,\boldsymbol{x}}\left[|y - \widehat{\mathfrak{f}}_{\tau}[\boldsymbol{x}]|^r \right] \le E_{y,\boldsymbol{x}}\left[|y - \mathfrak{f}^*[\boldsymbol{x}]|^r \right] + g(\tau) \right) \ge 1 - t, \qquad g(\tau) := r\tau^{r-1}B_n(t) + \frac{M_q^q}{\tau^{q-r}}.$$

It remains to minimize g over $(0,\infty)$, which is where $B_n(t) > 0$ and $M_q > 0$ enter. Setting to zero the derivative of g,

$$r(r-1)\tau^{r-2}B_n(t) = \frac{(q-r)M_q^q}{\tau^{q-r+1}},$$

yields equation 22. The critical point τ^* is indeed the unique global minimizer of g on $(0,\infty)$: factoring $g'(\tau) = \tau^{-(q-r+1)}\left[r(r-1)B_n(t)\,\tau^{q-1} - (q-r)M_q^q \right]$ shows that the bracket is strictly increasing in τ, so g decreases on $(0,\tau^*)$ and increases on (τ^*,∞) (note that g need not be convex for $r < 2$). Since τ^* is a deterministic level, the display holds at $\tau = \tau^*$, and substituting equation 22 into g and simplifying gives

$$g(\tau^*) = C(r,q)\big(B_n(t)\big)^{\frac{q-r}{q-1}} \left(M_q^q\right)^{\frac{r-1}{q-1}}, \qquad C(r,q) = (q-1)\, r^{\frac{q-r}{q-1}}\, (r-1)^{-\frac{r-1}{q-1}}\, (q-r)^{-\frac{q-r}{q-1}},$$

which is equation 23. □

A.5 Proof of Theorem 4

Proof of Theorem 4. Since equation AH implies equation AL, Proposition 4 applies at order q and gives $w_{\mathcal{F}}^{(q)} \le 2a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(q)} < \infty$; together with $s_{y|\boldsymbol{x}}^{(q)} < \infty$ this is equation HM. Since $s_{\boldsymbol{x}}^{(p)} < \infty$ by assumption, Proposition 5 gives $\mathfrak{R}_n \le \kappa_{\ell,p}\, a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_{\boldsymbol{x}}^{(p)}\, n^{-(p-1)/p}$, that is, $B_n(t) = 4\mathfrak{R}_n/t \le \bar{B}_n(t)$. As noted in the proof of Corollary 4, Corollary 3 applies to $\mathfrak{h}(a) = |a|^r$ with $\mathfrak{h}'(\tau) = r\tau^{r-1}$ and $C_{\mathfrak{h}} = 1$, and gives, for every $\tau > 0$ and $t \in (0,1)$, with probability at least $1 - t$,

$$\mathbb{E}_{y,\boldsymbol{x}}\left[|y - \widehat{\mathfrak{f}}_{\tau}[\boldsymbol{x}]|^r \right] \le \mathbb{E}_{y,\boldsymbol{x}}\left[|y - \mathfrak{f}^*[\boldsymbol{x}]|^r \right] + r\tau^{r-1}\, B_n(t) + \frac{M_q^q}{\tau^{q-r}};$$

as $r\tau^{r-1} \ge 0$, replacing $B_n(t)$ by $\bar{B}_n(t)$ proves equation 10.

For equation 11, note that $\bar{\tau}^*$ is a deterministic level, so equation 10 holds at $\tau = \bar{\tau}^*$ with probability at least $1 - t$, and the right-hand side there equals $\mathbb{E}[|y - \mathfrak{f}^*[\boldsymbol{x}]|^r] + \bar{g}(\bar{\tau}^*)$ with $\bar{g}(\tau) := r\tau^{r-1}\bar{B}_n(t) + M_q^q/\tau^{q-r}$. The function $\bar{g}$ is the function g of the proof of Corollary 4 with $\bar{B}_n(t) \in (0,\infty)$ in place of $B_n(t)$, and $\bar{\tau}^*$ is the expression equation 22 with the same replacement; the computation there, which uses only $\bar{B}_n(t) > 0$ and $M_q > 0$, therefore gives $\bar{g}(\bar{\tau}^*) = C(r,q)(\bar{B}_n(t))^{(q-r)/(q-1)}(M_q^q)^{(r-1)/(q-1)}$, which is equation 11. □

A.6 Analysis of the Barron loss

Robust residual losses are widely used in imaging applications. A popular example is the family introduced by Barron (2019),

$$\mathfrak{h}_{\alpha,\beta}(u) = \frac{|\alpha-2|}{\alpha}\left[\left(1+\frac{(u/\beta)^2}{|\alpha-2|}\right)^{\alpha/2}-1\right],$$

with continuous extensions at $\alpha = 0$ and $\alpha = 2$. The parameter α interpolates between several commonly used loss functions, including Welsch/Leclerc, Geman-McClure, Cauchy/Lorentzian ($\alpha = 0$), generalized Charbonnier, Charbonnier/pseudo-Huber/L_1- L_2 ($\alpha = 1$), and L_2 ($\alpha = 2$).

Lipschitz case. For $\alpha \le 1$, the loss is globally Lipschitz with constant

$$c_{\alpha,\beta} = \begin{cases} \frac{1}{\beta}\left(\frac{1-\alpha}{2-\alpha}\right)^{(1-\alpha)/2} & \alpha < 1 \\ 1/\beta & \alpha = 1 \end{cases} \tag{24}$$

Consequently, Theorem 1 immediately gives the noise-free bound

$$E_{y,\boldsymbol{x}}\left[\mathfrak{h}_{\alpha,\beta}[y-\widehat{\mathfrak{f}}[\boldsymbol{x}]]-\mathfrak{h}_{\alpha,\beta}[y-\mathfrak{f}^*[\boldsymbol{x}]]\right] \le \frac{4\kappa_{\ell,p}\, c_{\alpha,\beta}\, a_{\mathrm{Lip}}^{\ell} b_{\mathcal{M}}^{\ell+1} s_x^{(p)}}{t n^{(p-1)/p}}$$

with probability at least $1-t$. Moreover, the following proposition is an immediate consequence of Theorem 2 and Proposition 5.

Proposition 6 (Uniform generalization gap for deep neural networks under Barron losses)**.** *Let $\mathcal{F}$ be the feedforward neural-network class of Section 2 under equation $\mathcal{M}$ and equation AH, and let $\mathfrak{f}^* = \mathfrak{f}_{\Theta^*}$ for some $\Theta^* \in \mathcal{M}$. Then, for every $t \in (0,1)$, with probability at least $1-t$,*

$$\sup_{\mathfrak{f}\in\mathcal{F}}\left|E_{y,\boldsymbol{x}}\left[\mathfrak{h}_{\alpha,\beta}(y-\mathfrak{f}[\boldsymbol{x}])\right]-\frac{1}{n}\sum_{i=1}^{n}\mathfrak{h}_{\alpha,\beta}(y_i-\mathfrak{f}[\boldsymbol{x}_i])\right| \tag{25}$$

$$\le \frac{8c_{\alpha,\beta}\kappa_{\ell,p}a_{\mathrm{Lip}}^{\ell}b_{\mathcal{M}}^{\ell+1}s_{\boldsymbol{x}}^{(p)}}{t\, n^{(p-1)/p}} + 2^{3/p}c_{\alpha,\beta}\frac{s_{y|\boldsymbol{x}}^{(p)}}{n^{(p-1)/p}t^{1/p}}, \tag{26}$$

where $c_{\alpha,\beta}$ is the Lipschitz constant of $\mathfrak{h}_{\alpha,\beta}$.

Non-Lipschitz case. For $1 < \alpha \le 2$, the loss is no longer globally Lipschitz, but it is even, convex and of polynomial growth of order α. Hence, the results of Section 3.3 apply directly. In particular, for $\tau > 0$ consider its Lipschitz truncation

$$\mathfrak{h}_{\alpha,\beta,\tau}(u) = \begin{cases} \mathfrak{h}_{\alpha,\beta}(u), & |u| \le \tau, \\ \mathfrak{h}_{\alpha,\beta}(\tau) + \mathfrak{h}'_{\alpha,\beta}(\tau)(|u|-\tau), & |u| > \tau. \end{cases}$$

where

$$\mathfrak{h}'_{\alpha,\beta}(\tau) = \frac{\tau}{\beta^2}\left(1+\frac{(\tau/\beta)^2}{2-\alpha}\right)^{\alpha/2-1}, \qquad 1 < \alpha < 2,$$

with $\mathfrak{h}'_{2,\beta}(\tau) = \tau/\beta^2$ at $\alpha = 2$. Moreover, the polynomial growth condition

$$\mathfrak{h}_{\alpha,\beta}(u) \le C_{\alpha,\beta}|u|^{\alpha}.$$

is satisfied, with an explicit $C_{\alpha,\beta} > 0$. Therefore, under the higher-moment assumptions of Corollary 3 with $q > \alpha$, an empirical-risk minimizer $\widehat{\mathfrak{f}}_\tau$ under $\mathfrak{h}_{\alpha,\beta,\tau}$ satisfies

$$E_{y,\boldsymbol{x}}\left[\mathfrak{h}_{\alpha,\beta}(y-\widehat{\mathfrak{f}}_\tau[\boldsymbol{x}])\right] \le E_{y,\boldsymbol{x}}[\mathfrak{h}_{\alpha,\beta}(y-\mathfrak{f}^*[\boldsymbol{x}])] + \mathfrak{h}'_{\alpha,\beta}(\tau)B_n(t) + \frac{C_{\alpha,\beta}M_q^q}{\tau^{q-\alpha}}$$

with probability at least $1-t$. Using $\mathfrak{h}'_{\alpha,\beta}(\tau) \le \alpha C_{\alpha,\beta}\tau^{\alpha-1}$, the resulting polynomial upper bound can also be optimized explicitly. In particular, the same calculation as in Corollary 4, with $r = \alpha$, gives

$$\tau^* = \left(\frac{(q-\alpha)M_q^q}{\alpha(\alpha-1)B_n(t)}\right)^{1/(q-1)}.$$

Quadratic case. The case $\alpha = 2$ coincides with a scaled squared loss, and it can be treated by applying Corollary 4, $r = 2$.

A.7 Details of the sample-size experiment

We illustrate the dependence of the excess-risk bounds on the tail behavior of the covariates and of the response noise. We consider the regression model

$$y = f^\star(\boldsymbol{x}) + \varepsilon,$$

where $f^\star \in \mathcal{F}$ and $\mathcal{F}$ is the class one-hidden-layer ReLU network with input dimension $d = 10$ and hidden width $p_1 = 20$, $b_{\mathcal{M}} = 2$. For each sample size $n \in \{250, 500, 1000, 2000, 4000\}$ we compute an approximate empirical risk minimizer $\widehat{f}$ using the Huber loss.

We consider two input regimes. In the first, the coordinates of x are uniformly distributed on $[-1, 1]$, corresponding to the bounded-input setting of Corollary 1. In the second, the coordinates are heavy-tailed, sampled from $t_{1.25}$, with moments of order $p < 1.25$ finite, corresponding to the setting of Theorem 1. We compare Gaussian response noise $\epsilon \sim N(0, 1)$ with symmetric Pareto noise on $(-\infty, 1] \cup [1, \infty)$ with index 1.5, with finite mean and infinite variance. The networks are trained using full-batch Adam with learning rate 5×10^{-3}, projecting each weight matrix onto the Frobenius ball of radius 2 after every update.

For each configuration, the experiment is repeated over 100 independent datasets.

Performance is evaluated on an independent test sample through

$$\widehat{\mathbb{E}}(\widehat{f}) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \Big\{ h\big(y_i - \widehat{f}(x_i)\big) - h\big(y_i - f^\star(x_i)\big) \Big\}.$$

with $N_{\text{test}} = 10000$.

A.8 Details of the response-corruption experiment

To illustrate the role of Lipschitz surrogate training in Proposition 1 and Theorem 4, we study robust training under heavy-tailed corruption while evaluating performance under the squared loss.

We consider a regression model with bounded covariates $\boldsymbol{x} \in [-1, 1]^{10}$ and target function $f^\star \in \mathcal{F}$ given by a one-hidden-layer ReLU network of width 20, with layerwise Frobenius norms bounded by 2. For each experiment, we generate $n = 1000$ observations according to

$$y = f^\star(\boldsymbol{x}) + \sigma Z + B\xi,$$

where $Z \sim N(0, 1)$, $\sigma = 0.5$, $B \sim \text{Bernoulli}(\epsilon)$, and $\xi \sim t_2$, independent of $\boldsymbol{x}$. We vary the corruption probability ϵ between 0 and 0.2 and compare models fitted using L_2, L_1, Huber, Barron, and tangent-line truncated Barron losses. For the Barron losses, we set $\alpha = 1.5$ and $\beta = \sigma$, while the truncation threshold is fixed at $\tau = 3\sigma$. As before, the networks are trained using full-batch Adam with learning rate 5×10^{-3}, projecting each weight matrix onto the Frobenius ball of radius 2 after every update. All methods are trained on the same corrupted samples and evaluated on an independent test set of size 10^4 using squared prediction error. For each corruption probability, we replicate the experiment 100 times.

The results show a tradeoff between adapting to the regular observations and limiting extreme corrupted values. Since the cutoff $\tau = 3\sigma$, under low contamination, most residuals generated by the Gaussian component correctly fall inside the quadratic/untruncated region of the loss. Large residuals produced by the heavy-tailed component instead are treated only linearly due to the truncation. As the corruption probability increases, the expected proportion of corrupted observations with moderate residuals increases, and the cutoff does not directly distinguish regular observations from corrupted ones. Consequently, a threshold chosen according to the Gaussian noise scale need not provide the best robustness–efficiency balance at every contamination level, and L_1 training may become preferable. In these experiments, least-squares training deteriorates rapidly as the corruption probability increases.

This behaviour is consistent with the classical efficiency– robustness behavior of Huber loss under contamination (Huber & Ronchetti (2009)). Our experiments illustrate this perspective in neural-network regression and extend the empirical comparison to more general Lipschitz surrogates, such as the truncated Barron loss.

We include the untruncated Barron loss as an ablation to assess whether the observed improvement comes from the choice of the Barron loss itself or from its Lipschitz truncation. The untruncated Barron loss performs similarly to Huber, whereas its tangent-line truncation achieves lower average squared prediction error under corruption in the experiments considered. Since the two Barron objectives coincide below the cutoff and differ only in their treatment of larger residuals, this comparison highlights the empirical benefit of Lipschitz truncation. It illustrates how a Lipschitz surrogate can improve robustness while preserving the original loss below the cutoff.

A.9 Proof of Proposition 3

Proof. We want to show the following lower bound

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \geq \frac{m}{2^{1+1/p}\sqrt[p]{n^{p-1}t}}\right) > t,$$

which is equivalent to show that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n\mathbb{E}[X] + \frac{mn^{1/p}}{2(2t)^{1/p}}\right) > t. \tag{27}$$

The proof follows straight from the definition of X_i, which is taken to be $a = m/q^{1/p}$ times a Bernoulli random variable of parameter $q = 2t/n$. Thus, by definition of X_i, we have that

$$\mathbb{E}X = aq = \frac{m(2t)^{(p-1)/p}}{n^{(p-1)/p}} \implies n\mathbb{E}X = m(2t)^{1-1/p}n^{1/p}, \tag{28}$$

which implies that

$$\begin{aligned} n\mathbb{E}X + n\frac{m}{2^{1+1/p}\sqrt[p]{n^{p-1}t}} &= m(2t)^{1-1/p}n^{1/p} + \frac{mn^{1/p}}{2(2t)^{1/p}} \\ &= \frac{n^{1/p}m}{(2t)^{1/p}}\left(2t + \frac{1}{2}\right) \\ &= a\left(2t + \frac{1}{2}\right). \end{aligned} \tag{29}$$

Notice that for $t < 1/4$, $2t + 1/2 < 1$. Now, let A be the following event

$$A = \{X_i = a, \text{ for some } i = 1, 2, \ldots, n\}.$$

By equation 29, the inclusion of events below holds

$$A \subset \left\{\sum_{i=1}^{n} X_i \geq n\mathbb{E}[X] + \frac{mn^{1/p}}{2(2t)^{1/p}}\right\}.$$

Thus, the proof is concluded if we show that $\mathbb{P}(A) > t$. Indeed,

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \left(1 - \frac{2t}{n}\right)^n > 1 - e^{-2t} > t,$$

where in the last inequality we have used that $g(s) = 1 - e^{-2s} - s$ satisfies $g(0) = 0$ and $g'(0) = 1$, thus, it is positive at a neighborhood of zero. This concludes the proof. □

B Classical Results

We let $\{r_i\}_{i=1}^{\infty}$ be an i.i.d sequence of Rademacher random variables independent of the data and the fresh copy $(y, \boldsymbol{x})$.

Lemma 4 (Symmetrization. Lemma 2.3.1 of van der Vaart & Wellner (1996)). *Let $Z_1, \ldots, Z_n$ be i.i.d random elements of a measurable space $\mathcal{Z}$, and let $\mathcal{G}$ be a countable class of measurable functions $g : \mathcal{Z} \to \mathbb{R}$ with $\mathbb{E}|g(Z_1)| < \infty$ for every $g \in \mathcal{G}$. Then,*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}(g(Z_i) - \mathbb{E}g(Z_i))\right|\right] \leq 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} r_i g(Z_i)\right|\right]$$

Theorem 5 (Contraction principle; Theorem 4.12 in Ledoux & Talagrand (1991)). *Let $F : [0, \infty) \to [0, \infty)$ be a convex and increasing function, let $\psi_i : \mathbb{R} \to \mathbb{R}$ for $i \in \{1, 2, \ldots, n\}$ satisfy $\psi_i(0) = 0$ and $|\psi_i(t) - \psi_i(s)| \leq |s - t|$ for all $s, t \in \mathbb{R}$, and let T be a bounded set of $\mathbb{R}^n$. Then,*

$$\mathbb{E}F\left(\frac{1}{2}\sup_{(t_1,\ldots,t_n)\in T}\left|\sum_{i=1}^{n} r_i\psi_i(t_i)\right|\right) \leq \mathbb{E}F\left(\sup_{(t_1,\ldots,t_n)\in T}\left|\sum_{i=1}^{n} r_i t_i\right|\right).$$

Theorem 6 (Bounded differences inequality; McDiarmid (1989), Theorem 6.2 in Boucheron et al. (2013)). *Let $Z_1, \ldots, Z_n$ be independent random elements of a measurable space $\mathcal{Z}$, and let $\varphi : \mathcal{Z}^n \to \mathbb{R}$ be measurable. Suppose that there are constants $c_1, \ldots, c_n \geq 0$, not all zero, such that $|\varphi(z) - \varphi(z')| \leq c_i$ whenever $z, z' \in \mathcal{Z}^n$ differ in the i-th coordinate only. Then φ is bounded, and for every $\varepsilon > 0$,*

$$\mathbb{P}\Big(\varphi(Z_1, \ldots, Z_n) - \mathbb{E}\big[\varphi(Z_1, \ldots, Z_n)\big] \geq \varepsilon\Big) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Theorem 7 (von Bahr–Esseen inequality; Pinelis's form Pinelis (2015)). *Let $X_1, \ldots, X_n$ be centered random variables with finite p-th moment for some $p \in (1, 2]$, satisfying*

$$E[X_{i+1} \mid X_1 + \cdots + X_i] = 0, \text{a.s.}, \tag{30}$$

for all $i \in \{1, 2, \ldots, n-1\}$. Then,

$$E\left[\left|\sum_{i=1}^{n} X_i\right|^p\right] \leq 2^{2-p}\sum_{i=1}^{n} E|X_i|^p.$$