"DEEP LIFTS": NEURAL NETWORKS CAN GENERALIZE EVEN WITH THE DUMBEST OPTIMIZER

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a labeled dataset $\{(x_i, y_i)\}_{i=1}^n$, consider drawing a parameter vector $\hat{\theta}$ uniformly at random from those that interpolate the data within a width-W depth-D ReLU architecture with continuous weights bounded within $[-M, M]$. We study the generalization properties of $\hat{\theta}$ in the case where the dataset can, in fact, be explained by a smaller w-by-d ReLU network. Despite the obliviousness of the selection rule for $\hat{\theta}$, it adaptively generalizes at the rate $\tilde{O}\left(\frac{wdWD}{n}\right)$, rather than being governed by the VC dimension of the full W-by-D hypothesis class. The reason is due to a particular inductive bias in ReLU networks: the w-by-d parameter **witnesses** an exponential volume of W-by-D interpolating parameters, which dwarfs the volume of poorly generalizing parameters (conditioned on interpolation). We explicitly construct this set of near-equivalent W-by-D parameters by a proof technique called **deep lifting**, and then tie the achievable volume to stability properties of the w-by-d witness. Examples show that these techniques allow us to convert approximation-theoretic results into generalization guarantees for nonparametric classes. We further show that generalization of uniform-random interpolators can survive label corruptions.

1 INTRODUCTION

Modern deep learning generalization poses a problem for classical statistical learning theory (SLT). SLT typically formulates generalization bounds in terms of the complexity of the underlying hypothesis class, but the class of functions realized by over-parameterized neural networks is complex enough to fit even randomly assigned labels (Zhang et al., 2017). It is broadly believed that the explanation for this apparent contradiction is that modern learning pipelines implicitly constrain learned deep neural networks (DNNs) to be "simple" (Valle-Pérez et al., 2019; Berchenko, 2024; Wilson, 2025). Much recent work searches for this *simplicity bias* in the optimizers used to train DNNs. For example, gradient-based methods paired with specific loss functions have been shown to preferentially select low-norm/low-rank (Gunasekar et al., 2017; 2018; Arora et al., 2019), max-margin (Soudry et al., 2018; Lyu & Li, 2020; Chizat & Bach, 2020), or flat/stable (Mulayoff & Michaeli, 2020; Arora et al., 2022; Qiao et al., 2024; Liang et al., 2025) solutions, even when the surrounding hypothesis class contains far more complex alternatives.

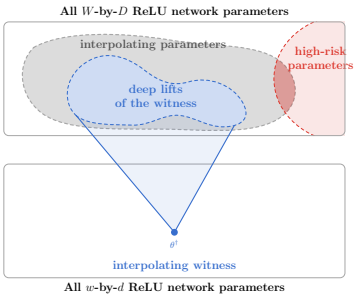

Figure 1: A small interpolating network "witnesses" many large interpolating networks. Blue volume dwarfs the intersection of the gray and red volumes once $n \gg wWD$.

At the same time, DNN generalization occurs robustly across many choices of loss function and optimizer, and it is unclear whether *optimizer-level* analyses are overindexing on properties of specific optimizer/loss-function pairs (Chiang et al., 2023). We ask the question: **do optimizers need to be secret geniuses to produce generalization?**

Consider as an alternate object of study the "learning algorithm" that chooses a width-W-depth-D DNN's parameters uniformly at random from those that interpolate the training

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.

labels. Although such a learner cannot be feasibly implemented, it represents the "dumbest possible" learning rule, because it chooses interpolating parameters without any regard for simplicity. **We study the generalization properties of this dumb learner, in order to isolate those aspects of generalization that are independent of optimizer bias.**

How can uniform random interpolating parameters be expected to generalize? Consider the scenario where there exist $w \ll W$ and $d \ll D$ such that some w-by-d network approximates the data generating mechanism. In the case where parameter weights are forced to belong to a discrete set, Buzaglo et al. (2024) show that uniform random interpolating parameters generalize faster than expected. The reason is architectural: any width-w depth-d DNN can be used to construct an exponential number of width-W depth-D networks that compute the same function, so that a w-by-d interpolating network **witnesses** a large number of W-by-D interpolators. A counting argument therefore establishes that a uniform random interpolating parameter for a W-by-D network generalizes at the rate $\tilde{O}(\frac{wWD}{n})$, where n is the number of data points, rather than the naive $\tilde{O}(\frac{W^2D}{n})$ one would expect from uniform convergence. The proof technique of "lifting" a small interpolating witness to furnish a vast set of larger interpolators has been used in several recent generalization studies (Buzaglo et al., 2024; Harel et al., 2024; Dziugaite & Roy, 2025a), and we call it **deep lifting**.

However, existing work in this direction is limited in at least two ways. The first limitation is that parameters are assumed to be drawn from a discrete set of values; this assumption is crucial in prior work, because the set of equivalent "lifted" parameters generally has zero probability under a continuous measure. The second limitation is that data is assumed to be noiseless and realizable — a necessary expedient for there to exist a small DNN that simultaneously interpolates and generalizes. Although results that address these limitations exist, they rely on further restrictions to the architecture or data model (e.g. shallow or boolean DNNs (Buzaglo et al., 2024; Harel et al., 2024)).

We extend the technique of deep lifting to (i) cover arbitrarily deep DNNs with continuous parameters drawn uniformly from a bounded interval, and (ii) handle label noise and unrealizability.

Suppose we have data $x_1 \dots x_n \in \mathbb{R}^p \overset{i.i.d.}{\sim} \mu_X$ and (multi-dimensional) labels $y_1 \dots y_n \in \mathbb{R}^s$. Let $\mathcal{C}_M(w,d)$ denote the class of parameter vectors for width-w, depth-d DNN with weights inside $[-M,M]$, and write $g_\theta^{w,d}$ for the function realized by $\theta \in \mathcal{C}_M(w,d)$. Given training data $\{(x_i,y_i)\}_{i=1}^n$, let

$$S_{\tau,n}^{W,D,M} := \left\{\phi \in \mathcal{C}_M(W,D) : \max_{i\in[n]} \|g_\phi^{W,D}(x_i) - y_i\|_\infty \le \tau\right\} \qquad U_{\tau,n}^{W,D,M} := \mathrm{Unif}\left(S_{\tau,n}^{W,D,M}\right). \tag{1}$$

define the uniform distribution over parameters that interpolate up to precision τ. See Appendix B for a formal definition of $U_{\tau,n}^{W,D,M}$. For a target f, define the "interpolation risk at precision τ" as

$$\Delta_\tau(\theta; f) := \Pr_{x\sim\mu_X}\left[\|g_\theta(x) - f(x)\|_\infty > \tau\right]. \tag{2}$$

We study $\Delta_\tau(\hat{\theta}; f)$ under $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$, under various data-generating mechanisms. Some of our main findings are the following.

Noiseless/realizable extension from continuous deep lifting. Assume $y_i = f(x_i)$. Furthermore, assume that f is realizable (up to precision $\tau/2$) by some witness $\theta \in \mathcal{C}_M(w,d)$, i.e. $\sup_{x\in[-X,X]^p} \|f(x) - g_\theta^{w,d}(x)\|_\infty \le \tau/2$. Then Theorem 3 shows

$$\Pr_{x^n,\hat{\theta}}\left[\Delta_\tau(\hat{\theta}; f) \ge \tilde{O}\left(\tfrac{wdWD}{n}\right)\right] \le \delta,$$

where $\tilde{O}(\cdot)$ hides a linear contribution from p and s, as well as log terms.

Noisy/unrealizable settings Assume that, of the n training points, m labels are (possibly adversarially) corrupted by a bounded perturbation, while all other labels are from the ground truth f. Then, under the same witness assumption as above, $\hat{\theta}$ still generalizes with respect to f at the rate:

$$\Pr_{x^n,\hat{\theta}}\left[\Delta_\tau(\hat{\theta}; f) \ge \tilde{O}\left(\tfrac{W(wD+m)d}{n}\right)\right] \le \delta$$

even though we force it to interpolate adversarial noise (Theorem 6).

By this token, when $m \asymp \epsilon n$ grows as a constant fraction of n, we have a tempered overfitting type rate. Corollary 7 shows that n can be chosen so that, with high probability

$$\Delta_\tau(\hat{\theta}; f) = \tilde{O}(W\epsilon).$$

when depth D_n grows with n. Similarly, we are able to certify generalization when f is not uniformly approximable by any DNN for any τ (Theorem 8).

Classification We derive analogous rates for classification. Here, the s network outputs are interpreted as logits for s classes, and a network parameter is said to interpolate the data if its argmax logit is always the correct class. As a corollary of these results, we are able to extend generalization results to $\hat{\theta}$ sampled from the Boltzmann distribution built on cross entropy loss and conditioned on interpolation. Due to space constraints, we report these results in Appendix H.

Key techniques The paper's proofs are based on a central technique that we call **deep lifts**. Heuristically: a "witness" $\theta \in \mathcal{C}_M(w, d)$ can be "lifted" into an equivalent parameter in $\mathcal{C}_M(W, D)$ in many ways. Thus, the existence of a single "good" width-w depth-d implies a large amount of probability mass on "good" width-W depth-D parameters (conditioned on interpolation) (Buzaglo et al., 2024; Harel et al., 2024; Dziugaite & Roy, 2025a).

The basic difficulty introduced by the continuous setting is producing a set of "deep lifts" with positive measure. While in the discrete case, some subset of the witness network's parameters could be frozen while maintaining exponential lifted mass, the continuous case requires that each weight vary in a neighborhood of the corresponding weight in the witness. For the witness behavior to survive such perturbations in the weights, we identify a sufficient condition that we call "structured stability," which we tie to the witness' ability to reject small perturbations injected at each layer.

Another crucial difficulty of the analysis is dealing with label noise in the interpolation setting. By requiring interpolation, we require that the network $g_{\hat{\theta}}$ explicitly disagree with the ground truth on the training set; but we proceed to evaluate $g_{\hat{\theta}}$ based on its population agreement with f. Here, our proof technique is to design a memorization gadget that learns localized spikes in order to fit corruptions. This gadget can be "bolted on" to a witness $\theta \in \mathcal{C}_M(w, d)$ that realizes f, generating an interpolating parameter $\theta^\dagger \in \mathcal{C}_M(w', D)$, upon which we may apply our analysis, whenever $W > w'$. Importantly, the memorization gadget appears only in the proof (to furnish a small interpolating witness), and it is not assumed or expected that $\hat{\theta}$ explicitly implements the gadget.

1.1 Related work

Valle-Pérez et al. (2019); De Palma et al. (2019) provide early evidence showing that uniform distributions over parameters push forward onto non-uniform distributions over implemented functions; Mingard et al. (2021; 2025) correspondingly analyzes SGD from the perspective of Bayesian sampling. Berchenko (2024) shows that random constructions can exhibit simplicity bias in overparameterized models. Unlike most of these works, we specifically consider uniformity *conditioned on interpolation*. Most relevant to our work are Buzaglo et al. (2024), who show that randomly sampled DNNs also generalize in the case of noiseless data from a realizable ground truth; Harel et al. (2024), who show that this analysis can be extended to noisy binary classification, but considers only DNNs with binary weights; and Dziugaite & Roy (2025a), who reformulate the proof technique of Buzaglo et al. (2024), and consider DNNs drawn from loss-based Boltzmann distributions rather than only interpolating DNNs. To our knowledge, prior work does not cover deep networks with continuous weights, or label noise in a general setting. Extended discussion of related work is in Appendix A.

2 Preliminaries

We are interested in approximating functions mapping $\mathbb{R}^p \to \mathbb{R}^s$ with ReLU Deep Neural Networks. Given p and s of interest[1], we define a width-w depth-d neural network to be given by a set of parameters, θ. We consider θ as organizing into a parameter vector the weight matrices

$$W_1^\theta \in \mathbb{R}^{w \times p}, W_k^\theta \in \mathbb{R}^{w \times w} \text{ for } k = 2 \dots d,$$

as well as biases $b_k^\theta \in \mathbb{R}^w$ for $k = 1 \dots d$ and a readout matrix $R^\theta \in \mathbb{R}^{s \times w}$. As such, θ is a vector with $N_{w,d} = w^2(d-1) + wd + wp + ws$ entries.

The neural network defined by θ processes inputs $x \in \mathbb{R}^p$ according to $x_k^\theta = \sigma(W_k^\theta x_{k-1}^\theta + b_k^\theta)$ with $x_0^\theta := x$ and $\sigma(z) = \max\{z, 0\}$ defined to be an elementwise application of ReLU. We call x_k^θ the layer-k activations. The function output defined by the neural network is $g_\theta^{w,d}(x) = R^\theta x_d^\theta(x) \in \mathbb{R}^s$.

[1] treated as constants in this paper

For a real number, $M > 0$, we define $\mathcal{C}_M(w, d)$ to be the set of all parameter vectors for width-w depth-d neural networks with real-valued elements in $[-M, M]$. Thus, $\mathcal{C}_M(w, d)$ may be identified with $[-M, M]^{N_{w,d}}$. For a sequence $M_n \to \infty$, we define $\mathcal{C}(w, d) = \bigcup_n \mathcal{C}_{M_n}(w, d)$, which may likewise be identified with $\mathbb{R}^{N_{w,d}}$.

3 Regression setting

In the (multi-valued) regression setting, we receive data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}^s$. We assume that the covariates x_i are drawn i.i.d. from a measure μ_X on p dimensions supported within $[-X, X]^p$ for some bound $X > 0$. For now, labels are left as general vectors.

We generally assume that there exists some ground truth $f : \mathbb{R}^p \to \mathbb{R}^s$ against which we measure the performance of a parameter $\theta \in \mathcal{C}(w, d)$. In the paper body, this performance is measured by **interpolation risk** at precision $\tau > 0$, or simply **risk** at τ, notated as $\Delta_\tau(\theta; f)$ and defined in Equation (2). Because f will always be clear from context in this paper, we often write only $\Delta_\tau(\theta)$. Given data $\{(x_i, y_i)\}_{i=1}^n$, target width W and depth D, and $\tau > 0$, we recall the definition of $S_{\tau,n}^{W,D,M}$ as the set of "τ-interpolating parameters," according to (1), and $U_{\tau,n}^{W,D,M}$ as the Lebesgue-induced uniform distribution over $S_{\tau,n}^{W,D,M}$ (technical definition in Appendix B). In this paper, our goal is to study the generalization properties of $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$, under various assumptions on the mechanism generating the labels $\{y_i\}$.

Remark 1 (Interpretation of τ). τ should be interpreted as a numerical precision. In practice, it is common to say a model "interpolates" if its predictions are within $\tau = 10^{-6}$ of the true values.

Remark 2 (Why study $\Delta_\tau(\hat{\theta}; f)$?). $\Delta_\tau(\hat{\theta}; f)$ has a clear interpretation as the population probability of making an incorrect continuous prediction, at the tolerance τ. Moreover, in the case where $s = 1$, $|f(x)| \le F$, and we define $\hat{f}$ by clamping $g_{\hat{\theta}}^{W,D}$ between $[-F, F]$, then a bound on $\Delta_\tau(\hat{\theta}; f)$ can be converted into an L_α-norm error bound for an exponent $\alpha \ge 1$:

$$\|\hat{f} - f\|_{L_\alpha(\mu_X)}^\alpha := \mathbb{E}_{x \sim \mu_X}\left[\left|\hat{f}(x) - f(x)\right|^\alpha\right] \le \tau^\alpha + (2F)^\alpha \Delta_\tau(\hat{\theta}; f).$$

In this case, depending on the behavior of $\Delta_\tau(\hat{\theta}; f)$ as a function of n and τ, we may choose τ *as a function of* n *and* α to derive L_α rates.

3.1 Regression with noiseless and realizable data

We now state our first result, which is an extension of Buzaglo et al. (2024) to continuous weights. Theorem 3 says that, if we receive noiseless observations from a function f that is uniformly approximated by a small neural network $\theta \in \mathcal{C}_M(w, d)$, then we generalize at the rate $\tilde{O}(\frac{wdWD}{n})$.

Theorem 3 (Continuous noiseless realizable setting)**.** *Let $\tau > 0$ and $\delta \in (0, 1)$. In the dataset $\{(x_i, y_i)\}_{i=1}^n$ assume that $y_i = f(x_i)$ for all $i = 1 \dots n$.*

Let $M \ge 1$. Assume that, for some w and d, there exists $\theta \in \mathcal{C}_M(w, d)$ so that $\|f(x) - g_\theta^{w,d}(x)\|_\infty \le \frac{\tau}{2}$ for all $x \in \mathrm{supp}(\mu_X)$. Let $D \ge d$ and $W > w$. Then for $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$

$$\Pr_{x^n, \hat{\theta}}\left[\Delta_\tau(\hat{\theta}) \ge t_n\right] \le \delta, \qquad t_n := \frac{C_0 wWD \log q_{\text{eff}} + \log \frac{1}{\delta}}{n}.$$

Here, C_0 is a constant depending linearly on p and s, and q_{eff} is a value satisfying

$$\log q_{\text{eff}} \le (d + 2) \log\left(16 M w D (pX + 2) \max\{1, \tau^{-1}\}\right).$$

We emphasize a key feature of Theorem 3: since the bound holds for every witness that (i) fits inside the ambient W-by-D architecture and (ii) explains the data-generating mechanism, the result may be optimized over all such witnesses; in this sense, the theorem is an adaptivity result, showing that the generalization rate adapts to the size of the *smallest* explanation available within the architecture. Though the size of the smallest network representing the ground truth is unknown, the cost of choosing an overlarge architecture is relatively mild, which motivates overparameterization.

In Theorem 3, the quantity q_{eff} plays a role analogous to the alphabet size in Buzaglo et al. (2024) (i.e. the cardinality of the discrete set from which weights are drawn). Unlike in the discrete setting

of Buzaglo et al. (2024), however, our generic bound incurs a linear dependence on the witness depth d through $\log q_{\text{eff}}$.[2] The source of this dependence is that, in the continuous setting, a witness must continue to implement nearly the same function when its parameters are perturbed within a small neighborhood. For generic networks, control of this "stability" can deteriorate rapidly (exponentially) with d. However, specific deep witnesses may in fact have stronger stability properties. Theorem 36, explicitly gives the rate in terms of the "stability" of the witness (as defined in Section 4), and the second example below shows that doing so can lead to sharper statistical rates when a deep but "stable" witness can be procured.

Example: Hölder functions Suppose $\text{supp}(\mu_X) \subseteq [0,1]^p$ and let $f : [0,1]^p \to \mathbb{R}$ satisfy Hölder smoothness with index $\beta > 0$ and radius $K > 0$, meaning it has bounded partial derivatives up to order $\lfloor \beta \rfloor$, and its $\lfloor \beta \rfloor$-order partial derivative are $\beta - \lfloor \beta \rfloor$ Hölder continuous (Schmidt-Hieber, 2020). Using the approximation theoretic results from Ohn & Kim (2019), we find that $\Delta_\tau(\hat{\theta}) = \tilde{O}\left(\frac{WD\tau^{-p/\beta}}{n}\right)$ with high probability, as long as W and D are large enough. If f is bounded by F, then following Remark 2 and tuning τ leads to the high probability L_1 bound

$$\left\|\hat{f} - f\right\|_{L^1(\mu_X)} = \tilde{O}\left(n^{-\frac{\beta}{\beta+2p}}\right).$$

For reference, without additional regularity assumptions on μ_X, even infinitely smooth functions have a lower bound of $\Omega\left(\frac{1}{n}\right)$ (Kohler & Krzyżak, 2013); the exponent in our rate approaches this barrier as β grows with p fixed. Further details for examples can be found in Appendix I.

Example: Barron functions Now assume $\text{supp}(\mu_X) \subseteq B_p$, where B_p is the Euclidean unit ball in p dimensions. ReLU-Barron functions $f : B_p \to \mathbb{R}$ are functions realizable as "infinite-width" ReLU neural networks:

$$f(x) = \mathbb{E}_{a,b,c\sim\rho}\left[a\sigma(b^T x + c)\right]. \tag{3}$$

The ReLU-Barron norm $\|f\|_{\mathcal{B}}$ is defined as $\inf_\rho \mathbb{E}\left[|a|(\|b\|_1 + |c|)\right]$, where the infimum is taken over measures ρ that realize (3). Let f be ground truth and $Q := \|f\|_{\mathcal{B}}$. In Appendix D.5, we find that a witness of depth $d_\tau \asymp (Q/\tau)^{\frac{2p}{p+3}}$ and width-$(2p+3)$, approximates f uniformly on B_p to precision $\tau/2$. Moreover, this witness's "structured stability" (to be defined in Section 4) deteriorates only linearly in d_τ. Taking $D \asymp d_\tau$, Theorem 36 (a strengthening of Theorem 3) gives $\Delta_\tau(\hat{\theta}) \leq \tilde{O}((W/n)(Q/\tau)^{\frac{2p}{p+3}})$ w.h.p whenever $W > 2p+3$. f is bounded by Q, and tuning τ with Remark 2 gives the rate

$$\left\|\hat{f} - f\right\|^2_{L_2(\mu_X)} = \tilde{O}\left(Q^2\left(\frac{W}{n}\right)^{\frac{p+3}{2p+3}}\right).$$

For fixed p, Q and $W > 2p+3$, this rate matches, up to logarithmic factors, the squared-L_2 minimax lower bound for Gaussian-noise regression (Yang & Zhou, 2025, Remark 4.3) and the best known upper-bound exponent for noiseless i.i.d. sampling (Parhi & Adcock, 2025, Equation (39), $k = 1$).

3.2 Regression with corrupted labels

We now extend the results of the previous section to cover data with noise. Specifically, we cover the case where an adversary is allowed to apply a bounded perturbation to any subset of the labels before $\hat{\theta}$ is drawn.

Assumption 4 ((m, Γ) adversarial corruptions)**.** Suppose the ground truth function is f. The dataset (x_i, y_i) is (m, Γ) adversarially corrupted if, after seeing $x_1 \dots x_n \sim \mu_X$, an adversary chooses any subset $I \subseteq [n]$ with size $|I| \leq m$, and $\{\delta_i\}_{i\in I} \subset [-\Gamma, \Gamma]^s$, and reveals to the learner the labels

$$y_i = f(x_i) + \delta_i \ (i \in I), \qquad y_i = f(x_i) \ (i \notin I).$$

Our result requires the following mild regularity condition on the distribution μ_X.

Assumption 5 (Λ marginal boundedness of μ)**.** For $\Lambda > 0$, we say that μ_X is Λ-marginally bounded if the marginal of the first coordinate of $x \sim \mu_X$ satisfies $\Pr[x[1] \in I] \leq \Lambda|I|$ for every interval $I \subseteq [-X, X]$.[3]

[2] This dependence may be considered mild, as w and d are viewed as small constants relative to W, D (Buzaglo et al., 2024).

[3] We take the first coordinate only for simplicity. Indeed, we can derive similar results under the assumption that there exists some fixed vector $a \in \mathbb{R}^p$ so that $x^T a$ has bounded density.

We now have a similar result to Theorem 3, which says that, even in the face of adversarial corruptions, if the ground truth f is approximated to precision $\tau/2$ by some $\theta \in \mathcal{C}_M(w,d)$, we may bound the generalization error of $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$ with high probability. A formal version of Theorem 6 is stated as Theorem 37 in the Appendix.

Theorem 6 (Regression with corrupted labels, informal)**.** *Let $\tau > 0$, $\delta \in (0,1)$, and $n \geq 2$. Suppose that our data-generating mechanism satisfies Assumptions 4 and 5, and assume that f is realizable up to precision $\tau/2$ by some witness $\theta \in \mathcal{C}_M(w,d)$.*

Assume moreover that $M \geq \mathrm{poly}(n, \Lambda, \Gamma, X, \delta^{-1})$, where poly is a fixed polynomial stated in the formal version. Let $D \geq d$ and $W \gtrsim w + s + \frac{m}{D}$. Then, for $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$

$$\Pr_{x^n,\hat{\theta}}\left[\Delta_\tau(\hat{\theta}; f) \geq C_0 \frac{W(wD+m)\log q_{\mathrm{eff},n}}{n}\right] \leq \delta,$$

where $\log q_{\mathrm{eff},n} \leq 2\log\left(\mathrm{poly}(n, m, W, w, d, \Gamma, \Lambda, X, p, s, \delta) DM \max\{1, \tau^{-1}\}\right)$ and C_0 is a constant depending only on p and s.

Let us interpret Theorem 6. The crucial witness assumption has remained unchanged: f is approximately realized by a small $\theta \in \mathcal{C}_M(w,d)$. The key conclusion is also the same; we get a generalization bound for $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$. However, since it is no longer assumed that all n labels come from f, we require additional assumptions on M and W. Since M enters the bound only under a log, choosing it to satisfy the required polynomial dependence is mild. There is additionally a new requirement on W. While in the noiseless setting, we needed only $W > w$, we now need $W > w + w_+$, for some slack w_+ that increases according to the ratio m/D. As we will see, this is because $\theta \in \mathcal{C}_M(w,d)$ does not itself interpolate the noisy points, and we use the extra width to memorize m/D points per layer. The upshot is that the noisy rate of Theorem 6 incurs an extra $\frac{Wmq_{\mathrm{eff},n}}{n}$ compared to the noiseless rate of Theorem 3.

If $m = o_n(n)$, then $\Delta_\tau(\hat{\theta}) = o_n(1)$ in Theorem 6. However, a key case of interest is when m scales as $m \asymp \epsilon n$. The following corollary finds that the assumptions of Theorem 6 are satisfied for $W > w + s + O(1)$ and $D = \Theta(n)$. Plugging in $m \asymp \epsilon n$ together with these values gives a tempered-overfitting type risk floor at $\tilde{O}(W\epsilon)$ when $W \gg w$. The formal statement is Theorem 40.

Corollary 7 (Tempered-type overfitting under constant label noise, informal)**.** *In the setting of the previous theorem, if $m \sim \mathrm{Binom}(n, \epsilon)$, $W \gtrsim w + s + B$ for some $B \in \mathbb{N}$, then there is some depth scaling $D = \Theta(n)$ so that the following holds.*

At $n = \max\left\{2, \left\lceil \frac{\log(2/\delta)}{\epsilon} \vee \frac{Bd - \log(2/\delta)}{\epsilon} \right\rceil\right\}$, we have

$$\Pr_{x^n \sim \mu_X^{\otimes}, \hat{\theta} \sim U_{\tau,n}^{W,D,M}}\left[\Delta_\tau(\hat{\theta}; f) > \tilde{O}\left(Wd\left(1 + \tfrac{w}{B}\right)\epsilon\right)\right] \leq \delta.$$

Empirical work finds that interpolating solutions typically exhibit "tempered overfitting," rather than converging to the optimal risk (Mallinar et al., 2022) in noisy settings, so it is not surprising that we observe a nonzero risk floor. We also point out a weakness. Corollary 7 is not an asymptotic result, but rather holds at a fixed value of n. Because of the $\log n$ dependence in Theorem 6, there are a wide range of n that give the $\tilde{O}(W\epsilon)$ rate, but eventually the guaranteed risk diverges as $\Theta(\log n)$.

3.3 Regression with unrealizable ground truth

We now come to results that cover the case where the ground truth, f, cannot be realized (i.e. uniformly approximated) by any ReLU neural network. Instead of assuming realizability, we assume that there is a sequence of witnesses $\theta_n \in \mathcal{C}_{M_n}(w_n, d_m)$ that approximate f to precision τ on some set of measure at least $1 - \epsilon_n$. Then, if our W_n-by-D_n architecture scales so that these witnesses remain admissible, we may generalize at a rate defined by ϵ_n.

Theorem 8 (Unrealizable ground truths, informal)**.** *Let $\delta \in (0,1)$ and $\tau > 0$. Assume the noiseless setting, i.e. $y_i = f(x_i)$ for all $i \in [n]$. Assume μ_X satisfies Assumption 5.*

Consider sequences $\epsilon_n \in (0,1]$, w_n, d_n, D_n, A_n with the requirements that $\epsilon_n = \Omega(\frac{1}{n})$ and $D_n \geq d_n$. Suppose that, for all $n \in \mathbb{N}_{\geq 2}$, there is $\theta_n \in \mathcal{C}_{A_n}(w_n, d_n)$ such that

$$\mu_X\left(\left\{x : \|g_{\theta_n}^{w_n,d_n}(x) - f(x)\|_\infty > \tfrac{\tau}{2}\right\}\right) \leq \epsilon_n, \qquad \sup_{x \in [-X,X]^p} \|g_{\theta_n}^{w_n,d_n}(x) - f(x)\|_\infty \leq \Gamma.$$

Assume moreover that $M_n \geq \mathrm{poly}(n, \Lambda, \Gamma, X, \delta^{-1}) \vee A_n$, *where* poly *is a fixed polynomial stated in the formal version, and* $W_n > w_n + w_+$, *where* $w_+ = \tilde{O}\left(1 + \frac{\epsilon_n n}{D_n}\right)$. *Consider* $\hat{\theta} \sim U_{\tau,n}^{W_n, D_n, M_n}$. *Then*

$$\Pr_{x^n, \hat{\theta}}\left[\Delta_\tau(\hat{\theta}) \geq \tilde{O}\left(W_n\left(\frac{w_n D_n + \log\frac{1}{\delta}}{n} + \epsilon_n\right)\log q_{\text{eff,n}}\right)\right] \leq \delta$$

where $\log q_{\text{eff}} = d_n \log \mathrm{poly}(n, W_n, D_n, w_n, d_n, \Lambda, \Gamma, M_n, \tau^{-1}, \delta^{-1})$.

Technically, Theorem 8 bears a resemblance to Corollary 7, due to the additive $\tilde{O}(W_n \epsilon_n)$ term. However, in the noisy case, ϵ is a constant rate parameter, whereas in the unrealizable case, we may have $\epsilon_n = o_n(1)$, allowing us to achieve $\Delta_\tau(\hat{\theta}; f) = o_n(1)$.

Example: Learning half spaces Consider the space of functions indicating p-dimensional half spaces:

$$\mathrm{Halfspaces}(R) = \left\{f_{a,b}(x) = \mathbf{1}\{a^T x - b > 0\} : a \in \mathbb{S}^{p-1},\ |b| \leq R\right\}.$$

Notice that for any $f \in \mathrm{Halfspaces}(R)$, the τ-realizability condition of Theorem 3 is unsatisfiable for any $\tau < 1$. However, under a mild regularity condition on μ_X, it is straightforward to define a sequence of $\theta_n \in \mathcal{C}_{A_n}(2, 1)$ with $A_n = \Theta(n)$ so that the hypotheses of Theorem 8 hold with $\epsilon_n = \frac{1}{n}$ and $\Gamma = 1/2$. These are the θ_n implementing ramp-type approximations to f.

Choose M_n polynomial in n and the problem parameters, large enough to satisfy the theorem statement, and take $D_n \geq 1$. For $\delta \in (0, 1)$, set

$$W_n > 6 + 4\left\lceil \frac{2 + \log\frac{4}{\delta}}{D_n} \right\rceil.$$

Then, for $\hat{\theta} \sim U_{\tau,n}^{W_n, D_n, M_n}$, we have $\Pr_{x^n, \hat{\theta}}[\Delta_\tau(\hat{\theta}) \geq \tilde{O}(\frac{W_n D_n}{n})] \leq \delta$ when $\tau \asymp 1/n$. According to Remark 2, the corresponding L_1 rate is

$$\|\hat{f} - f\|_{L_1(\mu_X)} \leq \tau + \Delta_\tau(\hat{\theta}).$$

Consequently, for $\tau \asymp 1/n$, with probability at least $1 - \delta$,

$$\|\hat{f} - f\|_{L_1(\mu_X)} = \tilde{O}\left(\frac{W_n D_n}{n}\right).$$

4 Proof Sketch

We now sketch the proofs for Theorems 3 and 6. Formal proofs are deferred to the Appendix

Structured Stability of Parameters We begin by introducing the notion of "structured stability," which quantifies how similarly a parameter $\theta \in \mathcal{C}_M(w, D)$ behaves to its nearby neighbors.

Definition 9 (Structured stability)**.** A parameter $\theta \in \mathcal{C}_M(u, D)$ has (H, L, C), with $H, L, C > 0$, *structured stability* if (i) $\max_{k\in[D]} \sup_{x\in\mathcal{X}} \|x_k^\theta(x)\|_\infty \leq H$ and (ii) for every $\tilde{\theta} \in \mathcal{C}_M(u, D)$ and $\rho \leq C^{-1}$ satisfying $\|\tilde{\theta} - \theta\|_\infty \leq \rho$, we have

$$\sup_{\substack{x\in\mathcal{X}\\ k\in[D]}} \|x_k^{\tilde{\theta}}(x)\|_\infty \leq H + 1 \quad \text{and} \quad \sup_{x\in\mathcal{X}} \|g_{\tilde{\theta}}^{u,D}(x) - g_\theta^{u,D}(x)\|_\infty \leq L\rho.$$

In other words, θ has bounded activations and all of its C^{-1}-close neighbors (i) enjoy a similar activation bound (ii) give similar outputs for all input in the support of μ_X. In the next section, we show that (H, L, C) govern the volume of a set of parameters $\phi \in \mathcal{C}_M(W, D)$ constructed to "behave like" θ, and therefore generalization in the larger architecture.

Deep Lifts for Stable Parameters Given a parameter $\theta \in \mathcal{C}_M(u, D)$, our proof technique relies on producing a large volume of parameters $\phi \in \mathcal{C}_M(W, D)$ that compute similar functions. We now construct such a set and write its volume in terms of the structured stability of θ.

Let $\theta \in \mathcal{C}_M(u, D)$ have (H, L, C) structured stability. Write the parameters of ϕ in block form

$$W_1^\phi = \begin{pmatrix} A_1 \\ C_1 \end{pmatrix}, \qquad W_{k\geq 2}^\phi = \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix}, \qquad b_k^\phi = \begin{pmatrix} a_k \\ c_k \end{pmatrix}, \qquad R^\phi = (Q_1\ Q_2),$$

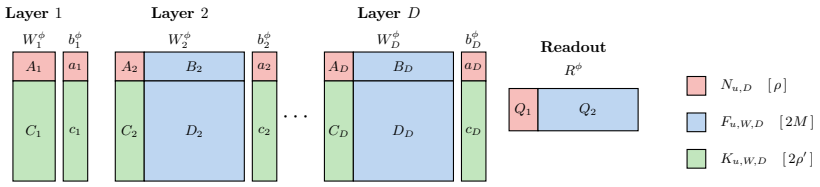

Figure 2: A visualization of the structure of a deep lift, $\phi \in \mathcal{C}_M(W, D)$, of $\theta \in \mathcal{C}_M(u, D)$. Most of ϕ's weights (in blue), may be chosen arbitrarily from $[-M, M]$. Those in green are chosen from an interval of size $2\rho'$. Those in red are chosen from an interval of size ρ (built around θ's weights). The result is a high-volume set of nearly equivalent lifts of θ.

where the block sizes follow from dividing the activation vector into a "carried" state of dimension u, and "added" state of dimension $W - u$. First, we explain that it is possible to arrange for the carried activations to emulate θ, and for the added activations to remain identically zero:

$$x_k^\phi[i] = \begin{cases} x_k^\theta[i], & i = 1, \dots, u, \\ 0, & i = u+1, \dots, W. \end{cases} \tag{4}$$

The proof is by induction. By structured stability, we have $\max_k \|x_k^\theta(x)\|_\infty \le H$. If (4) holds at k, then $\max_{j \ge u+1} \left| W_k^\phi x_{k-1}^\phi[j] \right| \le Hu\rho'$. Therefore, If ρ' is chosen small enough, we can choose the biases $c_k \in [-M, -M + 2\rho']$ so that the added pre-activations are negative, and the ReLU nonlinearity zeros them out. At the same time, under (4), the choice $A_k = W_k^\theta$ and $a_k = b_k^\theta$ suffices to establish $x_{k+1}^\phi[i \le u] = x_{k+1}^\theta$. Crucially, B_k and D_k have been left to choose their elements arbitrarily from $[-M, M]$. At readout, we may fix Q_2 arbitrarily.

This already gives many parameters ϕ implementing the same function as θ, but the construction has zero Lebesgue measure as a subset of $\mathbb{R}^{N_{W,D}}$, because the carried parameters A_k, a_k, Q_1 are fixed. The solution is to allow each carried weight to vary over an interval of length $\rho > 0$ around its corresponding value in θ. We call the constructed set of parameters $\phi \in \mathcal{C}_M(W, D)$ the set of "deep lifts of θ from u to W," notated as $\mathscr{L}_{u \to W}^{\rho,\rho'}(\theta)$. For suitably small radii ρ and ρ', Lemma 19 ensures that every $\phi \in \mathscr{L}_{u \to W}^{\rho,\rho'}(\theta)$ satisfies

$$\sup_{x \in [-X, X]^p} \|g_\theta^{u,D}(x) - g_\phi^{W,D}(x)\|_\infty \le \tfrac{\tau}{2}.$$

By construction, the Lebesgue measure of this set decomposes according the product of the intervals from which each individual weight is chosen, i.e.

$$\lambda\left(\mathscr{L}_{u \to W}^{\rho,\rho'}(\theta)\right) = \underbrace{(2M)^{F_{u,W,D}}}_{\text{from } B_k,\, D_k} \underbrace{\rho^{N_{u,D}}}_{\text{from } A_k,\, a_k} \underbrace{(2\rho')^{K_{u,W,D}}}_{\text{from } C_k,\, c_k}, \tag{5}$$

where $F_{u,W,D}$ counts the freely chosen parameters, $N_{u,D}$ the carried parameters chosen from intervals of length ρ, and $K_{u,W,D}$ the parameters chosen from intervals of length $2\rho'$. Notice ρ and ρ' both shrink as the structured stability parameters increase, so smaller stability parameters correspond to more lift volume. We summarize with the key upshot.

Lemma 10 (Informal). *Let $\theta \in \mathcal{C}_M(u, D)$ have (H, L, C)-structured stability and suppose it $\tau/2$-interpolates realized data $\{(x_i, y_i)\}_{i=1}^n$. Then, for $\rho = 1/C \wedge \tau/(2L)$, the set $\mathscr{L}_{u \to W}^{\rho,\rho'}(\theta) \subset \mathcal{C}_M(W, D)$ has measure given in Equation (5) and each $\phi \in \mathscr{L}_{u \to W}^{\rho,\rho'}(\theta)$ τ-interpolates data, with $\rho' \approx 1/H$. See Lemma 19 in the Appendix for the formal statement and precise radii.*

Generalization from Stable Interpolators We now state the main technical result that powers all of our results and sketch its proof.

Lemma 11 (Generalization from Structured Stability, Corollary 25 in Appendix). *Assume $y_i = f(x_i)$ for all labels except at most m (possibly chosen adversarially). Fix $u < W$ and D, M, H, L, C independently of the realized sample, with $M \ge 1$, $\tau > 0$, $m \le uD$, and $0 \le \eta < \delta < 1$. Suppose there is a training-data event $\mathcal{G}$ with probability at least $1 - \eta$ such that, on $\mathcal{G}$, there exists $\theta^\dagger \in \mathcal{C}_M(u, D)$ with (H, L, C) structured stability and*

$$\max_{i \in [n]} \|g_{\theta^\dagger}^{u,D}(x_i) - y_i\|_\infty \le \tfrac{\tau}{2}.$$

Then, for any $W > u$, we have that, for $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$:

$$\Pr_{\hat{\theta},x^n}[\Delta_\tau(\hat{\theta}) \geq t_n] \leq \delta, \qquad t_n := C_0 \frac{uWD \log \bar{q}_{\text{eff}} + \log \frac{1}{\delta - \eta}}{n},$$

where C_0 depends only on p and s, q_{eff} is defined in Theorem 23, and $\bar{q}_{\text{eff}} := \max\{en, q_{\text{eff}}\}$.

Lemma 11 tells us that if, with high probability over the draw of the dataset, there exists a (possibly data-dependent) parameter $\theta^\dagger \in \mathcal{C}_M(u, D)$ that (i) has structured stability (ii) interpolates the data, then $\hat{\theta}$ generalizes at the rate $\tilde{O}(uWD/n)$. We now outline the proof of Lemma.

Numerator/denominator decomposition Let $\mathcal{G}$ be the good training-data event in Lemma 11. On $\mathcal{G}$, the lift construction gives positive interpolation volume. Thus, for $\hat{\theta} \sim U_{\tau,n}^{W,D,M}$,

$$\Pr_{x^n,\hat{\theta}}[\Delta_\tau(\hat{\theta}; f) \geq t] \leq \eta + \Pr(\mathcal{G})\, \mathbb{E}\left[\frac{\lambda(B_t)}{\lambda(S_{\tau,n}^{W,D,M})} \,\middle|\, \mathcal{G} \right], \tag{6}$$

where λ is the Lebesgue measure defined on $\mathbb{R}^{N_{W,D}}$, and B_t is the set of interpolating parameters with risk larger than t:

$$B_t := S_{\tau,n}^{W,D,M} \cap \{\theta \in \mathcal{C}_M(W, D) \ : \ \Delta_\tau(\theta) \geq t\}.$$

We now consider the numerator terms and denominator terms separately. For the denominator, we focus on the event where there is a $\theta^\dagger \in \mathcal{C}_M(u, D)$ that $\tau/2$ interpolates the data. Under Lemma 10, we have $\mathscr{L}_{u\to W}^{\rho,\rho'}(\theta^\dagger) \subseteq S_{\tau,n}^{W,D,M}$, so that $\lambda\left(S_{\tau,n}^{W,D,M}\right) \geq \lambda\left(\mathscr{L}_{u\to W}^{\rho,\rho'}(\theta^\dagger)\right)$, and we know the exact volume on the RHS. This bound holds deterministically on $\mathcal{G}$, so we may pull it out of the expectation. Then we're left with the expectation of the numerator, and we interchange integration to express it as $\int_{\theta\,:\Delta_\tau(\theta)\geq t} \Pr_{x^n\sim\mu_X}\{\theta \in B_t\} d\lambda$. Lemma 22 show that the probability in the integrand can be shown to decay like $\binom{n}{m} e^{-t(n-m)}$, so that the whole numerator is bounded roughly as $(2M)^{N_{W,D}} \binom{n}{m} e^{-t(n-m)}$. Combining these bounds and simplifying, under the definition $q_{\text{eff}} = \left(\frac{2M}{\rho}\right) \vee \left(\frac{M}{\rho'}\right)$ gives Lemma 11.

Existence of parameters with structured stability. In light of Lemma 11, each result from the paper follows from showing that a $\tau/2$-interpolator $\theta \in \mathcal{C}_M(u, D)$ exists with high probability over the draw of the data, $\{(x_i, y_i)\}_{i=1}^n$, and quantifying its stability for each data model proposed. In Theorem 3, we assumed existence of $\theta \in \mathcal{C}_M(w, d)$ that $\tau/2$-interpolates every realization of the data, and we show (Lemma 31) that simply padding this witness to depth D results in an interpolating witness with stability parameters $\log(H \vee L \vee C) = \tilde{O}(d)$, which suffices to give the result. For Theorem 6, we explicitly construct a data-dependent network, or "memorization gadget," $\theta_{\text{mem}} \in \mathcal{C}_M(w_+, D)$ so that $g_{\theta_{\text{mem}}}^{w_+,D}$ is δ_i on all corrupted points, and 0 on all uncorrupted points. In order to memorize m points, the memorization gadget uses trapezoidal bumps to localize m/D points per layer, as constructed in Lemma 12. Assumption 5 allows us to choose these bumps to be sharp enough to successfully isolate only corrupted points with high probability (Lemma 13). Then we define $\theta^\dagger \in \mathcal{C}_M(w + w_+, D)$ to simply add these corrections into the function output (after padding the witness from depth d to D), and prove that the resulting $\theta^\dagger$ still satisfies structured stability such that $\log H, \log L, \log C = \tilde{O}(d)$ (Lemma 33). Notice that the data-dependence of $\theta^\dagger$ is not a problem, according to Lemma 11. An intermediate notion that we call "disturbance stability" — defined in Appendix D.4 — makes proving structured stability more convenient.

5 Conclusion

We extended the techniques of Buzaglo et al. (2024) to handle continuous weights and label noise, drawing connections between overparameterized generalization and the existence of narrow, *stable* witnesses. The principal limitation of this work is that we analyze an unrealistic learning rule; however, analyzing uniform random interpolating parameters allows us to isolate those aspects of generalization that are unequivocally due to deep neural architectures and not optimizer bias. A key technical takeaway of this work is that approximation-theoretic guarantees can be converted into generalization guarantees, which may depend on finegrained stability properties of the approximator.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with information retrieval, proofs and exposition. We have not used generative AI for any other purposes. The proofs of Theorems 3 and 6 were first written by authors, but proofs were substantially improved with the help of AI. For example, the proof that a depth-d witness, when padded to depth-D, has structured stability $H, L, C \approx e^d$ rather than e^D was initially proven separately for (i) a generic witness (ii) a generic witness plus a memorization gadget, by means of an ugly induction argument. Generative AI suggested the introduction of disturbance stability as a stronger (Lemma 27) but more modular (Lemma 32) type of stability, streamlining our proofs. AI was also used in order to speed up formalization; for example, the memorization gadget (Lemma 12) was formalized by AI at a request to build a narrow gadget using trapezoidal functions to localize points at each layer. We searched for prior work containing this memorization gadget, but found no exact matches, so we cite Zhang et al. (2017); Bubeck et al. (2020); Yun et al. (2019), which all share substantial DNA with the construction. At each stage, model output was verified by authors, and all textual content of the paper was manually written by humans. Figures 1 and 2 were generated by AI according to author specifications and with extensive back-and-forth interactions. AI was used to retrieve related work, which was then read by authors. This includes the approximation results used in Examples 1 and 2, such as Schmidt-Hieber (2020); E et al. (2022); Ohn & Kim (2019). Finally, AI was used in gathering feedback on writing quality and polishing prose. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yotam Alexander, Yonatan Slutzky, Yuval Ran-Milo, and Nadav Cohen. Do neural networks need gradient descent to generalize? a theoretical study. In *Advances in Neural Information Processing Systems*, 2025.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

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

Yakir Berchenko. Simplicity bias in overparameterized machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Sebastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and size of the weights in memorization with two-layers neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Gon Buzaglo, Itamar Harel, Mor Shpigel Nacson, Alon Brutzkus, Nathan Srebro, and Daniel Soudry. How uniform random weights induce non-uniform bias: Typical interpolating neural networks generalize with narrow teachers. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zaiwei Chen, Shancong Mou, and Siva Theja Maguluri. Stationary behavior of constant stepsize sgd type algorithms: An asymptotic characterization. *Proc. ACM Meas. Anal. Comput. Syst.*, 6 (1), 2022.

Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *International Conference on Learning Representations*, 2023.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

Eugenio Clerico, Tyler Farghly, George Deligiannidis, Benjamin Guedj, and Arnaud Doucet. Generalisation under gradient descent via deterministic PAC-Bayes. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 349–389. PMLR, 2025.

Giacomo De Palma, Bobak Kiani, and Seth Lloyd. Random deep neural networks are biased towards simple functions. In *Advances in Neural Information Processing Systems*, 2019.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. The size of teachers as a measure of data complexity: PAC-Bayes excess risk bounds and scaling laws. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, 2025a.

Gintare Karolina Dziugaite and Daniel M. Roy. The size of teachers as a measure of data complexity: Pac-bayes excess risk bounds and scaling laws. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3979–3987, 03–05 May 2025b.

Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022. doi: 10.1007/s00365-021-09549-y.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, 2018.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

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

Itamar Harel, William M. Hoza, Gal Vardi, Itay Evron, Nathan Srebro, and Daniel Soudry. Provable tempered overfitting of minimal nets and typical nets. In *Advances in Neural Information Processing Systems*, 2024.

Jason M. Klusowski and Andrew R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12): 7649–7656, 2018.

Michael Kohler and Adam Krzyżak. Optimal global rates of convergence for interpolation problems with random design. *Statistics & Probability Letters*, 83(8):1871–1879, 2013. doi: 10.1016/j.spl.2013.04.018.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, 2018.

Tongtong Liang, Dan Qiao, Yu-Xiang Wang, and Rahul Parhi. Stable minima of ReLU neural networks suffer from the curse of dimensionality: The neural shattering phenomenon. In *Advances in Neural Information Processing Systems*, 2025.

Tongtong Liang, Alexander Cloninger, Rahul Parhi, and Yu-Xiang Wang. Generalization below the edge of stability: The role of data geometry. In *International Conference on Learning Representations*, 2026.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, 2022.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. In *Advances in Neural Information Processing Systems*, 2022.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is SGD a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22, 2021.

Chris Mingard, Henry Rees, Guillermo Valle-Pérez, and Ard A. Louis. Deep neural networks have an inbuilt occam's razor. *Nature Communications*, 16:220, 2025.

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Ilsang Ohn and Yongdai Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7), 2019.

Ilsang Ohn and Yongdai Kim. Nonconvex sparse regularization for deep neural networks and its optimality, 2021.

Rahul Parhi and Ben Adcock. Upper bounds on averaged sampling numbers for general model classes. In *2025 International Conference on Sampling Theory and Applications (SampTA)*, pp. 1–5, 2025. doi: 10.1109/SampTA64769.2025.11133520. URL https://sparsity.ucsd.edu/publications/parhi2025upper.pdf.

Amit Peleg and Matthias Hein. Bias of stochastic gradient descent or the architecture: Disentangling the effects of overparameterization of neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 40154–40184, 2024a.

Amit Peleg and Matthias Hein. Bias of stochastic gradient descent or the architecture: Disentangling the effects of overparameterization of neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR MATH-AI Workshop*, 2021.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. In *Advances in Neural Information Processing Systems*, 2024.

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

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, (70), 2018.

Ryan Theisen, Jason Klusowski, and Michael Mahoney. Good classifiers are abundant in the interpolating regime. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Width is less important than depth in relu neural networks. In *Proceedings of Thirty Fifth Annual Conference on Learning Theory*, 2022.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of pac-bayesian bounds. *Machine Learning*, 113(2):519–604, Feb 2024. ISSN 1573-0565. doi: 10.1007/s10994-023-06391-0.

Felix Voigtlaender. L^p sampling numbers for the Fourier-analytic Barron space. arXiv:2208.07605, 2022. URL https://arxiv.org/abs/2208.07605.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 2011.

Andrew Gordon Wilson. Position: Deep learning is not so mysterious or different. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *Constructive Approximation*, 62:329–360, 2025. doi: 10.1007/s00365-024-09679-z.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In *Neural Information Processing Systems*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: A PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2019.

A Extended related work

Substantial effort has gone into investigating **architecture-induced** versus **optimizer-induced bias** (Buzaglo et al., 2024; Qiao et al., 2024; Alexander et al., 2025; Liang et al., 2026). We focus on related work in the former category, which are most relevant to this paper.

Simplicity Bias Valle-Pérez et al. (2019); De Palma et al. (2019) provide early evidence showing that uniform distributions over parameters push forward onto highly non-uniform distributions over implemented functions; related empirical work suggests that this architecture-induced function prior captures a significant component of the inductive bias observed under SGD (Mingard et al., 2021; 2025). Berchenko (2024) develops this perspective theoretically, by showing that random constructions can exhibit simplicity bias in overparameterized models. Unlike most of these works, we specifically consider uniformity *conditioned on interpolation*, which helps us derive rigorous generalization guarantees from architectural properties. At the same time, recent work has emphasized limits of a purely architectural explanation: Peleg & Hein (2024a) find settings in which increasing width benefits SGD but not random interpolators, while Alexander et al. (2025) show that gradient methods can outperform "dumb sampling" in matrix factorization problems.

Typical interpolating networks Buzaglo et al. (2024) shows that randomly sampled DNNs also generalize in the case of noiseless data from a realizable ground truth. Harel et al. (2024) shows that this analysis can be extended to noisy binary classification, but considers only DNNs with

binary weights. Dziugaite & Roy (2025a) reformulates the proof technique of Buzaglo et al. (2024), and considers DNNs drawn from loss-based Boltzmann distributions rather than only interpolating DNNs. To our knowledge, prior work does not cover deep networks with continuous weights, or label noise in this setting.

PAC-Bayes PAC-Bayes bounds analyze the risk of a hypothesis $h \sim Q_n$ sampled from a data-dependent distribution. Then, for any data-independent measure P, the worst-case complexity term in standard generalization bounds can be replaced by $\mathrm{KL}(Q_n||P)$, to get generalization bounds (McAllester, 1999; Alquier, 2024; Clerico et al., 2025). Although our setup can be formalized within the PAC-Bayes framework, we do not use PAC-Bayes results, and instead perform a direct analysis of the risk. Our primary technical contributions are orthogonal to PAC-Bayes theory, and relate specifically to generalization-promoting aspects of deep ReLU architectures. In practice, it has been observed that using PAC-Bayes bounds with priors P that favor **compressibility** leads to strong generalization bounds even at modern scales (Zhou et al., 2019; Lotfi et al., 2022; 2024), although these are numerical certificates and not analytical rates. The "deep lifting" procedure shows that large volumes of interpolating parameters "behave like" parameters for smaller DNNs (i.e. are compressible), possibly shedding light on the empirical success of these approaches.

B Definition of $U_{\tau,n}^{W,D,M}$

We define the uniform distribution over $S_{\tau,n}^{W,D,M}$, from which our estimate $\hat{\theta}$ will be drawn, as in (1). Since $\mathcal{C}_M(W,D) \subset \mathbb{R}^{N_{W,D}}$, we inherit Lebesgue measure from this ambient parameter space. For any measurable parameter set A, when $S_{\tau,n}^{W,D,M}$ has positive volume, define

$$U_{\tau,n}^{W,D,M}(A) := \mathrm{Unif}(S_{\tau,n}^{W,D,M})(A) := \frac{\lambda_{N_{W,D}}(A \cap S_{\tau,n}^{W,D,M})}{\lambda_{N_{W,D}}(S_{\tau,n}^{W,D,M})}.$$

When $S_{\tau,n}^{W,D,M}$ has zero volume, including when it is empty, define $U_{\tau,n}^{W,D,M}$ to be the point mass at the zero parameter vector.

C Memorization gadget

This section introduces the memorization gadget that memorizes m corruptions $\delta_1 \dots \delta_m \in \mathbb{R}^S$. Though there is prior work on memorization construction within DNNs (Zhang et al., 2017; Bubeck et al., 2020; Yun et al., 2019), the following construction suits our particular needs well.

Lemma 12 (Real-valued memorization gadget)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^p$, and suppose that the corrupted points are indexed by $[m]$. Assume $m \in \{0, \dots, n\}$, integer $D \geq 1$, and $|x_i[1]| \leq X$ for $i \in [n]$. Let the correction vectors satisfy $\|\delta_i\|_\infty \leq \Gamma$ for $i \in [m]$, where $X > 0$ and $\Gamma \geq 0$. Fix $\ell > 0$, and for each $i \in [m]$ choose $a_i \in \mathbb{R}$ such that*

$$x_i[1] \in I_i := [a_i, a_i + \ell].$$

Define the enlarged interval

$$\tilde{I}_i := [a_i - \ell, a_i + 2\ell].$$

Let $B := \lceil \frac{m}{D} \rceil$. Then there exists a width $w_{\mathrm{mem}} := 2 + 2s + 4B$ and depth-D ReLU network $\theta_{\mathrm{mem}} \in \mathcal{A}(w_{\mathrm{mem}}, D)$ such that

$$g_{\theta_{\mathrm{mem}}}^{w_{\mathrm{mem}},D}(x) = 0$$

whenever $x[1] \notin \bigcup_{i=1}^m \tilde{I}_i$.

Moreover, if $x_j[1] \notin \tilde{I}_i$ for every $i \in [m]$ and every $j \neq i$, then

$$g_{\theta_{\mathrm{mem}}}^{w_{\mathrm{mem}},D}(x_i) = \delta_i \qquad \text{for } i \in [m],$$

and

$$g_{\theta_{\mathrm{mem}}}^{w_{\mathrm{mem}},D}(x_j) = 0 \qquad \text{for } j \in \{m+1, \dots, n\}.$$

Moreover, θ_{mem} can be taken to lie within $\mathcal{C}_M(w_{\mathrm{mem}}, D)$ provided that

$$M \geq M_{X,\Gamma,\ell} := \max\{1, X + 2\ell, \frac{\Gamma}{\ell}\}.$$

Proof. The points $x_1, \ldots, x_m$ are the corrupted data points. For each such point, $\delta_i \in \mathbb{R}^s$ is the correction term that we wish to add at x_i.

For each $i \in [m]$, define the hinge points $\tau_i^1 = a_i - \ell$, $\tau_i^2 = a_i$, $\tau_i^3 = a_i + \ell$, and $\tau_i^4 = a_i + 2\ell$. Define

$$\psi_i(z) = \sigma\left(z - \tau_i^1\right) - \sigma\left(z - \tau_i^2\right) - \sigma\left(z - \tau_i^3\right) + \sigma\left(z - \tau_i^4\right).$$

Thus ψ_i is a trapezoidal function that takes value ℓ on $I_i = [a_i, a_i + \ell]$ and is zero outside $\tilde{I}_i = [a_i - \ell, a_i + 2\ell]$.

Divide the m corrupted points evenly among layers, into groups $C_1, \ldots, C_D$ of size at most $B = \lceil m/D \rceil$. We identify each C_k with a subset of $[m]$, but abuse notation by saying "$x_i \in C_k$" whenever $i \in C_k$.

Let $w_{\mathrm{mem}} = 2 + 2s + 4B$. We now construct θ_{mem} by describing how it evolves its internal activations with affine operations.

For internal activations in $\mathbb{R}^{w_{\mathrm{mem}}}$, we divide coordinates into channels $[Z^+, Z^-, P, N, S]$, where $Z^+, Z^- \in \mathbb{R}$, $P, N \in \mathbb{R}^s$, and $S \in \mathbb{R}^{4B}$. The two coordinates Z^+ and Z^- will carry the first input coordinate as $Z^+ = \sigma(x[1])$ and $Z^- = \sigma(-x[1])$, so $Z^+ - Z^- = x[1]$. The channels P and N will carry positive and negative "corrections". Write $\delta_i = \delta_i^+ - \delta_i^-$, where $\delta_i^+, \delta_i^- \in \mathbb{R}^s_{\geq 0}$ are the coordinatewise positive and negative parts of δ_i. The channel will S carry hinge points for groups of corrupted points. .

For a group C_k, let $\tau_{C_k} \in \mathbb{R}^{4B}$ denote the vector of hinge points associated to the points in C_k, padded with zero hinge locations and zero correction coefficients if $|C_k| < B$. We maintain by induction that, after layer k, the activation $[Z_k^+, Z_k^-, P_k, N_k, S_k]$ satisfies $Z_k^+ - Z_k^- = x[1]$,

$$P_k[j] = \sum_{u=1}^{k-1} \sum_{i \in C_u} \frac{\delta_i^+}{\ell}[j]\psi_i(x[1]), \qquad N_k[j] = \sum_{u=1}^{k-1} \sum_{i \in C_u} \frac{\delta_i^-[j]}{\ell}\psi_i(x[1]),$$

and S_k contains the hinge features $\sigma((x[1] - \tau_i^r))$ for every $i \in C_k$ and $r \in [4]$. We interpret the empty sum at $k = 1$ as 0.

Base case. At layer 1, set $Z_1^+ = \sigma(x[1])$, $Z_1^- = \sigma(-x[1])$, and $P_1 = N_1 = 0$. Also set

$$\tilde{S}_1 = \mathbf{1}_{4B} x[1] - \tau_{C_1}, \qquad S_1 = \sigma(\tilde{S}_1),$$

where $\mathbf{1}_{4B} \in \{1\}^{4B}$ is the vector of 1s. Thus S_1 contains the hinge features needed to form ψ_i for every $i \in C_1$, and the induction hypotheses hold at layer 1.

Induction. Now let $k \geq 2$ and assume the induction hypotheses hold at layer $k - 1$. Copy the first-coordinate channels by setting $\tilde{Z}_k^+ = Z_{k-1}^+$ and $\tilde{Z}_k^- = Z_{k-1}^-$. Since these coordinates are nonnegative, ReLU leaves them unchanged, so $Z_k^+ - Z_k^- = x[1]$.

For each $i \in C_{k-1}$, let

$$\Psi_i(S_{k-1}) := S_{k-1}^{i,1} - S_{k-1}^{i,2} - S_{k-1}^{i,3} + S_{k-1}^{i,4}.$$

By induction and the definition of the hinge features, $\Psi_i(S_{k-1}) = \psi_i(x[1])$. For every $j \in [s]$, set

$$\tilde{P}_k[j] = P_{k-1}[j] + \sum_{i \in C_{k-1}} \frac{\delta_i^+[j]}{\ell} \Psi_i(S_{k-1}),$$

and

$$\tilde{N}_k[j] = N_{k-1}[j] + \sum_{i \in C_{k-1}} \frac{\delta_i^-[j]}{\ell} \Psi_i(S_{k-1}).$$

Equivalently, $\tilde{P}_k[j] = P_{k-1}[j] + \sum_{i \in C_{k-1}} \frac{\delta_i^+[j]}{\ell}\psi_i(x[1])$, and similarly for n_k. These quantities are nonnegative, so $P_k = \tilde{P}_k$ and $N_k = \tilde{N}_k$ after ReLU.

Finally, prepare the hinge features for the next group by setting

$$\tilde{S}_k = \mathbf{1}_{4B}(Z_{k-1}^+ - Z_{k-1}^-) - \tau_{C_k}, \qquad S_k = \sigma(\tilde{S}_k).$$

Since $Z^+_{k-1} - Z^-_{k-1} = x[1]$, the channel S_k contains the normalized hinge features needed to form ψ_i for every $i \in C_k$. This completes the induction.

Readout. At layer D, the channels P_D and N_D contain the accumulated corrections from $C_1, \dots, C_{D-1}$, while S_D contains the hinge features for C_D. Define the readout coordinatewise by

$$g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x)[j] = P_D[j] - N_D[j] + \sum_{i \in C_D} \left(\frac{\delta^+_i[j]}{\ell} - \frac{\delta^-_i[j]}{\ell} \right) \Psi_i(S_D).$$

Since $\Psi_i(S_D) = \psi_i(x[1])$ for every $i \in C_D$, this gives

$$g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x) = \sum_{i=1}^{m} \delta_i \frac{1}{\ell} \psi_i(x[1]).$$

Now if $x[1] \notin \bigcup_{i=1}^m \tilde{I}_i$, then every $\psi_i(x[1])$ vanishes, so $g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x) = 0$.

Finally assume $x_j[1] \notin \tilde{I}_i$ for every $i \in [m]$ and every $j \neq i$. For a corrupted point x_i, we have $x_i[1] \in I_i$, so $\psi_i(x_i[1]) = \ell$, while $\psi_r(x_i[1]) = 0$ for every $r \neq i$. Hence $g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x_i) = \delta_i$. If $j \in \{m+1, \dots, n\}$, then $x_j[1] \notin \tilde{I}_i$ for every $i \in [m]$, so every $\psi_i(x_j[1])$ vanishes and $g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x_j) = 0$.

Since $a_i \in [x_i[1] - \ell, x_i[1]]$, all four hinge locations have magnitude at most $X + 2\ell$. The other nonzero weights have magnitudes 1 or at most Γ/ℓ; unused slots have zero coefficients. Thus all parameters are bounded by $M_{X,\Gamma,\ell}$. □

We now use the Λ-marginal boundedness assumption (Assumption 5) to show that the memorization gadget succeeds with high probability over the draw $x_1 \dots x_n$.

Lemma 13. *Assume that $x_1, \dots, x_n \sim \mu$ iid, and that μ is Λ-marginally bounded in its first coordinate. Assume $n \geq 2$, $\Lambda > 0$, $|x_i[1]| \leq X$ almost surely, $\Gamma \geq 0$, and integer $D \geq 1$. Fix a deterministic budget $m \in \{0, \dots, n\}$, $\eta \in (0,1)$, and choose*

$$0 < \ell \leq \frac{\eta}{3\Lambda n(n-1)}, \qquad M \geq M_{X,\Gamma,\ell} := \max\left\{1,\ X + 2\ell,\ \frac{\Gamma}{\ell}\right\}.$$

Define $w_{\mathrm{mem}} := 2 + 2s + 4\left\lceil \frac{m}{D} \right\rceil$.

Then, with probability at least $1 - \eta$ over $x_1, \dots, x_n$, the following holds simultaneously for every subset $I \subseteq [n]$ with $|I| \leq m$ and every collection of correction vectors

$$\{\delta_i : i \in I\}, \qquad \|\delta_i\|_\infty \leq \Gamma.$$

There exists $\theta_{\mathrm{mem}} \in \mathcal{C}_M(w_{\mathrm{mem}}, D)$ such that

$$g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x_i) = \delta_i \qquad \textit{for every } i \in I,$$

and

$$g^{w_{\mathrm{mem}},D}_{\theta_{\mathrm{mem}}}(x_j) = 0 \qquad \textit{for every } j \notin I.$$

Proof. For every $i \in [n]$, set

$$a_i := x_i[1], \qquad \tilde{I}_i := [a_i - \ell, a_i + 2\ell].$$

Let $\mathcal{B}$ be the event that there exist $i \neq j$ such that $x_j[1] \in \tilde{I}_i$. Conditional on x_i, the interval $\tilde{I}_i$ has length 3ℓ, so marginal boundedness gives

$$\Pr[x_j[1] \in \tilde{I}_i \mid x_i] \leq 3\Lambda\ell.$$

Taking a union bound over the $n(n-1)$ ordered pairs gives

$$\Pr[\mathcal{B}] \leq 3\Lambda\ell n(n-1) \leq \eta.$$

On $\mathcal{B}^c$, fix any $I \subseteq [n]$ with $|I| \le m$ and any collection of correction vectors satisfying the stated bound. Enumerate the points indexed by I and apply Lemma 12. The localization condition holds for every selected point, and the required width is at most

$$2 + 2s + 4\left\lceil \frac{m}{D} \right\rceil .$$

Unused memory slots are assigned zero correction coefficients. The resulting parameter lies in $\mathcal{C}_M(w_{\mathrm{mem}}, D)$ by the assumed lower bound on M, and it has the claimed values on all sampled points. Since $\mathcal{B}^c$ does not depend on I or on the correction vectors, the conclusion holds simultaneously for every such choice. □

Lemma 14. *Let $\eta \in (0,1)$, $\epsilon \in [0,1]$, and $m \sim \mathrm{Binom}(n, \epsilon)$. Then*

$$m \le 2\epsilon n + \log \frac{1}{\eta}$$

with probability at least $1 - \eta$.

Proof. Follows from the Chernoff bound. □

For $n \ge 2$, $\Lambda > 0$, and $\eta \in (0,1)$, we define $M_{X,\Gamma,\Lambda,n,\eta} := M_{X,\Gamma,\ell}$ with $\ell := \frac{\eta}{3\Lambda n(n-1)}$. Then we have the following corollary of Lemma 13.

Lemma 15 (Interpolation with memorization)**.** *Fix $\eta \in (0,1)$. Assume $n \ge 2$, $\Lambda > 0$, $\Gamma \ge 0$, and integer $D \ge 1$. Assume $x_1, \dots, x_n \sim \mu_X$ iid, where μ_X is Λ-marginally bounded in its first coordinate and satisfies $x[1] \in [-X, X]$ almost surely. Fix a deterministic budget $m \in \{0, \dots, n\}$, and let $I = I(x^n) \subseteq [n]$ be any subset with $|I| \le m$. Let $\{\delta_i : i \in I\}$ be any, possibly data-dependent, collection of correction vectors satisfying $\|\delta_i\|_\infty \le \Gamma$.*

Let $M \ge M_{X,\Gamma,\Lambda,n,\eta}$ and let $\theta \in \mathcal{C}_M(w, D)$. Then, with probability at least $1 - \eta$ over the draw $x_1 \dots x_n \sim \mu_X^{\otimes n}$, simultaneously for every such I and collection of correction vectors, there exists $\theta^\dagger \in \mathcal{C}_M(w', D)$ such that

$$g_{\theta^\dagger}^{w',D}(x_i) = g_\theta^{w,D}(x_i) + \delta_i \qquad \text{for every } i \in I,$$

and

$$g_{\theta^\dagger}^{w',D}(x_j) = g_\theta^{w,D}(x_j) \qquad \text{for every } j \notin I,$$

where $w' := w + w_{\mathrm{mem}}$, according to the definition of w_{mem} in Lemma 13.

Proof. On the good event of Lemma 13, there exists $\theta_{\mathrm{mem}} \in \mathcal{C}_M(w_{\mathrm{mem}}, D)$ such that $g_{\theta_{\mathrm{mem}}}^{w_{\mathrm{mem}},D}(x_i) = \delta_i$ for $i \in I$, and $g_{\theta_{\mathrm{mem}}}^{w_{\mathrm{mem}},D}(x_j) = 0$ for $j \notin I$. Run θ and θ_{mem} in parallel and add their outputs at the readout. This gives the claimed $\theta^\dagger \in \mathcal{C}_M(w', D)$. □

By Lemma 15, given $(n, m, \Lambda, \Gamma, \eta)$, there exists a memorization gadget θ_{mem} with width w_{mem}, that, with probability at least $1 - \eta$, fits the residuals for the m-of-n corrupted data points. We will refer to this as the $(n, m, \Lambda, \Gamma, \eta)$-memorization gadget. We will refer to the induced interpolating parameter $\theta^\dagger$ from Lemma 15 as the interpolating parameter induced by the $(n, m, \Lambda, \Gamma, \eta)$-memorization gadget.

D Structured Interpolators and Continuous Lifts

D.1 Structured Stability

Definition 16 (Structured stability)**.** Let $\theta \in \mathcal{C}_M(u, D)$. We say that θ has (H, L, C) structured stability, with $H, L, C > 0$, if it satisfies the following conditions.

(i) θ's hidden activations are bounded by H on $[-X, X]^p$: $\|x_k^\theta(x)\|_\infty \le H$ for every $x \in [-X, X]^p$ and $k \in [D]$,

(ii) If

$$\tilde{\theta} \in \mathcal{C}_M(u, D), \qquad \|\tilde{\theta} - \theta\|_\infty \le \rho, \qquad \rho \le \frac{1}{C},$$

then

$$\|x_k^{\tilde{\theta}}\|_\infty \le H + 1 \qquad \text{and} \qquad \sup_{x \in [-X,X]^p} \left\| g_{\tilde{\theta}}^{u,D}(x) - g_\theta^{u,D}(x) \right\|_\infty \le L\rho.$$

D.2 Lifting a Structured Parameter

Throughout this subsection assume $M \ge 1$ and integer widths and depth $1 \le u < W$, $D \ge 1$.

Let $\theta \in \mathcal{C}_M(u, D)$ have matrices $\{W_k^\theta\}$, readout R^θ, and biases $\{b_k^\theta\}$. Let $\phi \in \mathcal{C}_M(W, D)$ have matrices $\{W_k^\phi\}$, readout R^ϕ, and biases $\{b_k^\phi\}$. For $a \in [-M, M]$ and $0 < \rho \le 1$, define the inward interval

$$I_\rho(a) := \begin{cases} [a, a + \rho], & a \le 0, \\ [a - \rho, a], & a > 0. \end{cases}$$

For matrices and vectors, $I_\rho(A)$ denotes the Cartesian product of the corresponding intervals.

Write the parameters of ϕ in block form

$$W_1^\phi = \begin{pmatrix} A_1 \\ C_1 \end{pmatrix}, \qquad W_k^\phi = \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix} \quad (k \ge 2),$$

$$b_k^\phi = \begin{pmatrix} a_k \\ c_k \end{pmatrix}, \qquad R^\phi = (Q_1 \; Q_2),$$

where the first block has width u.

In the context of this section, we call $A_1 \dots A_D, a_1 \dots a_D, Q_1$ "carried" parameters, and the rest "added" parameters. Similarly, we say that the first u entries of the activations are "carried" coordinates and the remainder are "added" coordinates.

Definition 17 ((ρ, ρ')-lift)**.** Let $\theta \in \mathcal{C}_M(u, D)$, let $W > u$, and let $0 < \rho \le 1$ and $0 < \rho' \le M$. We say that ϕ is a (ρ, ρ')-lift of θ from width u to width W if:

I. $A_k \in I_\rho(W_k^\theta)$, $a_k \in I_\rho(b_k^\theta)$, and $Q_1 \in I_\rho(R^\theta)$;

II. B_k, D_k, Q_2 are arbitrary in $[-M, M]$ entrywise;

III. $C_1, C_k \in [-\rho', \rho']$ entrywise, and every coordinate of c_k lies in $[-M, -M + 2\rho']$.

We write $\mathscr{L}_{u \to W}^{\rho, \rho'}(\theta)$ for the set of all such lifts.

For $u \ge 1$ and $H > 0$, define

$$J_{u,H} := \max\{pX + 1,\ u(H + 1) + 1\}.$$

Lemma 18 (Added activations have no effect)**.** *Let $\phi \in \mathscr{L}_{u \to W}^{\rho, \rho'}(\theta)$, and let $\tilde{\theta} \in \mathcal{C}_M(u, D)$ be the carried parameter formed from $A_1, \dots, A_D, a_1, \dots, a_D, Q_1$. Assume that the activations of $\tilde{\theta}$ are bounded by $H + 1$. If*

$$\rho' \le \frac{M}{J_{u,H} + 1},$$

then

$$g_\phi^{W,D}(x) = g_{\tilde{\theta}}^{u,D}(x) \qquad \textit{for every } x \in [-X, X]^p.$$

Proof. Fix $x \in [-X, X]^p$. Write the layer-k activation of ϕ as $x_k^\phi = (\kappa_k, \iota_k)$, where $\kappa_k \in \mathbb{R}^u$ denotes the carried coordinates and $\iota_k \in \mathbb{R}^{W-u}$ the added coordinates. We show inductively that

$$\kappa_k = x_k^{\tilde{\theta}}, \qquad \iota_k = 0$$

for every $k \in [D]$.

At the first layer,

$$W_1^\phi x + b_1^\phi = \begin{pmatrix} A_1 \\ C_1 \end{pmatrix} x + \begin{pmatrix} a_1 \\ c_1 \end{pmatrix}.$$

Hence $\kappa_1 = \sigma(A_1 x + a_1) = x_1^{\tilde{\theta}}$. For each added coordinate, using $\|x\|_\infty \le X$ and the definition of the lift,

$$(C_1 x + c_1)[j] \le pX\rho' - M + 2\rho' \le (J_{u,H} + 1)\rho' - M \le 0.$$

Thus $\iota_1 = 0$.

Now suppose that $x_{k-1}^\phi = (\kappa_{k-1}, \iota_{k-1})$ where

$$\kappa_{k-1} = x_{k-1}^{\tilde{\theta}}, \qquad \iota_{k-1} = 0.$$

Then

$$W_k^\phi x_{k-1}^\phi + b_k^\phi = \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix} \begin{pmatrix} x_{k-1}^{\tilde{\theta}} \\ 0 \end{pmatrix} + \begin{pmatrix} a_k \\ c_k \end{pmatrix}.$$

The carried preactivation is therefore $A_k x_{k-1}^{\tilde{\theta}} + a_k$, so $\kappa_k = x_k^{\tilde{\theta}}$.

For every added coordinate, the assumed activation bound $\|x_{k-1}^{\tilde{\theta}}\|_\infty \le H + 1$ gives

$$(C_k x_{k-1}^{\tilde{\theta}} + c_k)[j] \le u(H+1)\rho' - M + 2\rho' \le (J_{u,H} + 1)\rho' - M \le 0.$$

Hence $\iota_k = 0$, completing the induction.

Finally,

$$x_D^\phi = \begin{pmatrix} x_D^{\tilde{\theta}} \\ 0 \end{pmatrix},$$

and therefore, using $R^\phi = (Q_1\ Q_2)$,

$$g_\phi^{W,D}(x) = (Q_1\ Q_2) \begin{pmatrix} x_D^{\tilde{\theta}} \\ 0 \end{pmatrix} = Q_1 x_D^{\tilde{\theta}} = g_{\tilde{\theta}}^{u,D}(x).$$

□

Lemma 19 (Lift of a structured parameter)**.** *Suppose $\theta \in \mathcal{C}_M(u, D)$ has (H, L, C) structured stability, where $M \ge 1$ and $W > u$. Let $\zeta > 0$, and suppose*

$$\rho \le \min\left\{1, \frac{1}{C}, \frac{\zeta}{L}\right\}, \qquad \rho' \le \frac{M}{J_{u,H} + 1}.$$

Then every $\phi \in \mathscr{L}_{u\to W}^{\rho,\rho'}(\theta)$ satisfies

$$\sup_{x \in [-X,X]^p} \|g_\phi^{W,D}(x) - g_\theta^{u,D}(x)\|_\infty \le \zeta.$$

Proof. The carried parameter $\tilde{\theta}$ satisfies $\|\tilde{\theta} - \theta\|_\infty \le \rho$. Structured stability therefore gives

$$\|x_k^{\tilde{\theta}}\|_\infty \le H + 1 \qquad \sup_x \|g_{\tilde{\theta}}(x) - g_\theta(x)\|_\infty \le L\rho \le \zeta.$$

Due to the activation bound, we can apply Lemma 18 to get $g_\phi(x) = g_{\tilde{\theta}}(x)$ for all $x \in [-X, X]^p$. □

For any $u \ge 1$, let

$$N_{u,D} := up + (D-1)u^2 + Du + us,$$
$$F_{u,W,D} := (D-1)(W-u)W + s(W-u),$$
$$K_{u,W,D} := (W-u)p + (D-1)(W-u)u + (W-u)D.$$

$N_{u,D}$ is the number of parameters in a neural network of width u and depth D, and is also the number of weights chosen from an interval of length ρ in a (ρ, ρ') lift from u to W. $F_{u,W,D}$ is the number of parameters in a (ρ, ρ')-lift to from u to W that are freely chosen in $[-M, M]$. $K_{u,W,D}$ is the number of parameters in a (ρ, ρ') lift chosen from an interval of length ρ'.

Thus

$$N_{W,D} = N_{u,D} + F_{u,W,D} + K_{u,W,D}. \tag{7}$$

Lemma 20 (Volume of structured lifts)**.** *Let $\theta \in \mathcal{C}_M(u, D)$, let $0 < \rho \leq 1$, and let $0 < \rho' \leq M$. Then*

$$\lambda\left(\mathscr{L}^{\rho,\rho'}_{u\to W}(\theta)\right) = (2M)^{F_{u,W,D}}(2\rho')^{K_{u,W,D}}\rho^{N_{u,D}}.$$

Proof. The lift is a Cartesian product of intervals: $F_{u,W,D}$ coordinates have length $2M$, $K_{u,W,D}$ coordinates have length $2\rho'$, and $N_{u,D}$ coordinates have length ρ. □

For fixed data, let

$$B_t := S^{W,D,M}_{\tau,n} \cap \{\theta \in \mathcal{C}_M(W, D) : \Delta_\tau(\theta) > t\}.$$

In the next Lemma, we assume without loss of generality that all labels except the first m_0 are from f. This is equivalent to assuming that the well-labeled points are chosen independently of the realization of $x_1 \dots x_n$.

Lemma 21 (Continuous numerator bound)**.** *Let $t \in [0, 1]$ and $x_1, \dots, x_n \sim \mu_X$ i.i.d. Assume $y_i = f(x_i)$ for $i = m_0 + 1, \dots, n$, with no restriction on the first m_0 labels. Then*

$$\mathbb{E}_{x^n\sim\mu_X^{\otimes n}}[\lambda(B_t)] \leq (2M)^{N_{W,D}}e^{-t(n-m_0)}.$$

Proof. By Tonelli's theorem,

$$\mathbb{E}[\lambda(B_t)] = \int \Pr[\phi \in B_t]\, d\lambda(\phi).$$

If $\Delta_\tau(\phi) > t$, then membership in $S^{W,D,M}_{\tau,n}$ requires τ-accurate prediction on all $n-m_0$ uncorrupted observations. Thus

$$\Pr[\phi \in B_t] \leq (1-t)^{n-m_0} \leq e^{-t(n-m_0)}.$$

Integrating over $[-M, M]^{N_{W,D}}$ proves the result. □

Using a union bound over subsets of size m_0, we may extend the previous lemma to the case where the subset of well-labeled points depends arbitrarily on $x_1 \dots x_n$.

Lemma 22 (Adversarial continuous numerator bound)**.** *Let $t \in [0, 1]$ and $x_1, \dots, x_n \sim \mu_X$ i.i.d. Suppose that, after observing x^n, there is a possibly data-dependent subset $I = I(x^n) \subseteq [n]$ with $|I| \leq m_0$ such that $y_i = f(x_i)$ for every $i \notin I$, with no restriction on the labels $\{y_i : i \in I\}$. Then*

$$\mathbb{E}_{x^n\sim\mu_X^{\otimes n}}[\lambda(B_t)] \leq (2M)^{N_{W,D}}\binom{n}{m_0}e^{-t(n-m_0)}.$$

Proof. For each fixed $J \subseteq [n]$ with $|J| = m_0$, define

$$B_{t,J} := \left\{\phi \in \mathcal{C}_M(W, D) : \Delta_\tau(\phi) > t,\ \|g^{W,D}_\phi(x_i) - f(x_i)\|_\infty \leq \tau \text{ for every } i \notin J\right\}.$$

Since $|I(x^n)| \leq m_0$,

$$B_t \subseteq \bigcup_{\substack{J\subseteq[n]\\|J|=m_0}} B_{t,J},$$

and therefore

$$\lambda(B_t) \leq \sum_{\substack{J\subseteq[n]\\|J|=m_0}} \lambda(B_{t,J}).$$

Taking expectations and applying Lemma 21 to each fixed J, after relabeling the observations, gives

$$\begin{aligned}\mathbb{E}[\lambda(B_t)] &\leq \sum_{\substack{J\subseteq[n]\\|J|=m_0}} \mathbb{E}[\lambda(B_{t,J})] \\ &\leq \binom{n}{m_0}(2M)^{N_{W,D}}e^{-t(n-m_0)}.\end{aligned}$$

□

D.3 Generalization from a Structured Interpolator

Theorem 23 (Generalization from a structured interpolator). *Let $x_1, \dots, x_n \sim \mu_X$ i.i.d., $t \geq 0$, and $0 \leq \eta < 1$. Assume $y_i = f(x_i)$ for $i = m_0 + 1, \dots, n$, with no restriction on the first m_0 labels. Fix $u < W$ and D, M, H, L, C independently of the realized sample, with $M \geq 1$, $\tau > 0$, and $H, L, C > 0$. Suppose there is a training-data event $\mathcal{G}$ with probability at least $1 - \eta$ such that, on $\mathcal{G}$, there exists $\theta^\dagger \in \mathcal{C}_M(u, D)$ with (H, L, C) structured stability and*

$$\max_{i \in [n]} \|g_{\theta^\dagger}^{u,D}(x_i) - y_i\|_\infty \leq \frac{\tau}{2}.$$

Define

$$\rho := \min\left\{1, \frac{1}{C}, \frac{\tau}{2L}\right\}, \qquad \rho' := \frac{M}{J_{u,H} + 1},$$
$$q_{\text{eff}} := \max\left\{\frac{2M}{\rho}, \frac{M}{\rho'}\right\}.$$

Then

$$\Pr_{x^n, \hat{\theta} \sim U_{\tau,n}^{W,D,M}}[\Delta_\tau(\hat{\theta}) > t] \leq \eta + \left(\frac{2M}{\rho}\right)^{N_{u,D}} \left(\frac{M}{\rho'}\right)^{K_{u,W,D}} e^{-t(n-m_0)}.$$

In particular,

$$\Pr[\Delta_\tau(\hat{\theta}) > t] \leq \eta + q_{\text{eff}}^{N_{u,D}+K_{u,W,D}} e^{-t(n-m_0)}.$$

Proof. For $t \geq 1$ the event is empty, so consider $0 \leq t < 1$. On $\mathcal{G}$, Lemma 19, with $\zeta = \tau/2$, gives

$$\mathscr{L}_{u \to W}^{\rho,\rho'}(\theta^\dagger) \subseteq S_{\tau,n}^{W,D,M}.$$

Lemma 20 therefore gives

$$\lambda(S_{\tau,n}^{W,D,M}) \geq V := (2M)^{F_{u,W,D}} (2\rho')^{K_{u,W,D}} \rho^{N_{u,D}} > 0 \qquad \text{on } \mathcal{G}.$$

Split the probability before using a volume ratio. With B_t as defined above,

$$\begin{aligned}\Pr[\Delta_\tau(\hat{\theta}) > t] &\leq \eta + \Pr(\mathcal{G})\, \mathbb{E}\left[\frac{\lambda(B_t)}{\lambda(S_{\tau,n}^{W,D,M})} \,\middle|\, \mathcal{G}\right] \\ &\leq \eta + \frac{\mathbb{E}[\mathbf{1}_{\mathcal{G}} \lambda(B_t)]}{V} \\ &\leq \eta + \frac{\mathbb{E}[\lambda(B_t)]}{V}.\end{aligned}$$

The conditional ratio is evaluated only on $\mathcal{G}$, where its denominator is positive. By Lemma 21, the last expression is at most

$$\eta + \frac{(2M)^{N_{W,D}} e^{-t(n-m_0)}}{(2M)^{F_{u,W,D}} (2\rho')^{K_{u,W,D}} \rho^{N_{u,D}}}.$$

Using (7) yields the stated bound. □

Corollary 24 (Structured interpolator rate). *In the setting of Theorem 23, assume $0 \leq \eta < \delta < 1$ and $m_0 \leq uD$. Then, for a constant C_0 depending only on p and s, we have*

$$\Pr[\Delta_\tau(\hat{\theta}) > t_n] \leq \delta, \qquad t_n := C_0 \frac{uWD \log q_{\text{eff}} + \log \frac{1}{\delta - \eta}}{n}$$

Here

$$q_{\text{eff}} = \max\{2M \max\{1, C, 2L/\tau\}, J_{u,H} + 1\}. \tag{8}$$

Proof. By definition,

$$N_{u,D} + K_{u,W,D} = Wp + (D-1)Wu + DW + su \leq C_1 uWD$$

for a constant C_1 depending only on p and s. If $t_n > 1$, the result is immediate. Otherwise,

$$t_n(n - m_0) \geq nt_n - m_0.$$

Since $m_0 \leq uD$, $W \geq 1$, and $q_{\text{eff}} \geq 2$, the term $uWD \log q_{\text{eff}}$ absorbs m_0. Taking C_0 sufficiently large makes the exponential term in Theorem 23 at most $\delta - \eta$.

The bound on q_{eff} follows from its definition in Theorem 23. □

Corollary 25 (Adaptive structured interpolator rate)**.** *In the setting of Theorem 23, with $M \geq 1$, replace the assumption that the first m_0 labels are unrestricted by the assumption that the labels are induced by an (m, Γ) adversarial corruption (Assumption 4), and set $m_0 := m$. Define*

$$\bar{q}_{\text{eff}} := \max\{en, q_{\text{eff}}\}.$$

Assume $0 \leq \eta < \delta < 1$ and $m_0 \leq uD$. Then, for a constant C_0 depending only on p and s,

$$t_n := C_0 \frac{uWD \log \bar{q}_{\text{eff}} + \log \frac{1}{\delta - \eta}}{n}$$

satisfies

$$\Pr[\Delta_\tau(\hat{\theta}) > t_n] \leq \delta.$$

Proof. The proof follows Theorem 23 and Corollary 24, using Lemma 22 in place of Lemma 21. If $m_0 = 0$, the additional factor $\binom{n}{m_0}$ is one. If $m_0 \geq 1$, it is absorbed by the complexity term since

$$\log \binom{n}{m_0} \leq m_0 \log \frac{en}{m_0} \leq uD \log(en) \leq uWD \log \bar{q}_{\text{eff}}.$$

□

By construction,

$$q_{\text{eff}} = \max\left\{2M \max\left\{1, C, \frac{2L}{\tau}\right\}, J_{u,H} + 1\right\}.$$

Thus the rate depends only logarithmically on the stability certificate. Every continuous interpolation theorem below follows by furnishing deterministic u, D, M, H, L, C and a high-probability $\tau/2$-interpolating structured parameter, and then invoking Corollary 24.

D.4 Disturbance Stability and Existence of Structured Interpolators

Structured Stability (Definition 16) relates to the stability of a computation of a DNN under a perturbation to its weights. We now draw a possibly surprising connection between Structured Stability and *Disturbance Stability*, which relates to perturbation of *inputs* to the network. A *disturbance* is a bounded adversarial perturbation injected to a DNN's computation at each layer, as a function of the current activations. For a parameter θ, we write

$$\theta = (W_1, \ldots, W_D, b_1, \ldots, b_D, R) \in \mathcal{C}_M(u, D)$$

and, for $x \in [-X, X]^p$:

$$x_0^\theta(x) := x, \qquad x_k^\theta(x) := \sigma(W_k x_{k-1}^\theta(x) + b_k), \quad k \in [D].$$

Given functions

$$\xi_k : [-X, X]^p \to \mathbb{R}^u, \quad k \in [D], \qquad \xi_{D+1} : [-X, X]^p \to \mathbb{R}^s,$$

we define the corresponding disturbed computation by

$$x_0^{\theta,\xi}(x) := x, \qquad x_k^{\theta,\xi}(x) := \sigma\left(W_k x_{k-1}^{\theta,\xi}(x) + b_k + \xi_k(x)\right), \quad k \in [D],$$

and

$$g_\theta^\xi(x) := R x_D^{\theta,\xi}(x) + \xi_{D+1}(x).$$

We call this a ν**-disturbed** computation if

$$\sup_{x \in [-X,X]^p} \|\xi_k(x)\|_\infty \leq \nu \qquad \text{for every } k \in [D+1].$$

Definition 26 (Disturbance Stability)**.** We say that θ has (H, l, c) disturbance stability, with $H, l, c > 0$, if its exact hidden activations satisfy

$$\sup_{x\in[-X,X]^p} \|x_k^\theta(x)\|_\infty \le H \qquad \text{for every } k \in [D],$$

and every ν-disturbed computation with $\nu \le 1/c$ satisfies

$$\sup_{x\in[-X,X]^p} \max_{k\in[D]} \|x_k^{\theta,\xi}(x) - x_k^\theta(x)\|_\infty \le l\nu, \quad \|x_k^{\theta,\xi}(x)\|_\infty \le H+1, \quad \sup_{x\in[-X,X]^p} \|g_\theta^{\xi,u,D}(x) - g_\theta^{u,D}(x)\|_\infty \le l\nu.$$

Definition 26 is related to Adversarial Robustness (Madry et al., 2018), but calls for an even stronger guarantee of robustness given possibly adaptive perturbations injected at each layer[4]. However, Disturbance Stability has useful compositional properties compared to Structured Stability, and we will be able to establish that it holds for parameters of interest, θ.

Notice also the extra requirement that intermediate activations remain similar under disturbances. Structured Stability does make any requirements on intermediate activations, but requiring it for Disturbance Stability facilitates constitutionality.

We now show that Disturbance Stability implies Structured Stability.

Define the quantity

$$J_{u,H} := \max\{pX + 1,\ u(H+1) + 1\}. \tag{9}$$

Lemma 27 (Disturbance stability implies structured stability)**.** *Suppose $\theta \in \mathcal{C}_M(u, D)$ has (H, l, c) disturbance stability. Then θ has (H, L, C) structured stability with*

$$C = J_{u,H}c, \qquad L = J_{u,H}l.$$

Proof. Let

$$\tilde\theta = (\tilde W_1, \dots, \tilde W_D, \tilde b_1, \dots, \tilde b_D, \tilde R) \in \mathcal{C}_M(u, D)$$

be a coordinatewise ρ-perturbation of θ, so that $\|\tilde\theta - \theta\|_\infty \le \rho$. Write

$$\tilde W_k = W_k + \Delta_k, \qquad \tilde b_k = b_k + \zeta_k, \qquad \tilde R = R + \Delta_R,$$

where every entry of Δ_k, ζ_k, and Δ_R has magnitude at most ρ.

We define a disturbance

$$\xi_1(x) := \Delta_1 x + \zeta_1, \qquad \xi_k(x) := \Delta_k x_{k-1}^{\tilde\theta}(x) + \zeta_k \quad (k \ge 2), \qquad \xi_{D+1}(x) := \Delta_R x_D^{\tilde\theta}(x),$$

such that, by construction,

$$x_k^{\theta,\xi}(x) = x_k^{\tilde\theta}(x) \quad (k \in [D]), \qquad g_\theta^\xi(x) = g_{\tilde\theta}(x).$$

We now must show that if $\rho \le \frac{1}{J_{u,H}c}$, then (ξ_k) is a disturbance of size at most $J_{u,H}\rho \le 1/c$. So assume $\rho \le \frac{1}{J_{u,H}c}$.

We begin by establishing the activation bound $\|x_k^{\tilde\theta}(x)\|_\infty \le H + 1$. Define

$$k_* := \min\left(\left\{k \in [D] : \sup_{x\in[-X,X]^p} \|x_k^{\tilde\theta}(x)\|_\infty > H + 1\right\} \cup \{D + 1\}\right),$$

so that $\sup_x \|x_j^{\tilde\theta}(x)\|_\infty \le H + 1$ for every $j < k_*$. At the first layer,

$$\|\xi_1(x)\|_\infty \le p\rho\|x\|_\infty + \rho \le (pX + 1)\rho \le J_{u,H}\rho \le \frac{1}{c}.$$

[4]The adversarial robustness literature focuses on learning adversarially robust DNNs. We deal exclusively with existence of robust DNNs. Our proof technique does not establish that $\hat\theta$ will have robustness.

For any $2 \le k \le \min\{k_*, D\}$, we have $k-1 < k_*$, and hence $\|x_{k-1}^{\tilde\theta}(x)\|_\infty \le H+1$. Therefore

$$\|\xi_k(x)\|_\infty \le u\rho\|x_{k-1}^{\tilde\theta}(x)\|_\infty + \rho \le \big(u(H+1)+1\big)\rho \le J_{u,H}\rho \le \frac{1}{c}.$$

Suppose that $k_* \le D$. Define a complete disturbance sequence $\bar\xi$ by

$$\bar\xi_k := \xi_k \quad (k \le k_*), \qquad \bar\xi_k := 0 \quad (k > k_*).$$

Then $\bar\xi$ is a $J_{u,H}\rho$-disturbed computation. Since $J_{u,H}\rho \le 1/c$, disturbance stability gives

$$\sup_x \|x_{k_*}^{\theta,\bar\xi}(x)\|_\infty \le H+1$$

But the computations of $(\theta, \bar\xi)$ and $\tilde\theta$ agree through layer k_*, so

$$x_{k_*}^{\theta,\bar\xi}(x) = x_{k_*}^{\tilde\theta}(x),$$

contradicting the assumption that $k_* \le D$. Thus $k_* = D+1$, and our previous calculations tell us that

$$\sup_x \|x_k^{\tilde\theta}(x)\|_\infty \le H+1, \qquad \sup_x \|\xi_k(x)\|_\infty \le J_{u,H}\rho \le \frac{1}{c} \qquad \text{for every } k \in [D].$$

Finally, at the output,

$$\|\xi_{D+1}(x)\|_\infty = \|\Delta_R x_D^{\tilde\theta}(x)\|_\infty \le u\rho\|x_D^{\tilde\theta}(x)\|_\infty \le u(H+1)\rho \le J_{u,H}\rho \le \frac{1}{c}.$$

Hence the computation of $\tilde\theta$ is a ν-disturbed computation of θ with $\nu := J_{u,H}\rho \le \frac{1}{c}$. Disturbance stability now gives

$$\sup_x \|g_{\tilde\theta}(x) - g_\theta(x)\|_\infty = \sup_x \|g_\theta^\xi(x) - g_\theta(x)\|_\infty \le l\nu = J_{u,H} l\rho.$$

Since the exact activations of θ are already bounded by H, this proves structured stability with

$$C = J_{u,H}c, \qquad L = J_{u,H}l.$$

□

In the previous proof, we did not need to use the closeness of the disturbed activations, since structured stability does not itself have any such requirement.

Lemma 28 (Activation bounds for bounded parameters and identity padding)**.** *Let $\theta_0 \in \mathcal{C}_M(w,d)$, and let $x_k^{\theta_0}(x)$ denote its layer-k hidden activation. Define*

$$A_1 := M(pX+1), \qquad A_k := M(wA_{k-1}+1).$$

Then

$$\sup_{x\in[-X,X]^p} \|x_k^{\theta_0}(x)\|_\infty \le A_k \qquad \textit{for every } k \in [d].$$

If $\theta \in \mathcal{C}_M(w,D)$ is the d-to-D identity padding of θ_0, then every hidden activation of θ is bounded by

$$H_{\mathrm{pad}} := \max\{1, A_d\}.$$

Moreover,

$$A_d \le d(Mw)^d(Xp+2).$$

Proof. At the first layer,

$$x_1^{\theta_0}(x) = \sigma(W_1x + b_1).$$

For every output coordinate j,

$$\begin{aligned}|(W_1x+b_1)[j]| &\le \sum_{a=1}^{p} |W_1[j,a]||x[a]| + |b_1[j]| \\ &\le pMX + M \\ &= M(pX+1) \\ &= A_1.\end{aligned}$$

Since $|\sigma(z)| \le |z|$, we obtain

$$\|x_1^{\theta_0}(x)\|_\infty \le A_1.$$

Now suppose $\|x_{k-1}^{\theta_0}(x)\|_\infty \le A_{k-1}$ for some $2 \le k \le d$. For each coordinate j,

$$\begin{aligned}|(W_k x_{k-1}^{\theta_0}(x) + b_k)[j]| &\le \sum_{a=1}^{w} |W_k[j,a]||x_{k-1}^{\theta_0}(x)[a]| + |b_k[j]| \\ &\le wMA_{k-1} + M \\ &= M(wA_{k-1}+1) \\ &= A_k.\end{aligned}$$

Applying ReLU gives

$$\|x_k^{\theta_0}(x)\|_\infty \le A_k \qquad (k \le d).$$

The layers after depth d have identity weight matrices and zero biases. Every hidden activation is coordinatewise nonnegative, so for $k > d$,

$$x_k^\theta(x) = \sigma(I_w x_{k-1}^\theta(x)) = x_{k-1}^\theta(x).$$

Thus

$$x_k^\theta(x) = x_d^\theta(x) \qquad \text{for every } k \ge d,$$

and every hidden activation is bounded by A_d, hence also by H_{pad}.

Finally, unrolling the recurrence gives

$$A_d = (Mw)^{d-1}A_1 + M\sum_{j=0}^{d-2}(Mw)^j.$$

Since $M, w \ge 1$,

$$A_d \le (Mw)^d(Xp+1) + d(Mw)^d \le d(Mw)^d(Xp+2).$$

□

We now show that, for a completely generic parameter, disturbance stability parameters decay exponentially in depth.

Lemma 29 (Structured stability of d-deep parameters)**.** *Fix $X > 0$, $M \ge 1$, and $D \ge d$. Let $\theta_0 \in \mathcal{C}_M(w,d)$.*

Define

$$A_1 := M(pX+1), \qquad A_k := M(wA_{k-1}+1), \qquad S_d := \sum_{j=0}^{d-1}(Mw)^j.$$

Then θ_0 has (H_0, l_0, c_0) disturbance stability with

$$H_0 := A_d, \qquad c_0 := \max\{1, S_d\}, \qquad l_0 := MwS_d + 1,$$

Proof. Consider the computation of θ_0 under a disturbance $\xi = \{\xi_k\}$ of size $\nu \le \frac{1}{c_0}$. Define

$$e_k := \sup_{x \in [-X,X]^p} \|x_k^{\theta_0,\xi}(x) - x_k^{\theta_0}(x)\|_\infty, \qquad e_0 := 0.$$

Then $e_k \le Mwe_{k-1} + \nu$,so that

$$e_k \le \left(\sum_{j=0}^{k-1}(Mw)^j\right)\nu \le S_d\nu \qquad \text{for every } k \in [d].$$

By our choice of ν, $e_k \le 1$ and so

$$\sup_x \|x_k^{\theta_0,\xi}(x)\|_\infty \le A_d + 1 = H_0 + 1.$$

Since $S_d \le l_0$, we also have that, for all $k \in [d]$

$$\sup_x \|x_k^{\theta_0,\xi}(x) - x_k^{\theta_0}(x)\|_\infty \le l_0\nu$$

Finally, at the readout,

$$\sup_x \|g_{\theta_0}^{\xi}(x) - g_{\theta_0}(x)\|_\infty \le Mwe_d + \nu \le (MwS_d + 1)\nu = l_0\nu.$$

Thus θ_0 has (H_0, l_0, c_0) disturbance stability. □

Although a generic d-deep parameter has stability that deteriorates as $d(Mw)^d$, if we pad that same parameter with identity layers so that it becomes an equivalent D-deep parameter, stability deteriorates only polynomially in D and exponentially in d. This is what the next Lemma and its corollary show.

Lemma 30 (Disturbance stability under identity padding)**.** *Fix $M \ge 1$ and $D \ge d$. Let $\theta_0 \in \mathcal{C}_M(w, d)$ have (H, l, c) disturbance stability, and let $\theta \in \mathcal{C}_M(w, D)$ be its d-to-D identity padding. Define*

$$H_{\text{pad}} := H, \qquad c_{\text{pad}} := \max\{1, c, l + D - d\}, \qquad l_{\text{pad}} := Mw(l + D - d + 1)$$

Then θ has $(H_{\text{pad}}, l_{\text{pad}}, c_{\text{pad}})$ disturbance stability. Consequently, it has $(H_{\text{pad}}, L_{\text{pad}}, C_{\text{pad}})$ structured stability with

$$C_{\text{pad}} = J_{w,H_{\text{pad}}} c_{\text{pad}}, \qquad L_{\text{pad}} = J_{w,H_{\text{pad}}} l_{\text{pad}}.$$

Proof. By construction,

$$x_k^{\theta}(x) = x_d^{\theta_0}(x), \qquad k \ge d.$$

Thus every activation of θ is bounded by $H \le H_{\text{pad}}$.

We now consider the computation of θ under a disturbance $\xi = \{\xi_k\}$ of size $\nu \le \frac{1}{c_{\text{pad}}}$.

For $k \le d$, disturbance stability of θ_0 and the definition of θ gives us

$$\sup_x \|x_k^{\theta,\xi}(x) - x_k^{\theta}(x)\|_\infty \le l\nu, \qquad \sup_x \|x_k^{\theta,\xi}(x)\|_\infty \le H + 1.$$

Moreover, for $k > d$, we now show $\sup_x \|x_k^{\theta,\xi}(x) - x_k^{\theta}(x)\|_\infty \le (l + k - d)$ by induction. The case $k = d$ is shown. For the inductive step, we have

$$\begin{aligned}
\sup_x \|x_k^{\theta,\xi}(x) - x_k^{\theta}(x)\|_\infty &= \sup_x \|\sigma(x_{k-1}^{\theta,\xi}(x) + \xi_k(x)) - \sigma(x_{k-1}^{\theta}(x))\|_\infty \\
&\le \sup_x \|x_{k-1}^{\theta,\xi}(x) + \xi_k(x) - x_{k-1}^{\theta}(x)\|_\infty \\
&\le \sup_x \|x_{k-1}^{\theta,\xi}(x) - x_{k-1}^{\theta}(x)\|_\infty + \nu \\
&\le (l + k - 1 - d)\nu + \nu = (l + k - d)\nu
\end{aligned}$$

In particular, we have shown that for all $k \in [D]$,

$$\sup_x \|x_k^{\theta,\xi}(x) - x_k^{\theta}(x)\|_\infty \le (l + D - d)\nu \le l_{\text{pad}}\nu$$

.

For the activation bound assumption, we have, using the size bound on ν and the above equation

$$\sup_x \|x_k^{\theta,\xi}(x)\|_\infty \le \sup_x \|x_k^{\theta}(x)\|_\infty + \sup_x \|x_k^{\theta}(x) - x_k^{\theta,\xi}(x)\|_\infty \le H + 1$$

Finally, it remains to show that the readout of θ is stable under the disturbance:

$$\sup_x \|g_\theta^{\xi}(x) - g_\theta(x)\|_\infty = \sup_x \|R(x_D^{\theta,\xi}(x) - x_D^{\theta})(x) + \xi_{D+1}(x)\|_\infty \le Mw(l + D - d + 1)\nu.$$

□

For the noiseless result, a direct parameter-perturbation argument gives a sharper structured-stability bound by retaining the activation size at each layer.

Lemma 31 (Direct structured stability under identity padding)**.** *Let $X > 0$, $M \geq 1$, and $D \geq d \geq 1$. Let $\theta^\dagger \in \mathcal{C}_M(w, D)$ be the d-to-D identity padding of $\theta_0 \in \mathcal{C}_M(w, d)$. Write $a := Mw$, $b := pX+1$, and*

$$A_1 := Mb, \qquad A_k := aA_{k-1} + M \quad (2 \leq k \leq d).$$

Then $\theta^\dagger$ has $(H_{\mathrm{pad}}, L_{\mathrm{pad}}, C_{\mathrm{pad}})$ structured stability with

$$\begin{aligned} H_{\mathrm{pad}} &:= A_d, \\ C_{\mathrm{pad}} &:= 3D(b+d)a^d, \\ L_{\mathrm{pad}} &:= (3aD+1)(b+d)a^d. \end{aligned}$$

In particular, all three parameters are bounded by $(Mw)^d \mathrm{poly}(X, M, p, d, D, w)$.

Proof. Abbreviate the claimed parameters by H, L, C and set $T := D-d$. The activation recurrence gives

$$A_k \leq M(b+k-1)a^{k-1}.$$

Since $a \geq 1$, the sequence A_k is nondecreasing. Thus all exact hidden activations are bounded by $H = A_d$, including the identity-padded layers.

Fix $\widetilde{\theta} \in \mathcal{C}_M(w, D)$ with $\|\widetilde{\theta} - \theta^\dagger\|_\infty \leq \rho \leq C^{-1}$, and define

$$e_k := \sup_{x \in [-X,X]^p} \|x_k^{\widetilde{\theta}}(x) - x_k^{\theta^\dagger}(x)\|_\infty.$$

For the first layer, $e_1 \leq b\rho$. For $2 \leq k \leq d$, decompose the preactivation difference as

$$\widetilde{W}_k(\widetilde{x}_{k-1} - x_{k-1}) + (\widetilde{W}_k - W_k)x_{k-1} + (\widetilde{b}_k - b_k).$$

The absolute row sums of $\widetilde{W}_k$ are at most a because the perturbed parameter remains in $\mathcal{C}_M(w, D)$. Since ReLU is 1-Lipschitz,

$$e_k \leq ae_{k-1} + (wA_{k-1} + 1)\rho = ae_{k-1} + \frac{A_k}{M}\rho.$$

Unrolling this recurrence yields, for $1 \leq k \leq d$,

$$e_k \leq B_k\rho, \qquad B_k := \left(kb + \frac{k(k-1)}{2}\right)a^{k-1} \leq k(b+k)a^{k-1}.$$

For each padded layer, $W_k = I_w$ and $b_k = 0$, so the absolute row sums of $\widetilde{W}_k$ are at most $1 + w\rho$. Therefore

$$e_k \leq (1+w\rho)e_{k-1} + (wH+1)\rho \qquad (k > d).$$

Since $C \geq wT$, we have $wT\rho \leq 1$ and hence $(1+w\rho)^T \leq \exp(wT\rho) \leq e < 3$. For $0 \leq t \leq T$, it follows that

$$\begin{aligned} e_{d+t} &\leq (1+w\rho)^t\big(B_d + t(wH+1)\big)\rho \\ &\leq 3\big(B_d + T(wH+1)\big)\rho. \end{aligned}$$

Using

$$wH + 1 \leq (b+d)a^d, \qquad B_d \leq d(b+d)a^d,$$

we obtain $e_k \leq C\rho \leq 1$ for every hidden layer (also $B_k \leq B_d$ for $k \leq d$). Thus every perturbed hidden activation is bounded by $H + 1$.

Finally, at the readout, the absolute row sums of $\widetilde{R}$ are at most a, and those of $\widetilde{R} - R$ are at most $w\rho$. Hence

$$\begin{aligned} \sup_x \|g_{\widetilde{\theta}}(x) - g_{\theta^\dagger}(x)\|_\infty &\leq ae_D + wH\rho \\ &\leq (aC + wH)\rho \leq L\rho. \end{aligned}$$

This proves structured stability. The claimed dependence follows from $H \leq (b+d)a^d$ and the explicit formulas for C, L. □

Lemma 32 (Disturbance stability under parallel composition). *Let $\theta_i \in \mathcal{C}_M(w_i, D)$ have (H_i, l_i, c_i) disturbance stability for $i \in \{1, 2\}$, and suppose they have the same output dimension. Let $\theta_1 \oplus \theta_2$ denote the parameter obtained by running them in parallel and summing their outputs. Then $\theta_1 \oplus \theta_2$ has*

$$(\max\{H_1, H_2\},\ l_1 + l_2,\ \max\{c_1, c_2\})$$

disturbance stability.

Proof. A ν-disturbed computation of the parallel network decomposes into ν-disturbed computations of the two subnetworks. If $\nu \le 1/\max\{c_1, c_2\}$, both disturbance-stability guarantees apply, so the combined hidden disturbed activations are bounded by $\max\{H_1, H_2\}+1$, and are $l_1 \vee l_2$-close to their undisturbed counterparts. Since the outputs are summed, their errors are at most $(l_1 + l_2)\nu$; assign the output disturbance to either one of the two subnetworks. The claim follows. □

Lemma 33 (Structured stability after memorization). *Let $\theta \in \mathcal{C}_M(w, D)$ have (H_0, l_0, c_0) disturbance stability, and assume $M \ge M_{X,\Gamma,\ell}$. Let θ_{mem} be Lemma 12's (n, m, ℓ) memorization gadget, and define*

$$b := \left\lceil \frac{m}{D} \right\rceil, \qquad w' := w + 2 + 2s + 4b.$$

The memorization gadget has $(H_{\mathrm{mem}}, l_{\mathrm{mem}}, c_{\mathrm{mem}})$ disturbance stability, where

$$H_{\mathrm{mem}} := \max\{1, 2X + 2\ell, \Gamma m\},$$

$$c_z := D, \qquad c_s := 2D + 1, \qquad c_{p,n} := D\left(1 + \frac{4b\Gamma}{\ell} c_s\right),$$

$$c_{\mathrm{mem}} := \max\{1, c_z, c_s, c_{p,n}\}, \qquad l_{\mathrm{mem}} := 2c_{p,n} + \frac{4b\Gamma}{\ell} c_s + 1.$$

Let $\theta^\dagger \in \mathcal{C}_M(w', D)$ be obtained by running θ and θ_{mem} in parallel and summing their outputs. Then $\theta^\dagger$ has (H, l, c) disturbance stability, where

$$H := \max\{H_0, H_{\mathrm{mem}}\} = \max\{H_0, 1, 2X + 2\ell, \Gamma m\},$$

$$c := \max\{c_0, c_{\mathrm{mem}}\} = \max\{1, c_0, c_z, c_s, c_{p,n}\},$$

and

$$l := l_0 + l_{\mathrm{mem}} = l_0 + 2c_{p,n} + \frac{4b\Gamma}{\ell} c_s + 1.$$

Consequently, $\theta^\dagger$ has (H, L, C) structured stability with

$$C = J_{w',H} c, \qquad L = J_{w',H} l.$$

If θ is the d-to-D identity padding of some $\theta_0 \in \mathcal{C}_M(w, d)$, then, for an absolute constant C_0,

$$H, C, L \le K\, p w' d D^2 X m^2 \Gamma^2 (1 + \ell^{-1})(Mw)^{2d+1}$$

where K is an absolute constant.

Proof. We first establish disturbance stability of θ_{mem} alone.

Write the exact layer-k hidden activation of the memorization gadget as

$$x_k^{\theta_{\mathrm{mem}}} = (z_k, p_k, n_k, s_k),$$

where $z_k = (Z_k^+, Z_k^-) \in \mathbb{R}^2$ is the first-coordinate copy channel, $p_k, n_k \in \mathbb{R}^s$ are the positive and negative accumulation channels, and $s_k \in \mathbb{R}^{4b}$ is the hinge-feature channel.

We first bound the exact activations. For the memorization gadget, the two copy coordinates are defined as $Z_k^+ = \sigma(x[1])$ and $Z_k^- = \sigma(-x[1])$, so that

$$\|z_k\|_\infty \le X.$$

Every hinge point associated with an active memory slot lies between $-X - 2\ell$ and $X + 2\ell$. For unused slots, choose the padded hinge points in this same interval and set their correction coefficients

to zero. Since the input coordinate lies in $[-X, X]$, every hinge feature is therefore bounded by $2X + 2\ell$, so

$$\|s_k\|_\infty \le 2X + 2\ell.$$

Each normalized trapezoidal contribution $\ell^{-1}\psi_i(x[1])$ lies in $[0, 1]$. Each coordinate of δ_i^+ and δ_i^- is at most Γ, and at most m corrections are accumulated. Consequently,

$$\|p_k\|_\infty, \|n_k\|_\infty \le \Gamma m.$$

Thus the exact hidden activations of the memorization gadget are bounded by

$$H_{\mathrm{mem}} = \max\{1, 2X + 2\ell, \Gamma m\}.$$

Now consider a ν-disturbed computation of θ_{mem}. Define

$$e_k^z := \sup_x \|z_k^{\theta_{\mathrm{mem}},\xi}(x) - z_k(x)\|_\infty, \qquad e_k^s := \sup_x \|s_k^{\theta_{\mathrm{mem}},\xi}(x) - s_k(x)\|_\infty,$$

$$e_k^p := \sup_x \|p_k^{\theta_{\mathrm{mem}},\xi}(x) - p_k(x)\|_\infty, \qquad e_k^n := \sup_x \|n_k^{\theta_{\mathrm{mem}},\xi}(x) - n_k(x)\|_\infty,$$

and

$$e_k^{p,n} := \max\{e_k^p, e_k^n\}.$$

Set the corresponding layer-zero errors equal to zero.

Disturbed activation errors. We prove simultaneously by induction on k that

$$e_k^z \le k\nu, \qquad e_k^s \le (2k-1)\nu,$$

and

$$e_k^{p,n} \le k\left(1 + \frac{4b\Gamma}{\ell} c_s\right)\nu,$$

where

$$c_s := 2D + 1.$$

At the first layer, the exact and disturbed memory channels receive the same input, so their differences arise only from the additive disturbance. Hence

$$e_1^z, e_1^s, e_1^p, e_1^n \le \nu.$$

The claimed bounds therefore hold for $k = 1$.

Now let $k \ge 2$ and assume the bounds hold at layer $k - 1$.

Copy channel. The copy channels are propagated by identity weights, so, since ReLU is 1-Lipschitz,

$$e_k^z \le e_{k-1}^z + \nu \le k\nu.$$

Hinge channel. Each hinge preactivation depends on $Z_{k-1}^+ - Z_{k-1}^-$ and a fixed bias. Thus

$$e_k^s \le 2e_{k-1}^z + \nu \le (2k-1)\nu.$$

Accumulation channels. Consider one coordinate of the positive accumulation channel. Its exact preactivation has the form

$$p_{k-1}[j] + \sum_{i \in C_{k-1}} \frac{\delta_i^+[j]}{\ell} \Psi_i(s_{k-1}),$$

where $|C_{k-1}| \le b$ and

$$\Psi_i(s) = s^{i,1} - s^{i,2} - s^{i,3} + s^{i,4}$$

is the trapezoid built from the four hinge points corresponding to point i. For every i,

$$|\Psi_i(s_{k-1}^{\theta_{\mathrm{mem}},\xi}) - \Psi_i(s_{k-1})| \le 4e_{k-1}^s.$$

Since $0 \le \delta_i^+[j] \le \Gamma$ and there are at most b active terms,

$$e_k^p \le e_{k-1}^p + \frac{4b\Gamma}{\ell} e_{k-1}^s + \nu.$$

The same bound holds for the negative accumulation channel, and hence

$$e_k^{p,n} \le e_{k-1}^{p,n} + \frac{4b\Gamma}{\ell} e_{k-1}^s + \nu.$$

Since

$$e_{k-1}^s \le (2k-3)\nu \le c_s \nu,$$

the inductive hypothesis gives

$$\begin{aligned} e_k^{p,n} &\le (k-1)\left(1 + \frac{4b\Gamma}{\ell} c_s\right)\nu + \frac{4b\Gamma}{\ell} c_s \nu + \nu \\ &= k\left(1 + \frac{4b\Gamma}{\ell} c_s\right)\nu. \end{aligned}$$

This completes the induction.

Stability of the memorization gadget. Define

$$c_z := D, \qquad c_{p,n} := D\left(1 + \frac{4b\Gamma}{\ell} c_s\right), \qquad c_{\mathrm{mem}} := \max\{1, c_z, c_s, c_{p,n}\}.$$

If $\nu \le 1/c_{\mathrm{mem}}$, then

$$e_k^z, e_k^s, e_k^p, e_k^n \le 1.$$

Combining these bounds with the exact activation bounds gives, for every $k \in [D]$,

$$\sup_x \|x_k^{\theta_{\mathrm{mem}},\xi}(x)\|_\infty \le H_{\mathrm{mem}} + 1.$$

Define

$$l_{\mathrm{mem}} := 2c_{p,n} + \frac{4b\Gamma}{\ell} c_s + 1.$$

Since

$$l_{\mathrm{mem}} \ge c_z, \qquad l_{\mathrm{mem}} \ge c_s, \qquad l_{\mathrm{mem}} \ge c_{p,n},$$

the preceding bounds also give

$$\sup_x \max_{k\in[D]} \|x_k^{\theta_{\mathrm{mem}},\xi}(x) - x_k^{\theta_{\mathrm{mem}}}(x)\|_\infty \le l_{\mathrm{mem}}\nu.$$

It remains to bound the gadget output. Evaluating the preceding bounds at $k = D$, the P and N channels contribute at most $2c_{p,n}\nu$, while the final hinge group contributes at most $\frac{4b\Gamma}{\ell} c_s \nu$. Including the output disturbance, we obtain

$$\sup_x \|g_{\theta_{\mathrm{mem}}}^{\xi}(x) - g_{\theta_{\mathrm{mem}}}(x)\|_\infty \le \left(2c_{p,n} + \frac{4b\Gamma}{\ell} c_s + 1\right)\nu = l_{\mathrm{mem}}\nu.$$

Thus θ_{mem} has $(H_{\mathrm{mem}}, l_{\mathrm{mem}}, c_{\mathrm{mem}})$ disturbance stability.

Finally, applying Lemma 32 to the (H_0, l_0, c_0) disturbance stability of θ and the $(H_{\mathrm{mem}}, l_{\mathrm{mem}}, c_{\mathrm{mem}})$ disturbance stability of θ_{mem}, we obtain disturbance stability of the parallelized parameter $\theta^\dagger$ with

$$\begin{aligned} H &= \max\{H_0, H_{\mathrm{mem}}\} = \max\{H_0, 1, 2X + 2\ell, \Gamma m\}, \\ c &= \max\{c_0, c_{\mathrm{mem}}\} = \max\{1, c_0, c_z, c_s, c_{p,n}\}, \end{aligned}$$

and

$$l = l_0 + l_{\mathrm{mem}} = l_0 + 2c_{p,n} + \frac{4b\Gamma}{\ell} c_s + 1.$$

From Lemma 27, we finally get Structured Stability with

$$C = J_{w',H}\, c, \qquad L = J_{w',H}\, l.$$

Substituting in the definitions and bounding coarsely gives

$$H, C, L \lesssim C\, pw' dD^2 X m^2 \Gamma^2 (1 + \ell^{-1})(Mw)^{2d+1}.$$

□

D.5 BARRON FUNCTIONS

We use the ReLU Barron norm studied in E et al. (2022). A function $f : B_p \to \mathbb{R}$ has Barron norm

$$\|f\|_{\mathcal{B}} := \inf_{\rho} \mathbb{E}_{\rho}\left[|a|(\|b\|_1 + |c|)\right], \tag{10}$$

where the infimum is over probability measures $\rho(da, db, dc)$ such that

$$f(x) = \mathbb{E}_{\rho}[a\sigma(b^\top x + c)] \qquad \text{for every } x \in B_p.$$

Lemma 34 (Barron approximation)**.** *Let $Q := \|f\|_{\mathcal{B}}$. For every $r \geq 1$, there are coefficients $a_k \in \mathbb{R}$ and $\beta_k \in \mathbb{R}^{p+1}$ such that*

$$\sum_{k=1}^{r} |a_k| \leq Q, \qquad \|\beta_k\|_2 = 1 \quad \textit{for every } k \in [r],$$

and

$$f_r(x) := \sum_{k=1}^{r} a_k \sigma(\beta_k^\top (x, 1))$$

satisfies

$$\|f - f_r\|_{L^\infty(B_p)} \leq C_p Q r^{-\frac{1}{2} - \frac{3}{2p}},$$

where C_p depends only on p.

Proof. If $Q = 0$, then $f = 0$ and we take all $a_k = 0$. Hence assume $Q > 0$. In (10), assume for simplicity that the infimum is attained[5], so that there is a representing measure ρ satisfying

$$\mathbb{E}_{\rho}\left[|a|(\|b\|_1 + |c|)\right] = Q.$$

For $(b, c) \neq 0$, define

$$s(b, c) := \sqrt{\|b\|_2^2 + c^2}, \qquad v(b, c) := \frac{(b, c)}{s(b, c)} \in \mathbb{S}^p.$$

When $(b, c) = 0$, define $v(b, c)$ arbitrarily. By positive homogeneity of ReLU,

$$a\sigma(b^\top x + c) = a\, s(b, c)\, \sigma\left(v(b, c)^\top (x, 1)\right).$$

Define a finite signed measure μ on $\mathbb{S}^p$ as the pushforward under $(a, b, c) \mapsto v(b, c)$ of $as(b, c)d\rho(a, b, c)$:

$$\mu(A) := \int \mathbf{1}\{v(b, c) \in A\}\, a\, s(b, c)\, d\rho(a, b, c), \qquad A \subseteq \mathbb{S}^p.$$

Then

$$f(x) = \int a\sigma(b^\top x + c)\, d\rho(a, b, c) = \int_{\mathbb{S}^p} \sigma\left(v^\top (x, 1)\right)\, d\mu(v).$$

and

$$\|\mu\| \leq \int |a|\, s(b, c)\, d\rho(a, b, c) \leq \int |a|(\|b\|_1 + |c|)\, d\rho(a, b, c) \leq Q.$$

Thus $f \in \mathcal{F}_\sigma(Q)$, in Equation (1.2) of Yang & Zhou (2025). Equivalently, $f/Q \in \mathcal{F}_\sigma(1)$. Lemma 2.3 of Yang & Zhou (2025) now gives $g_r \in \mathcal{F}_\sigma(r, 1)$ such that

$$\left\| \frac{f}{Q} - g_r \right\|_{L^\infty(B_p)} \leq C_p r^{-\frac{1}{2} - \frac{3}{2p}}.$$

Setting $f_r := Qg_r$ gives $f_r \in \mathcal{F}_\sigma(r, Q)$ and

$$\|f - f_r\|_{L^\infty(B_p)} \leq C_p Q r^{-\frac{1}{2} - \frac{3}{2p}}.$$

By the definition of $\mathcal{F}_\sigma(r, Q)$, f_r has the displayed form with $\|\beta_k\|_2 = 1$ and $\sum_{k=1}^{r} |a_k| \leq Q$. □

[5] If it is not attained, the same conclusion follows by a limiting argument.

We now use depth to implement the function f_r at constant width. The technique below has precedent in e.g. Vardi et al. (2022), but the key fact is that introducing a copy channel for the input and building features from that channel prevents disturbances from accumulating exponentially.

Lemma 35 (Deep Barron witness)**.** *The function f_r in Lemma 34 is realized by a width-$(2p+3)$ and depth-r ReLU parameter*

$$\theta_r \in \mathcal{C}_{Q\vee 1}(2p+3, r).$$

Moreover, θ_r has $(H_{p,Q}, l_{p,Q}r, c_{p,Q}r)$ disturbance stability, where

$$H_{p,Q},\ l_{p,Q},\ c_{p,Q} \le C_p(1+Q)$$

for a constant C_p depending only on p.

Proof. We describe a deep network that uses additional copy channels for the input x, as well as channels for a running sum that simulates the wide computation.

Defining the network Write $\beta_k = (b_k, c_k) \in \mathbb{R}^p \times \mathbb{R}$, and decompose $a_k = a_k^+ - a_k^-$ into its positive and negative parts. By Lemma 34,

$$\sum_{k=1}^{r}(a_k^+ + a_k^-) = \sum_{k=1}^{r}|a_k| \le Q, \qquad \|\beta_k\|_2 = 1.$$

We notate the exact layer-k hidden activation as $x_k^{\theta_r} = (z_k, p_k, n_k, h_k)$, where $z_k = (Z_k^+, Z_k^-) \in \mathbb{R}^{2p}$ carries the input, $p_k, n_k \in \mathbb{R}$ contain the positive and negative parts of the running sum, and $h_k \in \mathbb{R}$ prepares the next summand. At the first layer, set

$$Z_1^+ = \sigma(x), \qquad Z_1^- = \sigma(-x), \qquad p_1 = n_1 = 0, \qquad h_1 = \sigma(b_1^\top x + c_1).$$

For $k = 2, \dots, r$, copy the Z^+ and Z^- channels and set

$$p_k = \sigma\left(p_{k-1} + a_{k-1}^+ h_{k-1}\right), \qquad n_k = \sigma\left(n_{k-1} + a_{k-1}^- h_{k-1}\right), \qquad h_k = \sigma\left(b_k^\top Z_{k-1}^+ - b_k^\top Z_{k-1}^- + c_k\right).$$

Since $Z_k^+ - Z_k^- = x$, we have $h_k = \sigma(\beta_k^\top(x, 1))$ at all $k \in [r]$. Because p_k and n_k are nonnegative by definition, ReLU acts as the identity on those channels, and we have

$$p_r = \sum_{k=1}^{r-1} a_k^+ h_k, \qquad n_r = \sum_{k=1}^{r-1} a_k^- h_k.$$

We define our readout as

$$p_r - n_r + a_r h_r, \tag{11}$$

which gives $f_r(x)$ by construction.

Parameter magnitude In the above construction, all parameters have magnitude at most $Q \vee 1$: $\|\beta_k\|_2 = 1$ implies that every entry of b_k and c_k has magnitude at most one, while

$$|a_k| \le Q \implies a_k^+, a_k^- \le Q.$$

Therefore $\theta_r \in \mathcal{C}_{Q\vee 1}(2p+3, r)$.

Stability We now verify that θ_r satisfies disturbance stability with parameters depending only on p and (polynomially) on Q. Here, we verify stability for inputs, x, in the larger set $[-1,1]^p \supset B_p$.

First, activation bounds. Clearly $\|z_k\|_\infty \le 1$ for all k. Since $\|\beta_k\|_2 = 1$, we have $|h_k| \le \|(x,1)\|_2 \le \sqrt{p+1}$. Using $\sum_k |a_k| \le Q$, the p_k and n_k channels are bounded by $Q\sqrt{p+1}$. Hence all exact hidden activations are bounded by some

$$H_{p,Q} \le C_p(1+Q),$$

where C_p depends on p (as $\sqrt{p}$). Let $e_k^z, e_k^p, e_k^n, e_k^h$ denote the maximum coordinatewise errors in the corresponding blocks under any ν-disturbed computation, in analogy to the computations in previous sections. Also let $e_k^{p,n} = e_k^p \vee e_k^n$.

The copied input channels clearly satisfy $e_k^z \le k\nu$.

Since $\|b_k\|_1 \le \sqrt{p}$, we have (using ReLU's Lipschitzness)

$$e_k^h \le 2\|b_k\|_1 e_{k-1}^z + \nu \le (2\sqrt{p}(k-1)+1)\nu \le C_p k\nu.$$

For the accumulation channels,

$$e_k^{p,n} \le e_{k-1}^{p,n} + |a_{k-1}| e_{k-1}^h + \nu.$$

Since $\sum_{k=1}^r |a_k| \le Q$, it follows that

$$e_k^{p,n} \le C_p(1+Q)r\nu.$$

Choose $c_{p,Q} = C_p(1+Q)$, independently of r, with C_p sufficiently large. Then

$$\nu \le (c_{p,Q} r)^{-1} \quad \Longrightarrow \quad \max_{k\in[r]} \max\{e_k^z, e_k^h, e_k^p, e_k^n\} \le 1.$$

In particular, every disturbed hidden activation is bounded by $H_{p,Q}+1$.

Finally, using (11) and including the output disturbance,

$$\begin{aligned}\sup_x \left| g_{\theta_r}^{2p+3,r,\xi}(x) - g_{\theta_r}^{2p+3,r}(x) \right| &\le e_r^p + e_r^n + |a_r| e_r^h + \nu \\ &\le C_p(1+Q)r\nu.\end{aligned}$$

Thus, with $l_{p,Q} = c_{p,Q} = C_p(1+Q)$, the disturbance-stability parameters are $(H_{p,Q}, l_{p,Q}r, c_{p,Q}r)$. □

E Proofs of main results

E.1 Proofs for Section 3.1

E.1.1 Proof of Theorem 3

Let $\theta \in \mathcal{C}_M(w,d)$ be the witness satisfying $\sup_{x\in\mathrm{supp}(\mu_X)} \|g_\theta^{w,d}(x) - f(x)\|_\infty \le \tau/2$. Let $\theta^\dagger \in \mathcal{C}_M(w,D)$ be its d-to-D identity padding. Padding preserves the output, so $\theta^\dagger$ $\tau/2$-interpolates every realized sample. By Lemma 31, this padded parameter has structured stability with

$$\begin{aligned}H_{\mathrm{pad}} &= A_d, \\ C_{\mathrm{pad}} &= 3D(pX+1+d)(Mw)^d, \\ L_{\mathrm{pad}} &= (3MwD+1)(pX+1+d)(Mw)^d.\end{aligned}$$

In particular,

$$H_{\mathrm{pad}}, L_{\mathrm{pad}}, C_{\mathrm{pad}} \le (Mw)^d \mathrm{poly}(X, M, p, d, D, w).$$

Apply Corollary 24 with $u = w$, $m_0 = 0$, $\eta = 0$, and the padded witness $\theta^\dagger$. For completeness, use the exact definition in Theorem 23:

$$q_{\mathrm{eff}} = \max\left\{2M \max\{1, C_{\mathrm{pad}}, 2L_{\mathrm{pad}}/\tau\}, J_{w,H_{\mathrm{pad}}}+1\right\}.$$

Write $a := Mw$, $b := pX+1$, and $v := \max\{1, \tau^{-1}\}$. Since $d \le D$ and $a \ge 1$,

$$H_{\mathrm{pad}} \le D(b+1)a^d, \qquad 1 \le C_{\mathrm{pad}} \le L_{\mathrm{pad}} \le 4aD^2(b+1)a^d.$$

Thus the first term in the maximum defining q_{eff} is at most $4ML_{\mathrm{pad}}v \le 16Ma^{d+1}D^2(b+1)v$. The second term obeys

$$\begin{aligned}J_{w,H_{\mathrm{pad}}} + 1 &= \max\{b, w(H_{\mathrm{pad}}+1)+1\} + 1 \\ &\le b + wH_{\mathrm{pad}} + w + 2 \\ &\le 4aD(b+1)a^d,\end{aligned}$$

which is also at most the same upper bound. Consequently,

$$q_{\mathrm{eff}} \le 16Ma^{d+1}D^2(b+1)v \le \left(16MwD(pX+2)\max\{1,\tau^{-1}\}\right)^{d+2}.$$

Taking logarithms proves the displayed bound in the theorem. Substituting this into the structured-interpolator rate gives the claimed dependence on the witness depth.

E.1.2 Generalization with stable witness (Strengthening of Theorem 3)

Theorem 3 relies on a crude exponential-in-d bound on the structured stability of its witness. Below, we give a generic version of the theorem, which follows immediately from Corollary 24 and allows us to bound q_{eff} in terms of the stability of the provided witness. As we saw in the Barron example, approximation theory can often provide witnesses that are deep, yet stable, which makes results like Corollary 24 and the following Theorem useful.

Below, recall the definition that

$$J_{u,H} := \max\{pX + 1,\ u(H+1)+1\}.$$

Theorem 36 (Disturbance-stable interpolator rate)**.** *Assume $y_i = f(x_i)$ for $i = 1, \dots, n$. Fix $u < W$ and D, M, H, l, c independently of the realized sample, with $M \geq 1$, $\tau > 0$, and $H, l, c > 0$. Suppose there is a training-data event $\mathcal{G}$ with probability at least $1 - \eta$ such that, on $\mathcal{G}$, there exists $\theta^\dagger \in \mathcal{C}_M(u, D)$ with (H, l, c) disturbance stability and*

$$\max_{i\in[n]} \|g_{\theta^\dagger}^{u,D}(x_i) - y_i\|_\infty \leq \frac{\tau}{2}.$$

Assume $0 \leq \eta < \delta < 1$.

Then

$$\Pr[\Delta_\tau(\hat{\theta}) \geq t_n] \leq \delta; \qquad t_n := C_0 \frac{uWD \log q_{\text{eff}} + \log \frac{1}{\delta - \eta}}{n}.$$

Here, C_0 is a constant depending only (linearly) on p and s and

$$q_{\text{eff}} \leq \max\left\{2M \max\left\{1, J_{u,H}c, \frac{2J_{u,H}l}{\tau}\right\}, J_{u,H} + 1\right\}.$$

Proof. Take $\theta^\dagger \in \mathcal{C}_M(u, D)$ so that $\max_i \|g_{\theta^\dagger}^{u,D}(x_i) - y_i\|_\infty \leq \tau/2$. By Lemma 27, $\theta^\dagger$ has (H, L, C) structured stability with $C = J_{u,H}c$ and $L = J_{u,H}l$. Corollary 24 therefore applies with $m_0 = 0$. Substituting these values of C and L into the definition of q_{eff} gives the statement. □

F Proof of Theorem 6

Theorem 37 (Theorem 6, formal)**.** *Let $\tau > 0$, $\delta \in (0, 1)$, $\Gamma > 0$, and $n \geq 2$. Consider a dataset of n points $\{(x_i, y_i)\}_{i=1}^n$, where $x_1 \dots x_n \sim \mu_X^{\otimes n}$ and μ_X satisfies Assumption 5. Moreover, assume labels $\{y_i\}$ are (m, Γ) adversarial corruptions from a ground truth f, as in Assumption 4.*

Let $M \geq \max\{1, X + 2\ell_n, \frac{\Gamma}{\ell_n}\}$, with $\ell_n = \frac{\delta}{6\Lambda n(n-1)}$. Assume that, for some w and d, there exists $\theta \in \mathcal{C}_M(w, d)$ so that $\|f(x) - g_\theta^{w,d}(x)\|_\infty \leq \frac{\tau}{2}$ for all $x \in [-X, X]^p$. Let $D \geq d$ and $W > w + 2 + 2s + 4\lceil \frac{m}{D} \rceil =: w'$.

Then there is a constant C_0 depending only on p and s, so that for

$$\Pr_{x^n \sim \mu^{\otimes n}, \hat{\theta} \sim U_{\tau,n}^M} \left[\Delta_\tau(\hat{\theta}) \geq t_n\right] \leq \delta; \qquad t_n := C_0 \frac{(wWD + Wm) \log q_{\text{eff},n} + \log \frac{1}{\delta}}{n}.$$

Here, $q_{\text{eff},n}$ is a value satisfying

$$\log q_{\text{eff},n} \leq 2d \log\left(\text{poly}(\Gamma, \Lambda, X, n, m, w, p, s, \ell_n^{-1}) M \max\{1, \tau^{-1}\} D\right).$$

Proof. Set $\eta = \delta/2$. Take the assumed $\theta \in \mathcal{C}_M(w, d)$ and pad it from depth d to depth D with identity layers. Now apply Lemma 15, which tells us that on an event $\mathcal{G}$ (over the training data) with probability at least $1 - \eta$, for any I with $|I| = m$ and $\{\delta_i\}_{i=1}^m \subset [-\Gamma, \Gamma]^s$ we have a parameter $\theta^\dagger \in \mathcal{C}_M(w', D)$ so that $g_\theta^{w',D}(x_i) = g_\theta^{w,D}(x_i)$ for $i \notin I$, and $g_\theta^{w',D}(x_i) = g_\theta^{w,D}(x_i) + \delta_i$ for $i \in I$. Here,

$$w' = w + 2 + 2s + 4\left\lceil \frac{m}{D} \right\rceil.$$

By Lemma 33, any such $\theta^\dagger$ has structured stability with

$$H, C, L \lesssim pw'dD^2Xm^2\Gamma^2(1+\ell^{-1})(Mw)^{2d+1}.$$

Above, we clearly see that $m \le w'D$, so Corollary 25 applies with $u = w'$, $m_0 = m$, and $\eta = \delta/2$. Since

$$w'WD \le C_1(wWD + Wm)$$

for a constant C_1 depending only on p and s, we have

$$\Pr\left[\Delta_\tau(\hat{\theta}) \ge \frac{C_0(wWD + Wm)\log \bar{q}_{\mathrm{eff,n}} + \log\frac{2}{\delta}}{n}\right] \le \delta.$$

Here, $\log \bar{q}_{\mathrm{eff},n} = \max\{\log q_{\mathrm{eff}}, 1 + \log n\}$; the theorem uses this enlarged scale as $q_{\mathrm{eff},n}$, and so $\log q_{\mathrm{eff},n} \le 2d \log\left(\mathrm{poly}(\Gamma, \Lambda, X, n, m, w, p, s, \ell_n^{-1})M\max\{1, \tau^{-1}\}D\right)$ due to Lemma 33. □

G Proofs of Corollary 7, and Theorem 8

G.1 Proof of Theorem 7

In this section, we state formally and prove Theorem 8. In the following assumption, we formally introduce the corruption setting for Theorem 8.

Assumption 38 (Constant rate adversarial corruptions)**.** For an $\epsilon \in (0,1)$, assume $m \sim \mathrm{Binom}(n, \epsilon)$ independently of the realization of the data. Then we assume that the data undergoes an (m, Γ) adversarial corruption.

The above setting includes the following special case.

Assumption 39 (ϵ-corrupted data model)**.** Under the ϵ-corruption model, we assume that each label $y_i \in \mathbb{R}^s$ is corrupted by a bounded corruption $\delta_i \in [-\Gamma, \Gamma]^s$ with (independent) probability ϵ. Concretely, for $\delta_1 \dots \delta_n$ possibly depending on $x_1 \dots x_n$,

$$y_i = \begin{cases} f(x_i) & \text{with probability } 1-\epsilon \\ f(x_i) + \delta_i & \text{with probability } \epsilon \end{cases}$$

We may assume without loss of generality that the corrupted points are $y_1 \dots y_m$ for $m \sim \mathrm{Binom}(n, \epsilon)$, because the algorithm we are studying is invariant under the reordering of data points.

Also need Λ-marginal boundedness below. Log factors include many additional problem params.

Theorem 40 (Corollary 7, formal)**.** *Let $\epsilon, \delta \in (0,1)$ and $\tau > 0$. Assume μ_X satisfies Assumption 5, and that we observe labels according to Assumption 38 for some ground truth f. Assume that there exists $\theta \in \mathcal{C}_M(w, d)$ so that $f = g_\theta^{w,d}$ on $[-X, X]^p$.*

Let $W > w + 2 + 2s + 4B$ for some $B \in \mathbb{N}_{\ge 1}$. Let $m_n = \min\{n, \lceil 2\epsilon n + \log\frac{2}{\delta}\rceil\}$ and consider

$$D_n := \max\left\{d, 2, \left\lceil \frac{m_n}{B} \right\rceil\right\}.$$

Assume $M \ge \max\{1, X + 2\ell_n, \frac{\Gamma}{\ell_n}\}$, with $\ell_n = \frac{\delta}{12\Lambda n(n-1)}$.

Consider a parameter $\hat{\theta}$ sampled from the set of τ-interpolating solutions of width W and depth D_n:

$$\hat{\theta} \sim U_{\tau,n}^{W, D_n, M}.$$

Then for $n = \max\left\{2, \left\lceil \frac{\log(2/\delta)}{\epsilon} \vee \frac{Bd - \log(2/\delta)}{\epsilon} \right\rceil\right\}$, we have

$$\Pr_{x^n, \hat{\theta}, m}\left[\Delta_\tau(\hat{\theta}) > \tilde{O}\left(W\left(1 + \frac{w}{B}\right)\epsilon\right)\right] \le \delta$$

where $\tilde{O}(\cdot)$ hides only problem-dependent constants, and log factors of $\frac{1}{\epsilon}$ and $\log\frac{2}{\delta}$, and B.

Proof. Apply Lemma 14 with $\eta = \delta/2$. On the resulting event, the sample has at most m_n corruptions. Apply Theorem 37 with confidence parameter $\delta/2$, corruption budget m_n, and D_n. Since

$$W > w + 2 + 2s + 4B \geq w + 2 + 2s + 4\left\lceil \frac{m_n}{D_n} \right\rceil,$$

the width condition holds because B is an integer. Write $z = \log(2/\delta)$. The stated choice of n gives $\epsilon n \geq z$ and $\epsilon n \geq Bd - z$, hence $\epsilon n \geq Bd/2$. Consequently,

$$m_n \leq 2\epsilon n + z + 1 \leq 5\epsilon n, \qquad D_n = \max\{d, 2, \lceil m_n/B \rceil\} \leq 7\epsilon n/B.$$

Here $1 \leq 2\epsilon n/B$ and $d \leq 2\epsilon n/B$ control the ceiling and the minimum depth. Absorbing these constants into C_0,

$$t_n \leq C_0 W \left(1 + \frac{w}{B}\right) \epsilon \log q_{\text{eff},n} + C_0 W \left(1 + \frac{w}{B}\right) \frac{\log(2/\delta) \log q_{\text{eff},n}}{n} + \frac{\log(2/\delta)}{n}.$$

for the value of $q_{\text{eff},n}$ of Theorem 37.

At the stated value of n, this upper bound is

$$t_n \leq \tilde{O}\left(W \left(1 + \frac{w}{B}\right) \epsilon\right).$$

Union bounding the binomial event and the noisy theorem gives that the above holds with probability at least $1 - \delta$. □

G.2 Proof of Theorem 8

Theorem 41 (Theorem 8 formal)**.** *Let $\delta \in (0, 1)$ and $\tau > 0$. Suppose $x_1 \dots x_n \sim \mu_X$ i.i.d. and $y_i = f(x_i)$ for all $i \in [n]$. Assume μ_X satisfies Assumption 5.*

Consider sequences $\epsilon_n \in (0, 1]$, w_n, d_n, D_n, A_n with the requirements that $\epsilon_n = \Omega(\frac{1}{n})$ and $D_n \geq d_n$. Suppose that, for all $n \in \mathbb{N}_{\geq 2}$, there is $\theta_{0,n} \in \mathcal{C}_{A_n}(w_n, d_n)$ that has disturbance stability $(H_{0,n}, l_{0,n}, c_{0,n})$ and satisfies

$$\mu_X\left(\left\{x : \|g^{w_n,d_n}_{\theta_{0,n}}(x) - f(x)\|_\infty > \frac{\tau}{2}\right\}\right) \leq \epsilon_n, \qquad \sup_{x \in [-X,X]^p} \|g^{w_n,d_n}_{\theta_{0,n}}(x) - f(x)\|_\infty \leq \Gamma.$$

Let $m_n := \min\{n, \lceil 2\epsilon_n n + \log(4/\delta) \rceil\}$. Let $W_n > w_n + 2 + 2s + 4\lceil \frac{m_n}{D_n} \rceil$, and $M_n \geq \max\{1, A_n, X + 2\ell_n, \frac{\Gamma}{\ell_n}\}$, where $\ell_n = \frac{\delta}{12\Lambda n(n-1)}$.

Consider $\hat{\theta} \sim U^{W_n,D_n,M_n}_{\tau,n}$. Then

$$\Pr_{x^n,\hat{\theta}}\left[\Delta_\tau(\hat{\theta}) \geq \tilde{O}\left(W_n \left(\frac{w_n D_n + \log\frac{1}{\delta}}{n} + \epsilon_n\right) \log q_{\text{eff,n}}\right)\right] \leq \delta$$

where $q_{\text{eff}} = \text{poly}(n, H_{0,n}, c_{0,n}, l_{0,n}, D_n, w_n, d_n, M_n, X, \Lambda, \Gamma, \delta^{-1}, \tau^{-1})$.

Proof. Define the exceptional set

$$E_n := \left\{x \in [-X, X]^p : \|g^{w_n,d_n}_{\theta_{0,n}}(x) - f(x)\|_\infty > \frac{\tau}{2}\right\}.$$

By assumption, $\mu_X(E_n) \leq \epsilon_n$. Define $I_{E_n} = \{i \in [n] : x_i \in E_n\}$ to be the (data-dependent) set of indices corresponding to exceptional points. Lemma 14, applied with $\epsilon = \mu_X(E_n)$ and $\eta = \delta/4$, gives

$$|I_{E_n}| \leq m_n$$

with probability at least $1 - \delta/4$. On this event, for each $i \in I_{E_n}$ define

$$\delta_i := f(x_i) - g^{w_n,d_n}_{\theta_{0,n}}(x_i).$$

The uniform approximation bound gives $\|\delta_i\|_\infty \le \Gamma$.

Pad $\theta_{0,n}$ from depth d_n to depth D_n to get θ_n, which has disturbance stability with $H_{\mathrm{pad},n}, l_{\mathrm{pad},n}, c_{\mathrm{pad},n}$ by Lemma 30) applied to $\theta_{0,n}$ and $(H_{0,n}, l_{0,n}, c_{0,n})$. Given the lower bound on M_n, we may apply Lemma 15 with failure probability $\delta/4$, budget m_n, subset I_{E_n}, and base parameter θ_n. In particular, memorization gadget used uses $\ell = \frac{\delta}{12\Lambda n(n-1)}$.

By a union bound, on an event $\mathcal{G}_n$ with probability at least $1 - \delta/2$, there exists

$$\theta_n^\dagger \in \mathcal{C}_{M_n}(u_n, D_n), \qquad u_n := w_n + 2 + 2s + 4\left\lceil \frac{m_n}{D_n} \right\rceil,$$

such that

$$\max_{i\in[n]} \|g_{\theta_n^\dagger}^{u_n, D_n}(x_i) - f(x_i)\|_\infty \le \frac{\tau}{2}.$$

(H_n, l_n, c_n) disturbance stability stability for $\theta_n^\dagger$ follows from applying Lemma 33 with $\theta = \theta_n$ and its associated disturbance-stable parameters $(H_{\mathrm{pad},n}, l_{\mathrm{pad},n}, c_{\mathrm{pad},n})$. Substituting $\ell = \frac{\delta}{12\Lambda n(n-1)}$ into its expression tells us that $\theta_n^\dagger$ has structured stability with parameters

$$H_n, C_n, L_n \lesssim \mathrm{poly}(H_{0,n}, l_{0,n}, c_{0,n}, p, s, u_n, d_n, D, M, \Gamma, \Lambda, X, n, \delta^{-1}). \tag{12}$$

Since $W_n > u_n$, we may now apply Corollary 24 with $u = u_n$, $D = D_n$, $m_0 = 0$, and $\eta = \delta/2$. The numerator uses all n observations because every observed label equals $f(x_i)$. We obtain

$$\Pr_{x^n\sim\mu_X^{\otimes n},\, \hat\theta\sim U_{\tau,n}^{W_n, D_n, M_n}} \left[\Delta_\tau(\hat\theta) \ge t_n\right] \le \delta, \qquad t_n = C_0 \frac{u_n W_n D_n \log q_{\mathrm{eff},n} + \log\frac{2}{\delta}}{n}.$$

where Theorem 8's bound on $q_{\mathrm{eff,n}}$ follows by combining (12) with (8).

We now further bound t_n. Notice that $D_n\left\lceil \frac{m_n}{D_n}\right\rceil \le m_n + D_n \le m_n + w_n D_n$ ($w_n \ge 1$), so that, for some constant C_1 depending only on s:

$$u_n D_n \le C_1\big(w_n D_n + m_n\big).$$

Therefore, we have a p and s-dependent constant C_0 such that

$$t_n \le C_0 W_n \left(\frac{w_n D_n + m_n}{n}\right)\log q_{\mathrm{eff},n} + \frac{\log\frac{2}{\delta}}{n}. \tag{13}$$

Consider the term $\frac{m_n}{n} \le \frac{\lceil 2\epsilon_n n\rceil}{n} + \frac{\lceil \log\frac{4}{\delta}\rceil}{n}$. We have

$$\frac{\lceil 2\epsilon_n n\rceil}{n} = O\left(\max\left\{\epsilon_n, \frac{1}{n}\right\}\right) = O(\epsilon_n)$$

since $\epsilon_n = \Omega(n^{-1})$. Thus, plugging back into (13)

$$\frac{m_n}{n} = O\left(\epsilon_n + \frac{\log\frac{1}{\delta}}{n}\right).$$

This allows us to complete the proof by noting

$$t_n = O\left(W_n\left(\frac{w_n D_n + \log\frac{1}{\delta}}{n} + \epsilon_n\right)\log q_{\mathrm{eff},n}\right).$$

□

H CLASSIFICATION

In the classification setting, we assume we observe labeled data $\{(x_i, \bar y_i)\}_{i=1}^n$. As before, we assume $x_i \in [-X, X]^p$ are random features distributed according to μ. However, we now assume that

$\bar{y}_i \in [s]$ is a discrete observation indicating membership in one of s classes. As such, we now interpret the s outputs of the neural network as scores over s classes, with class predictions then given by returning the class with the highest score. We notate the corresponding decision rule as

$$c_\theta^{w,d}(x) := \arg\max_{j \in [s]} g_\theta^{w,d}(x)[j]$$

with ties broken arbitrarily. In the classification setting, a parameter $\theta \in \mathcal{C}_M(w,d)$ **interpolates** if $c_\theta^{w,d}$ correctly classifies each data point. Given data $\{(x_i, \bar{y}_i)\}$, a weight-bound M, and a target width and depth W, D, we define

$$S_n^{W,D,M} := \left\{\theta \in \mathcal{C}_M(W,D) : c_\theta^{W,D}(x_i) = \bar{y}_i \text{ for all } i \in [n]\right\}, \quad U_n^{W,D,M} := \mathrm{Unif}(S_n^{W,D,M}),$$

where the technical details behind the definition of $U_n^{W,D,M}$ are as before (refer to Appendix B).

In this section, we focus on the case where $\bar{y}_i$ reflect partially corrupted labels from a ground truth labeling function, $c : \mathcal{X} \to [s]$. We evaluate any estimate $\theta \in \mathcal{C}_M(W,D)$ in terms of its classification error with respect to c:

$$\Delta^{\mathrm{cls}}(\theta; c) = \Pr_x\left[c_\theta^{W,D}(x) \neq c(x)\right].$$

Parallel to the regression setting, we give results for a noisy setting where m data points can be corrupted arbitrarily. These can be converted into tempered overfitting-type results exactly as in the regression setting.

Assumption 42 (Adversarial noise, classification)**.** The labels $\{(x_i, \bar{y}_i)\}_{i=1}^n$ are generated as follows: $\{x_i\}_{i=1}^n \overset{iid}{\sim} \mu_X$, and $y_i := c(x_i)$. Then an adversary can change the class label for any m points, resulting in the dataset $\{(x_i, \bar{y}_i)\}_{i=1}^n$.

We also introduce a margin assumption on the witness: if we wish to measure error with respect to a classifier, c, then some $\theta \in \mathcal{C}_M(w,d)$ must induce $c_\theta^{w,d} = c$ with margin on the logits.

Assumption 43 (Realizability with a small witness at margin γ)**.** Assume there exists $\theta \in \mathcal{C}_M(w,d)$ so that, letting $g := g_\theta^{w,d}$, we have

$$g(x)[c(x)] > \max_{j \neq c(x)} g(x)[j] + \gamma$$

for every $x \in \mathrm{supp}(\mu)$. Furthermore, assume $\|g(x)\|_\infty \leq B/2$ for all $x \in \mathrm{supp}(\mu)$.

As an example, it suffices for the Bayes probability to have margin for (almost) all x, and be realizable by $\theta \in \mathcal{C}_M(w,d)$. In binary classification, such a requirement is referred to as a *Massart condition*.

We now state the analog of Theorem 6 for the classification setting. First, for simplicity we state an informal version that can be compared to the informal statement of Theorem 6.

Theorem 44 (Noisy classification setting, informal)**.** *Let $\delta \in (0,1)$ and $n \geq 2$. Assume data is gathered according to Assumption 42's adversarial corruption model. Also suppose μ_X satisfies marginal boundedness (Assumption 5) with parameter Λ.*

Assume that Assumption 43 holds for $\theta \in \mathcal{C}_M(w,d)$ with margin $\gamma > 0$ and bound B. Let $M \geq \mathrm{poly}(n, \Lambda, B, X, \delta^{-1})$, where poly *is a fixed polynomial stated in the formal version.*

Let $D \geq d$ and $W > w + w_+$, where $w_+ \leq C\left(1 + \frac{m}{D}\right)$ for some constant C depending on s.

Then, for $\hat{\theta} \sim U_n^{W,D,M}$, we have

$$\Pr_{x^n, \hat{\theta}}\left[\Delta^{\mathrm{cls}}(\hat{\theta}) \geq C_0 \frac{W(wD+m)\log q_{\mathrm{eff},n}}{n}\right] \leq \delta.$$

Here, $q_{\mathrm{eff},n} = \tilde{O}(d)$.

Lemma 45 (Classification numerator bound)**.** *Let $t \in [0,1]$. Let*

$$B_{t,n} := S_n^{W,D,M} \cap \{\phi \in \mathcal{C}_M(W,D) : \Delta^{\mathrm{cls}}(\phi) \geq t\}.$$

Under the noisy classification setting with at most m corrupted labels,

$$\mathbb{E}_{x^n}[\lambda(B_{t,n})] \leq (2M)^{N_{W,D}} \binom{n}{m} e^{-t(n-m)}.$$

Proof. The proof is identical to Lemma 22, replacing τ-accurate prediction by correct classification. □

For the formal version of the theorem, we write the rate directly in terms of an arbitrary stability certificate for θ, which we know to be at worst exponential-in-d, but which may be better.

Theorem 46 (Theorem 44, formal)**.** *Let $\gamma > 0$, $\delta \in (0,1)$. Let $n \geq 2$ be the number of data points and m be the maximum number of corrupted data points. Assume the noisy classification setting of Assumption 42. Also suppose Assumption 5 holds with parameter Λ. Finally assume that Assumption 43 holds with margin γ and bound B, and further assume that the witness θ has disturbance stability with parameters (H_0, l_0, c_0).*

Let $M \geq \max\{1, X + 2\ell_n, B/\ell_n\}$, with $\ell_n = \frac{\delta}{6\Lambda n(n-1)}$. Let $D \geq d$ and $W > w + 2 + 2s + 4\left\lceil \frac{m}{D} \right\rceil$.

Let $\hat{\theta} \sim U_n^{W,D,M}$. Then there is a constant C_0 depending only on p and s, so that

$$\Pr_{x^n, \hat{\theta}}\left[\Delta^{\mathrm{cls}}(\hat{\theta}) \geq t_n\right] \leq \delta, \qquad t_n := C_0 \frac{(wWD + Wm)\log q_{\mathrm{eff},n} + \log\frac{1}{\delta}}{n},$$

Here, $q_{\mathrm{eff},n}$ is a value satisfying

$$\log q_{\mathrm{eff},n} \leq \log\left(\mathrm{poly}(H_0, l_0, c_0, p, s, w, d, D, M, B, \Lambda, X, n, \delta^{-1})\max\{1, \gamma^{-1}\}\right).$$

Proof. Let $I \subseteq [n]$ denote the set of corrupted points, with $|I| \leq m$. For each class $a \in [s]$, define $v^a \in \mathbb{R}^s$ by

$$v^a[a] = \frac{B}{2}, \qquad v^a[j] = -\frac{B}{2} \quad (j \neq a).$$

For each corrupted point, set

$$\delta_i^{\mathrm{cls}} := v^{\bar{y}_i} - g_\theta^{w,d}(x_i).$$

Then $\|\delta_i^{\mathrm{cls}}\|_\infty \leq B$, so the memorization gadget applies with $\Gamma = B$. Let

$$w' := w + 2 + 2s + 4\left\lceil \frac{m}{D} \right\rceil.$$

Let $\mathcal{G}$ be the event that the $(n, m, \Lambda, B, \delta/2)$-memorization gadget succeeds. Then $\Pr(\mathcal{G}) \geq 1 - \delta/2$, and on $\mathcal{G}$ padding the witness and appending the gadget gives $\theta^\dagger \in \mathcal{C}_M(w', D)$ such that

$$g_{\theta^\dagger}^{w',D}(x_i) = v^{\bar{y}_i} \quad (i \in I), \qquad g_{\theta^\dagger}^{w',D}(x_i) = g_\theta^{w,d}(x_i) \quad (i \notin I).$$

Thus $\theta^\dagger$ classifies every observed label with margin at least B on the corrupted points and at least γ on the uncorrupted points.

Give the choice of ℓ_n from the Theorem statement, Lemma 30 together with Lemma 33 tells us that $\theta^\dagger$ has structured stability with

$$H, C, L \lesssim \mathrm{poly}(H_0, l_0, c_0, p, s, w, d, D, M, B, \Lambda, X, n, \delta^{-1}). \tag{14}$$

Define

$$\rho_{\mathrm{cls}} := \min\left\{1, \frac{1}{C}, \frac{\gamma}{3L}\right\}, \qquad \rho'_{\mathrm{cls}} := \frac{M}{J_{w',H} + 1}, \tag{15}$$

and

$$\widetilde{q}_{\mathrm{eff},n} := \max\left\{\frac{2M}{\rho_{\mathrm{cls}}}, \frac{M}{\rho'_{\mathrm{cls}}}\right\}. \tag{16}$$

By Lemma 19, every $(\rho_{\text{cls}}, \rho'_{\text{cls}})$-lift lies within $\gamma/3$ of $\theta^\dagger$. Since $B \geq \gamma$, every such lift classifies the sample correctly. Therefore, on $\mathcal{G}$,

$$\lambda(S_n^{W,D,M}) \geq (2M)^{F_{w',W,D}} (2\rho'_{\text{cls}})^{K_{w',W,D}} \rho_{\text{cls}}^{N_{w',D}}.$$

Splitting off $\mathcal{G}^c$ first, as in Theorem 23, and combining this positive denominator bound on $\mathcal{G}$ with Lemma 45 gives

$$\Pr[\Delta^{\text{cls}}(\hat{\theta}) \geq t] \leq \frac{\delta}{2} + \binom{n}{m} (\widetilde{q}_{\text{eff},n})^{N_{w',D}+K_{w',W,D}} e^{-t(n-m)}. \tag{17}$$

Set

$$q_{\text{eff},n} := \max\{en, \widetilde{q}_{\text{eff},n}\}.$$

$q_{\text{eff},n}$ satisfies the bound from the Theorem statement, as we may see by combining Equations (14), (15) and (16). By the same calculations as in Corollary 25, Equation (17) can be manipulated into the given bound (using $w'WD \lesssim (wWD + Wm)$). □

To recover the informal version of the theorem statement, we can simply use the concrete $(Mw)^d$-type stability bound on the witness $\theta \in \mathcal{C}_M(w, d)$.

H.1 Boltzmann distribution

A more realistic model for the distribution of learned parameter vectors $\hat{\theta} \in \mathcal{C}(W, D)$ for fixed training data $x_1 \ldots x_n$ is a loss-based Boltzmann distribution with temperature $\beta > 0$ (Mingard et al., 2021):

$$b_n^{W,D,M}(\theta) \propto e^{-\beta \hat{L}_n(\theta)} \mathbf{1}\{\theta \in \mathcal{C}_M(W, D)\}, \tag{18}$$

where $\hat{L}_n = \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, \theta)$ is the empirical loss corresponding to a *loss-function* ℓ.

Here, we focus on classification, where a common choice of ℓ is *cross-entropy* (CE)-loss. To define the loss function, we first consider the logits induced by any $\theta \in \mathcal{C}(W, D)$:

$$\text{logit}_\theta(x)[i] = \frac{\exp(g_\theta^{W,D}(x)[i])}{\sum_{j=1}^s \exp(g_\theta^{W,D}(x)[j])}.$$

Then, when $\bar{y}_i \in [s]$ indicates membership in one of s classes, ℓ_{CE} is defined as

$$\ell_{CE}(x_i, \bar{y}_i, \theta) := -\log\left(\text{logit}_\theta(x_i)[\bar{y}_i]\right) = -g_\theta^{W,D}(x_i)[\bar{y}_i] + \log \sum_{j=1}^s \exp(g_\theta^{W,D}(x_i)[j]).$$

Now take

$$\hat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_{CE}(x_i, y_i, \theta),$$

and consider the definition of $b^{W,D,M}$ in (18). We may *condition* $b^{W,D,M}$ on interpolation. Specifically, we take

$$\pi_n^{W,D,M}(\theta) \propto b_n^{W,D,M}(\theta)\mathbf{1}\{\theta \in S_n^{W,D,M}\} \propto e^{-\beta \hat{L}_n(\theta)}\mathbf{1}\{\theta \in S_n^{W,D,M}\}. \tag{19}$$

This density is used when $S_n^{W,D,M}$ has positive ambient volume; otherwise, return the zero parameter vector, as for $U_n^{W,D,M}$.

The next result shows that $\hat{\theta} \sim \pi_n^{W,D,M}(\theta)$ generalizes under the same assumptions as Theorem 46.

Theorem 47. *Adopt the setting of Theorem 46, but suppose $\hat{\theta} \sim \pi_n^{W,D,M}$ rather than $\hat{\theta} \sim U_n^{W,D,M}$. Then there is a constant C_0 depending on p and s so that*

$$\Pr_{x^n,\hat{\theta}}\left[\Delta^{\mathrm{cls}}(\hat{\theta}) \geq t_n\right] \leq s^\beta \delta, \qquad t_n := C_0 \frac{(wWD + Wm)\log q_{\mathrm{eff},n} + \log\frac{1}{\delta}}{n},$$

Here, $q_{\mathrm{eff},n}$ is a value satisfying

$$\log q_{\mathrm{eff},n} \leq \log\left(\mathrm{poly}(H_0, l_0, c_0, p, s, w, d, D, M, B, \Lambda, X, n, \delta^{-1})\max\{1, \gamma^{-1}\}\right).$$

Proof.

$$\Pr_{\hat{\theta}\sim\pi_n^{W,D,M},x_n}\left[\Delta^{\mathrm{cls}}(\hat{\theta}) \geq t\right] \leq \mathbb{E}_{x^n}\mathbb{E}_{\hat{\theta}\sim U_n^{W,D,M}}\left[\underbrace{\frac{e^{-\beta\hat{L}_n(\hat{\theta})}}{\mathbb{E}_{\theta\sim U_n^{W,D,M}}\left[e^{-\beta\hat{L}_n(\theta)}\right]}}_{r(\hat{\theta})}\mathbf{1}\{\Delta^{\mathrm{cls}}(\hat{\theta}) \geq t\}\right] \tag{20}$$

On zero-volume datasets the two laws coincide and $r = 1$ at their common output. Otherwise, let $\theta \in S_n^{W,D,M}$ be an interpolating solution, in the sense that $c_\theta^{W,D}(x_i) = \bar{y}_i$ for all $i \in [n]$. Notice that this implies $\mathrm{logit}_\theta(x_i)[\bar{y}_i] \geq \frac{1}{s}$. Therefore $\ell_{CE}(x_i, y_i, \theta) \leq \log s$ for each $i \in [n]$, and hence $\hat{L}_n(\theta) \leq \log s$. Moreover, we definitionally have $\hat{L}_n(\theta) \geq 0$ for any θ. Hence

$$r(\hat{\theta}) \leq e^{\beta \log s} = s^\beta.$$

Since the indicator in Equation (20) is positive, we have the upper bound

$$\Pr_{\hat{\theta}\sim\pi_n^{W,D,M},x_n}\left[\Delta^{\mathrm{cls}}(\hat{\theta}) \geq t\right] \leq s^\beta \Pr_{x^n,\hat{\theta}\sim U_n^{W,D,M}}\left[\Delta^{\mathrm{cls}}(\hat{\theta}) \geq t\right].$$

But the quantity on the right hand side is now bounded by Theorem 46. □

Remark 48. Setting $\delta' = \delta/s^\beta$ and appealing to Theorem 47 says that, with probability δ, we achieve risk $t_n = \tilde{O}\left(\frac{\beta(wWD+Wm)}{n}\right)$, since $1/\delta$ only appears in log terms.

Why is the Boltzmann model important? To heuristically see why the Boltzmann distribution is a natural model for gradient-based optimizers, first ignore the interpolation constraint. In the interior of $\mathcal{C}_M(W, D)$, Langevin dynamics targeting the ambient law $b_n^{W,D,M}$ takes the form

$$\theta_{k+1} = \theta_k - \eta\beta\nabla_\theta L_n(\theta_k) + \sqrt{2\eta}\,\xi_k, \qquad \xi_k \sim N(0, I),$$

which is a noised version of gradient descent. $\pi_n^{W,D,M}$ essentially considers conditioning on the set $S_n^{W,D,M}$, which (still heuristically) can be viewed as assuming the above process finds an interpolating solution (i.e. repeating the process, allowing it to mix each time). More rigorously, the connection between Boltzmann/Bayesian sampling and gradient-methods has a long history in the literature (Welling & Teh, 2011; Mingard et al., 2021; Chen et al., 2022).

I Full examples

I.1 Examples for Section 3.1

Example: Hölder functions Suppose $\mathrm{supp}(\mu_X) \subseteq [0,1]^p$ and let $f : [0,1]^p \to \mathbb{R}$ satisfy Hölder smoothness with index $\beta > 0$ and radius $K > 0$, meaning it has bounded partial derivatives up to order $\lfloor\beta\rfloor$, and its $\lfloor\beta\rfloor$-order partial derivative are $\beta - \lfloor\beta\rfloor$ Hölder continuous (Schmidt-Hieber, 2020). Then it is known from Ohn & Kim (2019) that for some L_0, N_0, M_0 depending only on p, β and K, there exists a ReLU network $\theta \in \mathcal{C}_{M_0}(w, d)$ with width $w = N_0\tau^{-p/\beta}$, depth $d = L_0 \log\frac{1}{\tau}$ such that $|g_\theta^{w,d} - f(x)| \leq \tau$ for $x \in [0,1]^p$. Thus, we have the following high probability guarantee whenever $W > w$ and $D \geq d$:

$$\Delta_\tau(\hat{\theta}) \leq \tilde{O}_{p,\beta,K}\left(\frac{WD\tau^{-p/\beta}}{n}\right).$$

If we choose $W \asymp w$ and $D \asymp d$, we get $\Delta_\tau(\hat{\theta}) \leq \tilde{O}_{p,\beta,K}\left(\frac{\tau^{-2p/\beta}}{n}\right)$. If f is bounded by F, following Remark 2 and tuning τ leads to the high probability L_1 bound:

$$\left\|\hat{f} - f\right\|_{L^1(\mu_X)} = \tilde{O}\left(n^{-\frac{\beta}{\beta+2p}}\right)$$

For reference, the minimax rate is $O(n^{-\beta/p})$, usually achieved by specialized algorithms that are aware of β (unlike our agnostic sampling rule), and requiring additional regularity assumptions on μ_X not made by us (Kohler & Krzyżak, 2013). Without these regularity assumptions, there is an $\Omega\left(\frac{1}{n}\right)$ lower bound even for infinitely smooth f (Kohler & Krzyżak, 2013); the exponent in our rate approaches this barrier as β grows with p fixed.

Example: Barron functions Now assume $\mathrm{supp}(\mu_X) \subseteq B_p$, where B_p is the Euclidean unit ball in p dimensions, so that we may take $X = 1$. ReLU-Barron functions $f : B_p \to \mathbb{R}$ are functions realizable as "infinite-width" ReLU neural networks:

$$f(x) = \mathbb{E}_{a,b,c\sim\rho}\left[a\sigma(b^T x + c)\right]. \tag{21}$$

The ReLU-Barron norm $\|f\|_{\mathcal{B}}$ is defined as $\inf_\rho \mathbb{E}\left[|a|(\|b\|_1 + |c|)\right]$, where the infimum is taken over measures ρ that realize (21). Let $Q := \|f\|_{\mathcal{B}}$. In Appendix D.5, we use the results of Yang & Zhou (2025) to show that for every $r \geq 1$ there is a function f_r satisfying

$$\|f - f_r\|_{L^\infty(B_p)} \leq C_p Q r^{-\frac{p+3}{2p}},$$

and that f_r can be represented exactly by a ReLU network $\theta_r \in \mathcal{C}_{2Q\vee 1}(2p+3, r)$ with $(H_{p,Q}, L_{p,Q}r, C_{p,Q}r)$-structured stability (to be defined in Section 4). Crucially, the stability is linear in r rather than exponential.

Thus, taking $d_\tau \asymp (Q/\tau)^{\frac{2p}{p+3}}$ gives a width-$(2p+3)$, depth-d_τ structured-stable parameter whose realization lies within $\tau/2$ of f uniformly on B_p. Taking $D \asymp d_\tau$, Theorem 36 — a strengthening of Theorem 3 — gives the following high probability guarantee whenever $W > 2p+3$:

$$\Delta_\tau(\hat{\theta}) \leq \tilde{O}\left(\frac{W}{n}\left(\frac{Q}{\tau}\right)^{\frac{2p}{p+3}}\right),$$

where $\tilde{O}(\cdot)$ hides polynomial dependence on p. Since $\|f(x)\|_\infty \leq Q$, under the procedure of Remark 2, we obtain that $\tau \asymp Q(\frac{W}{n})^{\frac{p+3}{4p+6}}$ yields

$$\left\|\hat{f} - f\right\|^2_{L_2(\mu_X)} = \tilde{O}\left(Q^2\left(\frac{W}{n}\right)^{\frac{p+3}{2p+3}}\right).$$

For fixed p, Q and $W > 2p+3$, this rate matches, up to logarithmic factors, the squared-L_2 minimax lower bound for Gaussian-noise regression (Yang & Zhou, 2025, Remark 4.3) and the best known upper-bound exponent for noiseless i.i.d. sampling (Parhi & Adcock, 2025, Equation (39), $k = 1$).

In particular, the L_2-norm exponent tends to $1/4$ as $p \to \infty$, showing that the curse of dimensionality does not apply for Barron classes.

I.2 Half space example for Section 3.3

Example: Learning half spaces Consider the space of functions indicating p-dimensional half spaces:

$$\mathrm{Halfspaces}(R) = \left\{f_{a,b}(x) = \mathbf{1}\{a^T x - b > 0\} : a \in \mathbb{S}^{p-1},\ |b| \leq R\right\}.$$

Assume that our data is distributed i.i.d. according to μ_X, and that any linear marginal of μ_X has bounded measure: i.e. there exists Λ so that for any $a \in \mathbb{S}^{p-1}$, it holds that $\Pr_{x\sim\mu_X}\left[a^T x \in I\right] \leq$

$\Lambda|I|$[6]. This assumption is satisfied by e.g. any uniform distribution on a bounded positive volume set, and vectors with i.i.d. entries from bounded-density measures

Take any $f \in \text{Halfspaces}(R)$ with $f(x) = \mathbf{1}\{a^T x - b > 0\}$. Because f is discontinuous and ReLU networks are continuous, the τ-realizability condition of Theorem 3 is unsatisfiable for any $\tau < 1$. However, we may define an approximant $f_h(x) := \frac{1}{2h}\left(\sigma(a^T x - b + h) - \sigma(a^T x - b - h)\right)$ that agrees with f wherever $|a^T x - b| \geq h$. Notice that f_h is realized by some $\theta_h \in \mathcal{C}_{A_h}(2, 1)$, where $A_h = \max\left\{1, \frac{1}{2h}, R + h\right\}$. We also have $\sup_x |f(x) - f_h(x)| \leq \frac{1}{2}$ so that the bounded difference hypothesis is satisfied with $\Gamma = \frac{1}{2}$. The measure of the disagreement set between f_h and f is given by $\Pr_{x\sim\mu_X}\left[a^T x \in (b - h, b + h)\right] \leq 2\Lambda h$. Taking $h_n = \frac{1}{2\Lambda n}$ allows us to apply Theorem 8 with $\epsilon_n = \frac{1}{n}$, $w_n = 2$, $d_n = 1$, $A_n = A_{h_n}$, and $D_n = D \geq 1$.

Set $m_\delta := \lceil 2 + \log(4/\delta)\rceil$ and require

$$W > 6 + 4\left\lceil \frac{m_\delta}{D}\right\rceil.$$

We choose M_n polynomial in n and the problem parameters and large enough to satisfy the lower bound in the theorem statement. Then, for $\hat{\theta} \sim U_{\tau,n}^{W,D,M_n}$, we have

$$\Pr_{x^n,\hat{\theta}}\left[\Delta_\tau(\hat{\theta}) \geq \tilde{O}\left(\left(\frac{WD}{n}\right)\log q_{\text{eff},n}\right)\right] \leq \delta,$$

where $q_{\text{eff},n}$ is polynomial in $n, D, \Lambda, \delta^{-1}, R, \tau^{-1}$ and the fixed problem parameters. Since f takes values in $[0, 1]$, the convention of Remark 2 gives $\|\hat{f}_n - f\|_{L_1(\mu_X)} \leq \tau + \Delta_\tau(\hat{\theta})$. Thus, if $\tau \asymp 1/n$, with probability at least $1 - \delta$,

$$\|\hat{f}_n - f\|_{L_1(\mu_X)} = \tilde{O}\left(\frac{WD}{n}\right).$$

As an extension, results can be derived for learning intersections of halfspaces using additional witness depth (i.e. polytopes), which can be used to approximate general convex bodies with continuous boundaries.