Sparsity-aware generalization theory for deep neural networks

Ramchandran Muthukumar RMUTHUK1@JHU.EDU
Department of Computer Science & Mathematical Institute for Data Science, Johns Hopkins University

Jeremias Sulam JSULAM1@JHU.EDU
Department of Biomedical Engineering & Mathematical Institute for Data Science, Johns Hopkins University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Deep artificial neural networks achieve surprising generalization abilities that remain poorly understood. In this paper, we present a new approach to analyzing generalization for deep feed-forward ReLU networks that takes advantage of the degree of sparsity that is achieved in the hidden layer activations. By developing a framework that accounts for this reduced effective model size for each input sample, we are able to show fundamental trade-offs between sparsity and generalization. Importantly, our results make no strong assumptions about the degree of sparsity achieved by the model, and it improves over recent norm-based approaches. We illustrate our results numerically, demonstrating non-vacuous bounds when coupled with data-dependent priors in specific settings, even in over-parametrized models.

1. Introduction

Statistical learning theory seeks to characterize the generalization ability of machine learning models, obtained from finite training data, to unseen test data. The field is by now relatively mature, and several tools exist to provide upper bounds on the generalization error, $R(h)$. Often the upper bounds depend on the empirical risk, $\hat{R}(h)$, and different characterizations of complexity of the hypothesis class as well as potentially specific data-dependent properties. The renewed interest in deep artificial neural network models has demonstrated important limitations of existing tools. For example, VC dimension often simply relates to the number of model parameters and is hence insufficient to explain generalization of overparameterized models (Bartlett et al., 2019). Traditional measures based on Rademacher complexity are also often vacuous, as these networks can indeed be trained to fit random noise (Zhang et al., 2017). Margin bounds have been adapted to deep non-linear networks (Bartlett et al., 2017; Golowich et al., 2018; Neyshabur et al., 2015, 2018), albeit still unable to provide practically informative results.

An increasing number of studies advocate for non-uniform data-dependent measures to explain generalization in deep learning (Nagarajan and Kolter, 2019a; Pérez and Louis, 2020; Wei and Ma, 2019). Of particular interest are those that employ the sensitivity of a data-dependent predictor to parameter perturbations – sometimes also referred to as *flatness* (Shawe-Taylor and Williamson, 1997; Neyshabur et al., 2017; Dziugaite and Roy, 2017; Arora et al., 2018; Li et al., 2018; Nagarajan and Kolter, 2019b; Wei and Ma, 2019; Sulam et al., 2020; Banerjee et al., 2020). This observation has received some empirical validation as well (Zhang et al., 2017; Keskar et al., 2017; Izmailov et al., 2018; Neyshabur et al., 2019; Jiang* et al., 2020; Foret et al., 2021). Among the theoretical results of this line of work, Arora et al. (2018) study the generalization properties of a *compressed* network, and Dziugaite and Roy (2017); Neyshabur et al. (2017) study a stochastic perturbed version of the

original network. The work in (Wei and Ma, 2019) provides improved bounds on the generalization error of neural networks as measured by a low Jacobian norm with respect to training data, while Wei and Ma (2020) capture the sensitivity of a neural network to perturbations in intermediate layers. PAC-Bayesian analysis provides an alternate way of studying generalization by incorporating prior knowledge on a distribution of well-performing predictors in a Bayesian setting (McAllester, 1998; Guedj, 2019; Alquier, 2021). Recent results (Dziugaite and Roy, 2017, 2018; Zhou et al., 2019) have further strengthened the standard PAC-Bayesian analysis by optimizing over the posterior distribution to generate non-vacuous bounds on the expected generalization error of stochastic neural networks. Derandomized versions of PAC-Bayes bounds have also been recently developed (Nagarajan and Kolter, 2019b; Banerjee et al., 2020) relying on the sensitivity or *noise resilience* of an obtained predictor. All of these works are insightful, alas important gaps remain in understanding generalization in non-linear, over-parameterized networks (Pérez and Louis, 2020).

Our contributions. In this work we employ tools of sensitivity analysis and PAC-Bayes bounds to provide generalization guarantees on deep ReLU feed-forward networks. Our key contribution is to make explicit use of the sparsity achieved by these networks across their different layers, reflecting the fact that only sub-networks, of reduced sizes and complexities, are active at every sample. Similar in spirit to the observations in Muthukumar and Sulam (2022), we provide conditions under which the set of active neurons (smaller than the number of total neurons) is stable over suitable distributions of networks, with high-probability. In turn, these results allow us to instantiate recent de-randomized PAC-Bayes bounds (Nagarajan and Kolter, 2019b) and obtain new guarantees that do not depend on the global Lipschitz constant, nor are they exponential in depth. Importantly, our results provide data-dependent non-uniform guarantees that are able to leverage the structure (sparsity) obtained on a specific predictor. As we show experimentally, this degree of sparsity – the reduced number of active neurons – need not scale linearly with the width of the model or the number of parameters, thus obtaining bounds that are significantly tighter than known results. We also illustrate our generalization results on MNIST for models of different width and depth, providing non-vacuous bounds in certain settings.

Manuscript organization. After introducing basic notation, definitions and problem settings, we provide a detailed characterization of stable inactive sets in single-layer feed-forward maps in Section 2. Section 3 presents our main results by generalizing our analysis to multiple layers, introducing appropriate distributions over the hypothesis class and tools from de-randomized PAC-Bayes theory. We demonstrate our bounds numerically in Section 4, and conclude in Section 5.

1.1. Notation And Definitions

Sets and spaces are denoted by capital (and often calligraphic) letters, with the exception of the set $[K] = \{1, \dots, K\}$. For a Banach space $\mathcal{W}$ embedded with norm $\|\cdot\|_{\mathcal{W}}$, we denote by $\mathcal{B}_r^{\mathcal{W}}(\mathbf{W})$, a bounded ball centered around $\mathbf{W}$ with radius r. Throughout this work, scalar quantities are denoted by lower or upper case (not bold) letters, and vectors with bold lower case letters. Matrices are denoted by bold upper case letters: $\mathbf{W}$ is a matrix with *rows* $\mathbf{w}[i]$. We denote by $\mathcal{P}_I$, the index selection operator that restricts input to the coordinates specified in the set I. For a vector $\mathbf{x} \in \mathbb{R}^d$ and $I \subset [d]$, $\mathcal{P}_I : \mathbb{R}^d \to \mathbb{R}^{|I|}$ is defined as $\mathcal{P}_I(\mathbf{x}) := \mathbf{x}[I]$. For a matrix $\mathbf{W} \in \mathbb{R}^{p \times d}$ and $I \subset [p]$, $\mathcal{P}_I(\mathbf{W}) \in \mathbb{R}^{|I| \times d}$ restricts $\mathbf{W}$ to the *rows* specified by I. For row and column index sets $I \subset [p]$ and $J \subset [d]$, $\mathcal{P}_{I,J}(\mathbf{W}) \in \mathbb{R}^{|I| \times |J|}$ restricts $\mathbf{W}$ to the corresponding sub-matrix. Throughout this work, we refer to *sparsity* as the *number of zeros* of a vector, so that for $\mathbf{x} \in \mathbb{R}^d$ with degree of sparsity s,

$\|\mathbf{x}\|_0 = d - s$. We denote the induced operator norm by $\|\cdot\|_2$, and the Frobenius norm by $\|\cdot\|_F$. In addition, we will often use operator norms of reduced matrices induced by sparsity patterns. To this end, the following definition will be used extensively.

Definition 1 *(Sparse Induced Norms) Let* $\mathbf{W} \in \mathbb{R}^{d_2 \times d_1}$ *and* (s_2, s_1) *be sparsity levels such that* $0 \leq s_1 \leq d_1 - 1$ *and* $0 \leq s_2 \leq d_2 - 1$. *We define the* (s_2, s_1) *sparse induced norm* $\|\cdot\|_{(s_2,s_1)}$ *as*

$$\|\mathbf{W}\|_{(s_2,s_1)} := \max_{|J_2|=d_2-s_2} \max_{|J_1|=d_1-s_1} \|\mathcal{P}_{J_2,J_1}(\mathbf{W})\|_2 .$$

The sparse induced norm $\|\cdot\|_{(s_2,s_1)}$ measures the induced operator norm of a worst-case sub-matrix. For any two sparsity vectors $(s_2, s_1) \preceq (\hat{s}_2, \hat{s}_1)$, one can show that $\|\mathbf{W}\|_{(\hat{s}_2,\hat{s}_1)} \leq \|\mathbf{W}\|_{(s_2,s_1)}$ for any matrix $\mathbf{W}$ (see Lemma 17). In particular,

$$\max_{i,j} |\mathbf{W}[i,j]| = \|\mathbf{W}\|_{(d_2-1,d_1-1)} \leq \|\mathbf{W}\|_{(s_2,s_1)} \leq \|\mathbf{W}\|_{(0,0)} = \|\mathbf{W}\|_2 .$$

Thus, the sparse norm interpolates between the maximum absolute entry norm and the operator norm. Frequently in our exposition we rely on the case when $s_2 = d_2 - 1$, thus obtaining $\|\mathbf{W}\|_{(d_2-1,s_1)} = \max_{i\in[d_2]} \max_{|J_1|=d_1-s_1} \|\mathcal{P}_{J_1}(\mathbf{w}[i])\|_2$, the maximum norm of any reduced row of matrix $\mathbf{W}$.

Outside of the special cases listed above, computing the sparse norm for a general (s_2, s_1) has combinatorial complexity. Instead, a modified version of the babel function (see Tropp et al. (2003)) provides computationally efficient upper bounds[1].

Definition 2 *(Reduced Babel Function (Muthukumar and Sulam, 2022)) Let* $\mathbf{W} \in \mathbb{R}^{d_2 \times d_1}$, *the reduced babel function at row sparsity level* $s_2 \in \{0, \ldots, d_2 - 1\}$ *and column sparsity level* $s_1 \in \{0, \ldots, d_1 - 1\}$ *is defined as*[2],

$$\mu_{s_2,s_1}(\mathbf{W}) := \frac{1}{\|\mathbf{W}\|^2_{(d_2-1,s_1)}} \max_{\substack{J_2\subset[d_2],\\ |J_2|=d_2-s_2}} \max_{j\in J_2} \left[\sum_{\substack{i\in J_2,\\ i\neq j}} \max_{\substack{J_1\subseteq[d_1]\\ |J_1|=d_1-s_1}} \left|\mathcal{P}_{J_1}(\mathbf{w}[i])\mathcal{P}_{J_1}(\mathbf{w}[j])^T\right| \right].$$

For the special case when $s_2 = 0$, the reduced babel function is equivalent to the babel function from Tropp et al. (2003) on the transposed matrix $\mathbf{W}^T$. We show in Lemma 18 that the sparse-norm can be bounded using the reduced babel function and the maximum reduced row norm $\|\cdot\|_{(d_2-1,s_1)}$,

$$\|\mathbf{W}\|_{s_2,s_1} \leq \|\mathbf{W}\|_{d_2-1,s_1} \sqrt{1 + \mu_{s_2,s_1}(\mathbf{W})}. \tag{1}$$

See Appendix D for a computationally efficient implementation of the reduced babel function.

1.2. Learning Theoretic Framework

We consider the task of multi-class classification with a bounded input space $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{d_0} \mid \|\mathbf{x}\|_2 \leq \mathsf{M}_{\mathcal{X}}\}$ and labels $\mathcal{Y} = \{1, \ldots, C\}$ from an unknown distribution $\mathcal{D}_{\mathcal{Z}}$ over $\mathcal{Z} := (\mathcal{X} \times \mathcal{Y})$. We search for a hypothesis in $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}'\}$ that is an accurate predictor of label y given input $\mathbf{x}$. Note

1. The particular definition used in this paper is weaker but more computationally efficient than that introduced in Muthukumar and Sulam (2022).
2. When $s_2 = d_2 - 1, |J_2| = 1$, we simply define $\mu_{(s_2,s_1)}(\mathbf{W}) := 0$.

that $\mathcal{Y}$ and $\mathcal{Y}'$ need not be the same. In this work, we consider $\mathcal{Y}' = \mathbb{R}^C$, and consider the predicted label of the hypothesis h as $\hat{y}(\mathbf{x}) := \operatorname{argmax}_j [h(\mathbf{x})]_j$[3]. The quality of prediction of h at $\mathbf{z} = (\mathbf{x}, y)$ is informed by the margin defined as $\rho(h, \mathbf{z}) := \left([h(\mathbf{x})]_y - \operatorname{argmax}_{j \neq y} [h(\mathbf{x})]_j\right)$. If the margin is positive, then the predicted label is correct. For a threshold hyper-parameter $\gamma \geq 0$, we define a γ-threshold 0/1 loss ℓ_γ based on the margin as $\ell_\gamma(h, \mathbf{z}) := \mathbb{1}\{\rho(h, \mathbf{z}) < \gamma\}$. Note that ℓ_γ is a stricter version of the traditional zero-one loss ℓ_0, since $\ell_0(h, \mathbf{z}) \leq \ell_\gamma(h, \mathbf{z})$ for all $\gamma \geq 0$. With these elements, the *population risk* (also referred to as *generalization error*) of a hypothesis R_γ is the expected loss it incurs on a randomly sampled data point, $R_\gamma(h) := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_\mathcal{Z}} \left[\ell_\gamma\left(h, \mathbf{z}\right)\right]$. The goal of supervised learning is to obtain a hypothesis with low population risk $R_0(h)$, the probability of misclassification. While the true distribution $\mathcal{D}_\mathcal{Z}$ is unknown, we assume access to an i.i.d training set $\mathtt{S}_T = \{\mathbf{z}^{(i)}, \ldots, \mathbf{z}^{(m)}\} \sim (\mathcal{D}_\mathcal{Z})^m$ and we seek to minimize the *empirical risk* $\hat{R}_\gamma$, the average loss incurred on the training sample $\mathtt{S}_T$, i.e. $\hat{R}_\gamma(h) := \frac{1}{m} \sum_{i=1}^m \ell_\gamma\left(h, \mathbf{z}^{(i)}\right)$. We shall later see that for any predictor, $R_0(h)$ can be upper bounded using the stricter empirical risk $\hat{R}_\gamma(h)$ for an appropriately chosen $\gamma > 0$.

In this work, we study the hypothesis class $\mathcal{H}$ containing feed-forward neural networks with K hidden layers. Each hypothesis $h \in \mathcal{H}$ is identified with its weights $\{\mathbf{W}_k\}_{k=1}^{K+1}$, and is a sequence of K linear maps $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ composed with a nonlinear activation function $\sigma(\cdot)$ and a final linear map $\mathbf{W}_{K+1} \in \mathbb{R}^{C \times d_K}$,

$$h(\mathbf{x}_0) := \mathbf{W}_{K+1} \sigma\left(\mathbf{W}_k \sigma\left(\mathbf{W}_{K-1} \cdots \sigma\left(\mathbf{W}_1 \mathbf{x}_0\right) \cdots\right)\right).$$

We exclude bias from our definitions of feed-forward layers for simplicity[4]. We denote by $\mathbf{x}_k$ the k^{th} hidden layer representation of network h at input $\mathbf{x}_0$, so that $\mathbf{x}_k := \sigma\left(\mathbf{W}_k \mathbf{x}_{k-1}\right)\ \forall 1 \leq k \leq K$, and $h(\mathbf{x}) := \mathbf{W}_{K+1} \mathbf{x}_K$. Throughout this work, the activation function is assumed to be the Rectifying Linear Unit, or ReLU, defined by $\sigma(x) = \max\{x, 0\}$, acting entrywise on an input vector.

2. Warm Up: Sparsity In Feed-Forward Maps

As a precursor to our sensitivity analysis for multi-layer feed-forward networks, we first consider a generic feed-forward map $\Phi(\mathbf{x}) := \sigma(\mathbf{W}\mathbf{x})$. A naïve bound on the norm of the function output is $\|\Phi(\mathbf{x})\|_2 \leq \|\mathbf{W}\|_2 \|\mathbf{x}\|_2$, but this ignores the sparsity of the output of the feed-forward map (due to the ReLU). Suppose there exists a set I of inactive indices such that $\mathcal{P}_I(\Phi(\mathbf{x})) = \mathbf{0}$, i.e. for all $i \in I$, $\mathbf{w}[i] \cdot \mathbf{x} \leq 0$. In the presence of such an index set, clearly $\|\Phi(\mathbf{x})\|_2 \leq \|\mathcal{P}_{I^c}(\mathbf{W})\|_2 \|\mathbf{x}\|_2$[5]. Thus, estimates of the effective size of the feed-forward output, and other notions such as sensitivity to parameter perturbations, can be refined by accounting for the sparsity of activation patterns. Note that the inactive index set I varies with each input, $\mathbf{x}$, and with the parameters of predictor, $\mathbf{W}$.

For some $\zeta_0, \xi_1, \eta_1 > 0$ and sparsity levels s_1, s_0, let $\mathcal{X}_0 = \{\mathbf{x} \in \mathbb{R}^{d_0} \mid \|\mathbf{x}\|_2 \leq \zeta_0,\ \|\mathbf{x}\|_0 \leq d_0 - s_0\}$ denote a bounded sparse input domain and let $\mathcal{W}_1 := \{\mathbf{W} \in \mathbb{R}^{d_1 \times d_0} \mid \|\mathbf{W}\|_{(d_1-1, s_0)} \leq \xi_1,\ \mu_{s_1, s_0}(\mathbf{W}) \leq \eta_1\}$ denote a parameter space. We now define a radius function that measures the amount of relative perturbation within which a certain inactive index set is stable.

3. The argmax here is assumed to break ties deterministically.
4. This is a standard choice in related works, e.g. Bartlett et al. (2017). Our analysis can be expanded to account for bias.
5. I^c is the complement of the index set I, also referred to as J when clear from context.

Definition 3 *(Sparse local radius[6]) For any weight $\mathbf{W} \in \mathbb{R}^{d_1 \times d_0}$, input $\mathbf{x} \in \mathbb{R}^{d_0}$ and sparsity level $1 \leq s_1 \leq d_1$, we define a sparse local radius and a sparse local index set as*

$$r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) := \sigma\left(\text{SORT}\left(-\frac{\mathbf{W}\cdot\mathbf{x}}{\xi_1\zeta_0},\ s_1\right)\right), \quad I(\mathbf{W}, \mathbf{x}, s_1) := \text{TOP-K}\left(-\frac{\mathbf{W}\cdot\mathbf{x}}{\xi_1\zeta_0}, s_1\right). \tag{2}$$

Here, $\text{TOP-K}(\mathbf{u}, j)$ is the index set of the top j entries in $\mathbf{u}$, and $\text{SORT}(\mathbf{u}, j)$ is its j^{th} largest entry.

We note that when evaluated on a weight $\mathbf{W} \in \mathcal{W}_1$ and input $\mathbf{x} \in \mathcal{X}_0$, for all sparsity levels the sparse local radius $r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) \in [0, 1]$. We denote the sparse local index set as I when clear from the context. We now analyze the stability of the sparse local index set and the resulting reduced sensitivity of model output. For brevity, we must defer all proofs to the appendix.

Lemma 4 *Let $\epsilon_0 \in [0, 1]$ be a relative input corruption level and let $\epsilon_1 \in [0, 1]$ be the relative weight corruption. For the feed-forward map Φ with weight $\mathbf{W} \in \mathcal{W}_1$ and input $\mathbf{x} \in \mathcal{X}_0$, the following statements hold for any output sparsity level $1 \leq s_1 \leq d_1$,*

1. ***Existence of an inactive index set and bounded outputs**: If $r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) > 0$, then the index set $I(\mathbf{W}, \mathbf{x}, s_1)$ is inactive for $\Phi(\mathbf{x})$. Moreover, $\|\Phi(\mathbf{x})\|_2 \leq \xi_1\sqrt{1+\eta_1}\cdot\zeta_0$.*

2. ***Stability of an inactive index set to input and parameter perturbations**: Suppose $\hat{\mathbf{x}}$ and $\hat{\mathbf{W}}$ are perturbed inputs and weights respectively such that, $\|\hat{\mathbf{x}} - \mathbf{x}\|_0 \leq d_0 - s_0$ and,*

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_2}{\zeta_0} \leq \epsilon_0 \ \text{ and } \ \max\left\{\frac{\left\|\hat{\mathbf{W}} - \mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}, \frac{\left\|\hat{\mathbf{W}} - \mathbf{W}\right\|_{(s_1,s_0)}}{\xi_1\sqrt{1+\eta_1}}\right\} \leq \epsilon_1,$$

   and denote $\hat{\Phi}(\mathbf{x}) = \sigma(\hat{\mathbf{W}}\mathbf{x})$. If $r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) \geq -1 + (1+\epsilon_0)(1+\epsilon_1)$, then the index set $I(\mathbf{W}, \mathbf{x}, s_1)$ is inactive and stable to perturbations, i.e.[7] $\mathcal{P}_I(\Phi(\mathbf{x})) = \mathcal{P}_I(\Phi(\hat{\mathbf{x}})) = \mathcal{P}_I(\hat{\Phi}(\hat{\mathbf{x}})) = \mathbf{0}$. Moreover, $\left\|\hat{\Phi}(\hat{\mathbf{x}}) - \Phi(\mathbf{x})\right\|_2 \leq (-1 + (1+\epsilon_0)(1+\epsilon_1))\cdot\xi_1\sqrt{1+\eta_1}\cdot\zeta_0$.

3. ***Stability of sparse local radius**: For a perturbed input $\hat{\mathbf{x}}$ such that $\|\hat{\mathbf{x}} - \mathbf{x}\|_0 \leq d_0 - s_0$, and perturbed weight $\hat{\mathbf{W}}$, the difference between sparse local radius is bounded*

$$\left|r_{\text{sparse}}(\hat{\mathbf{W}}, \hat{\mathbf{x}}, s_1) - r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1)\right| \leq -1 + \left(1 + \frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_2}{\zeta_0}\right)\left(1 + \frac{\left\|\hat{\mathbf{W}} - \mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right).$$

A key takeaway of this Lemma (see Appendix A.1.1 for its proof) is that one can obtain tighter bounds, on both the size of the network output as well as its sensitivity to corruptions, if the corresponding sparse local radius is sufficiently large. The results above quantify these notions for a given sample. In the next section, we will leverage this characterization within the framework of PAC-Bayes analysis to provide a generalization bound for feed-forward networks.

6. The definition here is inspired by Muthukumar and Sulam (2022) but stronger.
7. For notational ease we suppress arguments and let $I = I(\mathbf{W}, \mathbf{x}, s_1)$.

$\mathbf{s} = \{s_1, \ldots, s_k\}, \;\; 0 \leq s_k \leq d_k - 1$	Layer wise sparsity vector
$\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_{K+1}\}, \;\; 0 \leq \xi_k$	Layer wise bound on $\|\cdot\|_{(d_k-1, s_{k-1})}$
$\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_K\}, \;\; 0 \leq \eta_k$	Layer wise bound on $\mu_{s_k, s_{k-1}}(\cdot)$
$\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_{K+1}\}, \;\; 0 \leq \epsilon_k$	Layer wise bound on relative perturbation

Table 1: Independent base hyper-parameters

3. A Sparsity-Aware Generalization Theory

We shall construct non-uniform data-dependent generalization bounds for feed-forward networks based on a local sensitivity analysis of deep ReLU networks, employing the intuition from the previous section. To do so, we will first study the size of the layer outputs using Definition 2, then measure the sensitivity in layer outputs to parameter perturbations using Lemma 4 across multiple layers, and finally leverage a derandomized PAC-Bayes result from Nagarajan and Kolter (2019b) (see Appendix C.2). Before embarking on the analysis, we note the following convenient property of the margin for any two predictors $h, \hat{h}$ from (Bartlett et al., 2017, Lemma A.3),

$$\left| \left(h(\mathbf{x})_y - \max_{j \neq y} h(\mathbf{x})_j \right) - \left(\hat{h}(\mathbf{x})_y - \max_{j \neq y} \hat{h}(\mathbf{x})_j \right) \right| \leq 2 \left\| \hat{h}(\mathbf{x}) - h(\mathbf{x}) \right\|_\infty .$$

Hence, quantifying the sensitivity of the predictor outputs will inform the sensitivity of the loss. Similar to other works (Nagarajan and Kolter, 2019b; Banerjee et al., 2020), our generalization bound will be derived by studying the sensitivity of neural networks upon perturbations to the layer weights.

For the entirety of this section, we fix a set of *base hyper-parameters* that determine a specific class of neural networks, the variance of a posterior distribution over networks, and the resolution (via a sparsity vector) at which the generalization is measured – see Table 1 for reference. We denote by $\mathbf{s} = \{s_1, \ldots, s_K\}$ a vector of layer-wise sparsity levels, which reflects the inductive bias of the learner on the potential degree of sparsity of a trained network on the training data. Next we define two hyper-parameters, $\boldsymbol{\xi} := \{\xi_1, \ldots, \xi_{K+1}\}$ where $\xi_k > 0$ bounds the sparse norm $\|\cdot\|_{(d_k-1, s_{k-1})}$ of the layer weights and $\boldsymbol{\eta} := \{\eta_1, \ldots, \eta_K\}$ where $\eta_k > 0$ bounds the reduced babel function $\mu_{s_k, s_{k-1}}(\cdot)$ of the layer weights. Finally, we let $\boldsymbol{\epsilon} := \{\epsilon_1, \ldots, \epsilon_{K+1}\}$ with $\epsilon_k > 0$ bound the amount of relative perturbation in the weights. This section treats the quartet $(\mathbf{s}, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\epsilon})$ as constants[8], while in the next section we shall discuss appropriate values for these hyper-parameters.

Definition 5 *(Norm bounded feed-forward networks) We define below the parameter domain $\mathcal{W}_k$ and a class of feed-forward networks $\mathcal{H}_{K+1}$ with K-hidden layers,*

$$\mathcal{W}_k := \left\{ \mathbf{W} \in \mathbb{R}^{d_k \times d_{k-1}} \mid \|\mathbf{W}\|_{(d_k-1, s_{k-1})} \leq \xi_k, \quad \mu_{s_k, s_{k-1}}(\mathbf{W}) \leq \eta_k, \right\}, \; \forall \; k \in [K],$$

$$\mathcal{H} := \left\{ h(\cdot) := \mathbf{W}_{K+1} \sigma \left(\mathbf{W}_K \cdots \sigma \left(\mathbf{W}_1 \cdot \right) \right) \mid \|\mathbf{W}_{K+1}\|_{(C-1, s_K)} \leq \xi_{K+1}, \; \mathbf{W}_k \in \mathcal{W}_k, \; \forall \; k \in [K] \right\}.$$

To measure the local sensitivity of the network outputs, it will be useful to formalize a notion of local neighborhood for networks.

8. Unless otherwise specified we let $s_0 = s_{K+1} = 0$ and $\epsilon_0 = 0$.

$\zeta_k := \xi_k\sqrt{1+\eta_k}\cdot\zeta_{k-1}\,, \quad \forall\, k \in [K]$	Bound on norm of layer outputs
$\zeta_{K+1} := \xi_{K+1}\zeta_K$	Bound on norm of network output
$\gamma_k := -1 + \prod_{n=1}^{k}(1+\epsilon_n), \quad \forall\, k \in [K+1]$	Layer wise threshold for local radius
$r_k(h,\mathbf{z}) := \sigma\left(\text{sort}\left(-\left[\frac{\mathbf{w}_k[i]\cdot\mathbf{x}_{k-1}}{\xi_k\zeta_{k-1}}\right]_{i=1}^{d_k},\ d_k - s_k\right)\right)$	Layer-wise sparse local radius

Table 2: Layer-wise bounds and thresholds.

Definition 6 *(Local Neighbourhood) Given $h \in \mathcal{H}$, define $\mathcal{B}(h,\boldsymbol{\epsilon})$ to be the local neighbourhood around h containing perturbed networks $\hat{h}$ with weights $\{\hat{\mathbf{W}}_j\}_{k=1}^{K+1}$ such that at each layer k*[9]*,*

$$\max\left\{\frac{\left\|\hat{\mathbf{W}}_k - \mathbf{W}_k\right\|_{(s_k,s_{k-1})}}{\xi_k\sqrt{1+\eta_k}}, \frac{\left\|\hat{\mathbf{W}}_k - \mathbf{W}_k\right\|_{(d_k-1,s_{k-1})}}{\xi_k}\right\} \leq \epsilon_k.$$

It will be useful to understand the probability that $\hat{h} \in \mathcal{B}(h,\boldsymbol{\epsilon})$ when the perturbations to each layer weight are random, in particular from Gaussian distributions over feed-forward networks:

Definition 7 *(Entrywise Gaussian) Let $h \in \mathcal{H}$ be any network with $K+1$ layers, and let $\boldsymbol{\sigma}^2 := \{\sigma_1^2,\ldots,\sigma_{K+1}^2\}$ be a layer-wise variance. We denote by $\mathcal{N}(h,\boldsymbol{\sigma}^2)$ a distribution with mean network h such that for any $\hat{h} \sim \mathcal{N}(h,\boldsymbol{\sigma}^2)$ with layer weights $\hat{\mathbf{W}}_k$, each entry $\hat{\mathbf{W}}_k[i,j] \sim \mathcal{N}(\mathbf{W}_k[i,j],\sigma_k^2)$.*

3.1. Sensitivity Of Network Output

Given a predictor $h \in \mathcal{H}$, note that the size of a network output for any given input is bounded by $\|h(\mathbf{x}_0)\|_2 \leq \prod_{k=1}^{K+1}\|\mathbf{W}_k\|_2\,\mathsf{M}_{\mathcal{X}}$, which ignores the sparsity of the intermediate layers. We will now generalize the result in Lemma 4 by making use of the inactive index sets at every layer I_k, such that $\mathcal{P}_{I_k}(\mathbf{x}_k) = \mathbf{0}$, obtaining a tighter (input dependent) characterization of sensitivity to perturbations of the network. For notational convenience, we define two additional dependent notations: we let $\zeta_0 := \mathsf{M}_{\mathcal{X}}$ and $\zeta_k := \xi_k\sqrt{1+\eta_k}\cdot\zeta_{k-1} = \mathsf{M}_{\mathcal{X}}\prod_{n=1}^{k}\xi_n\sqrt{1+\eta_n}$ denote a bound on the layer-wise size of the outputs. At the final layer, we let $\zeta_{K+1} := \xi_{K+1}\zeta_K$ as a bound on the network output. Additionally, we define $\gamma_k := -1 + \prod_{n=1}^{k}(1+\epsilon_n)$ as a threshold on the sparse local radius evaluated at each layer – see Table 2 for a summary. In the last layer, we let this value γ_{K+1} represent the desired margin. For networks $\hat{h}$ with perturbed weights $\hat{\mathbf{W}}$, we denote by $\hat{\mathbf{x}}_k := \sigma\left(\hat{\mathbf{W}}_k\hat{\mathbf{x}}_{k-1}\right)$ the perturbed layer representation corresponding to input $\mathbf{x}_0$.

Definition 8 *(Layer-wise sparse local radius) Let h be any feed-forward network with weighs $\mathbf{W}_k \in \mathbb{R}^{d_k\times d_{k-1}}$, and let $\mathbf{x}_0 \in \mathbb{R}^{d_0}$. We define a layer-wise sparse local radius and a layer-wise inactive index set as below,*

$$I_k(h,\mathbf{x}_0) := \textsc{Top-k}\left(-\frac{\mathbf{W}_k\cdot\mathbf{x}_{k-1}}{\xi_k\zeta_{k-1}}, s_k\right), \quad r_k(h,\mathbf{x}_0) := \sigma\left(\text{sort}\left(-\frac{\mathbf{W}_k\cdot\mathbf{x}_{k-1}}{\xi_k\zeta_{k-1}},\ s_k\right)\right).$$

Definition 8 now allows us, by employing Lemma 4, to generalize our previous observations to entire network models, as we now show.

9. For the last layer we only require $\left\|\hat{\mathbf{W}}_{K+1} - \mathbf{W}_{K+1}\right\|_{C-1,s_K} \leq \epsilon_{K+1}\cdot\xi_{K+1}$.

Proposition 9 *Let $h \in \mathcal{H}$, if at each layer k the layer-wise sparse local radius is nontrivial, i.e. $\forall\, k \in [K],\ \ r_k(h, \mathbf{x}_0) > 0$. Then the index sets $I_k(h, \mathbf{x}_0)$ are inactive at layer k and the size of the hidden layer representations and the network output are bounded as follows,*

$$\forall\, k \in [K], \quad \|\mathbf{x}_k\|_2 \leq \zeta_k, \quad \textit{and} \quad \|h(\mathbf{x}_0)\|_\infty \leq \zeta_{K+1}. \tag{3}$$

In a similar vein, we can characterize the sensitivity of the network to parameter perturbations.

Proposition 10 *Let $h \in \mathcal{H}$ and let $\hat{h} \in \mathcal{B}(h, \boldsymbol{\epsilon})$ be a nearby perturbed predictor with weights $\{\hat{\mathbf{W}}_k\}$. If each layer-wise sparse local radius is sufficiently large, i.e. $\forall\, k \in [K],\ r_k(h, \mathbf{x}_0) \geq \gamma_k$, then the index sets $I_k(h, \mathbf{x}_0)$ are inactive for the perturbed layer representations $\hat{\mathbf{x}}_k$ and the distance between the layer representations and the network output are bounded as follows,*

$$\forall\, k \in [K], \quad \|\hat{\mathbf{x}}_k - \mathbf{x}_k\|_2 \leq \zeta_k \cdot \gamma_k, \quad \textit{and} \quad \left\|\hat{h}(\mathbf{x}_0) - h(\mathbf{x}_0)\right\|_\infty \leq \zeta_{K+1} \cdot \gamma_{K+1}. \tag{4}$$

Proofs of the above propositions can be found in A.1.2 and A.1.3 respectively.

3.2. Sparsity-Aware Generalization

We are now ready to state our main theorem on generalization of feed-forward networks that leverages improved sensitivity of network outputs due to stable inactive index sets.

Theorem 11 *Let $\mathcal{P}$ be any prior distribution over depth-$(K+1)$ feed-forward network chosen independently of the training sample. Let $h \in \mathcal{H}$ be any feed-forward network (possibly trained on sample data), with $\mathcal{H}$ determined by fixed base hyper-parameters $(\mathbf{s}, \boldsymbol{\epsilon}, \boldsymbol{\xi}, \boldsymbol{\eta})$, and denote the sparse loss by $\ell_{\text{sparse}}(h, \mathbf{x}) = \mathbb{1}\{\exists\ k,\ r_k(h, \mathbf{x}) < 3\gamma_k\}$. With probability at least $(1-\delta)$ over the choice of i.i.d training sample S_T of size m, the generalization error of h is bounded as follows,*

$$R_0(h) \leq \hat{R}_{4\zeta_{K+1}\gamma_{K+1}}(h) + \frac{2K}{m} \sum_{\mathbf{x}^{(i)} \in S_T} \ell_{\text{sparse}}(h, \mathbf{x}^{(i)}) + \tilde{\mathcal{O}}\left(\sqrt{\frac{\text{KL}\left(\mathcal{N}\left(h, \boldsymbol{\sigma}^2_{\text{sparse}}\right)\ ||\ \mathcal{P}\right)}{m}}\right)$$

where $\boldsymbol{\sigma}_{\text{sparse}} = \{\sigma_1, \ldots, \sigma_K\}$ is defined by $\sigma_k := \epsilon_k \cdot \frac{\xi_k}{4\sqrt{2d_{\text{eff}} + \log\left(2(K+1)\sqrt{m}\right)}}$, and where $d_{\text{eff}} := \max_{k \in [K]} \frac{(d_k - s_k)\log(d_k) + (d_{k-1} - s_{k-1})\log(d_{k-1})}{2}$ is an effective layer width[10].

The notation $\tilde{\mathcal{O}}$ above hides logarithmic factors (see Appendix A.3 for a complete version of the bound). This result bounds the generalization error of a trained predictor as a function of three terms. Besides the empirical risk with margin threshold $4\zeta_{K+1}\gamma_{K+1}$, the risk is upper bounded by an empirical sparse loss that measures the proportion of samples (in the training data) that do not achieve a sufficiently large sparse radius at any layer. Lastly, as is characteristic in PAC-Bayes bounds, we see a term that depends on the distance between the prior and posterior distributions, the latter centered at the obtained (data-dependent) predictor. The posterior variance $\boldsymbol{\sigma}^2_{\text{sparse}}$ is determined entirely by the base hyper-parameters. Finally, note that the result above holds for any prior distribution $\mathcal{P}$. Before moving on, we comment on the specific factors influencing this bound.

10. We note the effective width is at worst $\max_k d_k \log(d_k)$ and could be larger than actual width depending on the sparsity vector $\mathbf{s}$. In contrast, for large $\mathbf{s}$, $d_{\text{eff}} \ll \max_k d_k$.

Sparsity. The result above depends on the sparsity by the choice of the parameter $\mathbf{s}$. One can always instantiate the above result for $\mathbf{s} = \mathbf{0}$, corresponding to a global sensitivity analysis. At this trivial choice, the sparsity loss vanishes (because the sparse radius is infinite) and the bound is equivalent to an improved (derandomized) version of the results by Neyshabur et al. (2018). The formulation in Theorem 11 enables a continuum of choices (via hyper-parameters) suited to the trained predictor and sample data. A larger degree of sparsity at every layer results in a tighter bound since the upper bounds to the sensitivity of the predictor is reduced (as only reduced matrices are involved in its computation). In turn, this reduced sensitivity leads to a lower empirical margin risk by way of a lower threshold $4\zeta_{K+1}\gamma_{K+1}$. Furthermore, the effective width – determining the scale of posterior – is at worst $\max_k d_k \log(d_k)$ (for $\mathbf{s} = 0$) , but for large $\mathbf{s}$, $d_{\text{eff}} \ll \max_k d_k$.

Sensitivity. Standard sensitivity-based generalization bounds generally depend directly on the global Lipschitz constant that scales as $\mathcal{O}(\prod_{k=1}^{K} \|\mathbf{W}_k\|_2)$. For even moderate-size models, such dependence can render the bounds vacuous. Further recent studies suggest that the layer norms can even increase with the size of the training sets showing that, even for under-parameterized models, generalization bounds may be vacuous (Nagarajan and Kolter, 2019a). Our generalization bound does *not* scale with the reduced Lipschitz constant ζ_{K+1}: while larger (reduced) Lipschitz constants can render the empirical sparse loss closer to its maximum value of 1, the bound remains controlled due to our choice of modelling *relative* perturbations of model parameters.

Dependence On Depth. Unlike recent results (Bartlett et al., 2017; Neyshabur et al., 2015, 2018, 2019), our bound is not exponential with depth. However, the sensitivity bounds ζ_k and radius thresholds γ_k are themselves exponential in depth. While the empirical risk and sparse loss terms in the generalization bounds depend on ζ_k, γ_k, they are bounded in $[0, 1]$. In turn, by choosing the prior to be a Gaussian $P = \mathcal{N}(h_{\text{prior}}, \boldsymbol{\sigma}^2_{\text{sparse}})$, the KL-divergence term can be decomposed into layer-wise contributions, $\text{KL}\left(\mathcal{N}\left(h, \boldsymbol{\sigma}^2_{\text{sparse}}\right) \,||\, \mathcal{N}(h_{\text{prior}}, \boldsymbol{\sigma}^2_{\text{sparse}})\right) = \sum_{k=1}^{K+1} \frac{\|\mathbf{W}_k - \mathbf{W}_{\text{prior},k}\|_F^2}{2\sigma_k^2}$. Hence, the KL divergence term does not scale with the product of the relative perturbations (like γ_k) or the product of layer norms (like ζ_k).

Comparison To Related Work. Besides the relation to some of the works that have been mentioned previously, our contribution is most closely related to those approaches that employ different notions of reduced effective models in developing generalization bounds. Arora et al. (2018) do this via a *compression* argument, alas the resulting bound holds for the compressed network and not the original one. Neyshabur et al. (2017) develops PAC-Bayes bounds that clearly reflect the importance of *flatness*, which in our terms refers to the loss effective sensitivity of the obtained predictor. Similar in spirit to our results, Nagarajan and Kolter (2019b) capture a notion of reduced active size of the model and presenting their derandomized PAC-Bayes bound (which we centrally employ here). While avoiding exponential dependence on depth, their result depends inversely with the minimum absolute pre-activation level at each layer, which can be arbitrarily small (and thus, the bound becomes arbitrarily large). Our analysis, as represented by Lemma 4, circumvents this limitation. Our constructions on normalized sparse radius have close connections with the *normalized margins* from Wei and Ma (2020), and our use of augmented loss function (such as our *sparse loss*) resemble the ones proposed in Wei and Ma (2019). Most recently, Galanti et al. (2023) analyze the complexity

of compositionally sparse networks, however the sparsity stems from the convolutional nature of the filters rather than as a data-dependent (and sample dependent) property.

3.3. Hyper-Parameter Search

For any fixed predictor h, there can be multiple choices of $\mathbf{s}, \boldsymbol{\xi}, \boldsymbol{\eta}$ such that h is in the corresponding hypothesis class. In the following, we discuss strategies to search for suitable hyper-parameters that can provide tighter generalization bounds. To do so, one can instantiate a grid of candidate values for each hyper-parameter that is independent of data. Let the grid sizes be $(T_{\mathbf{s}}, T_{\boldsymbol{\xi}}, T_{\boldsymbol{\eta}}, T_{\boldsymbol{\epsilon}})$, respectively. We then instantiate the generalization bound in Theorem 11 for each choice of hyper-parameters in the cartesian product of grids with a reduced failure probability $\delta_{\text{red}} = \frac{\delta}{T_{\mathbf{s}} T_{\boldsymbol{\xi}} T_{\boldsymbol{\eta}} T_{\boldsymbol{\epsilon}}}$. By a simple union-bound argument, all these bounds hold simultaneously with probability $(1-\delta)$. In this way, for a fixed δ, the statistical cost above is $\sqrt{\log(T_{\mathbf{s}} T_{\boldsymbol{\xi}} T_{\boldsymbol{\eta}} T_{\boldsymbol{\epsilon}})}$ as the failure probability dependence in Theorem 11 is $\sqrt{\log\left(\frac{1}{\delta_{\text{red}}}\right)}$. The computational cost of a naïve search is $\mathcal{O}(T_{\mathbf{s}} T_{\boldsymbol{\xi}} T_{\boldsymbol{\eta}} T_{\boldsymbol{\epsilon}})$. In particular, for multilayer networks, to exhaustively search for a sparsity vector requires a grid of size $T_{\mathbf{s}} := \prod_{k=1}^{K} d_k$ rendering the search infeasible. Nonetheless, we shall soon show that by employing a greedy algorithm one can still obtain tighter generalization bounds with significantly lesser computational cost. Moreover, these hyper-parameters are not independent, and so we briefly describe here how this optimization can be performed with manageable complexity.

Norm Hyper-Parameters $(\boldsymbol{\xi}, \boldsymbol{\eta})$: One can choose $(\boldsymbol{\xi}, \boldsymbol{\eta})$ from a grid (fixed in advance) of candidate values, to closely match the true properties of the predictor. For networks with zero bias, w.l.o.g. one can normalize each layer weight $\mathbf{W}_k \to \tilde{\mathbf{W}}_k := \frac{1}{\|\mathbf{W}_k\|_{(d_k-1, s_{k-1})}} \mathbf{W}_k$ to ensure that $\left\|\tilde{\mathbf{W}}_k\right\|_{(d^k-1, s_{k-1})} = 1$ without changing the prediction[11]. The predicted labels, babel function, sparse local radius , margin and the generalization bound in Theorem 11 are all invariant to such a scaling. For the normalized network we can simply let $\xi_k := 1$ for all k. Fixing $\boldsymbol{\xi}$ this way results in no statistical or computational cost (beyond normalization). For discretizing $\boldsymbol{\eta}$, we can leverage the fact that for all (s_k, s_{k-1}), the reduced babel function is always less than $d_k - s_k - 1$ – since the inner products are scaled by the square of the sparse norms. Thus, we can construct a grid in $[0, 1]$ with T_η elements, which can be searched efficiently (see Appendix B for further details).

Sparsity Parameter $\mathbf{s}$: The sparsity vector $\mathbf{s}$ determines the degree of structure at which we evaluate the generalization of a fixed predictor. For a fixed predictor and relative sensitivity vector $\boldsymbol{\epsilon}$, a good choice of $\mathbf{s}$ is one that has sufficiently large sparse local radii on the training sample resulting in small average sparse loss, $\frac{1}{m} \sum_{\mathbf{x}^{(i)} \in S_T} \ell_{\text{sparse}}(h, \mathbf{x}^{(i)})$. At the trivial choice of sparsity $\mathbf{s} = \mathbf{0}$, for any choice of $\boldsymbol{\epsilon}$, the above loss is exactly zero. In general, at a fixed $\boldsymbol{\epsilon}$, this loss increases with larger (entrywise) $\mathbf{s}$. At the same time, the empirical margin loss term $\hat{R}_{4\zeta_{K+1}\gamma_{K+1}}(h)$ decreases with increasing $\mathbf{s}$ (since ζ_{K+1} grows). This reflects an inherent tradeoff in the choice of $(\mathbf{s}, \boldsymbol{\epsilon})$ to balance the margin loss and the sparse loss (in addition to the KL-divergence).

For any $\boldsymbol{\epsilon}$ and a data point $\mathbf{z} = (\mathbf{x}, y)$, we employ a greedy algorithm to find a sparsity vector $s^*(\mathbf{x}, \boldsymbol{\epsilon})$ in a layer wise fashion such that the loss incurred is zero, i.e. so that $r_k(h, \mathbf{x}) \geq 3\gamma_k$ for

11. This is not true for networks with non-zero bias. In networks with bias, one can still employ a grid search like in Bartlett et al. (2017).

all k. At each layer, we simply take the maximum sparsity level with sufficiently large radius. The computational cost of such an approach is $\log_2\left(\prod_{k=1}^{K} d_k\right)$. One can thus collect the sparsity vectors $s^*(\mathbf{x}, \boldsymbol{\epsilon})$ across the training set and choose the one with sample-wise minimum, so that the average sparse loss vanishes. Of course, one does not necessarily need the sparse loss to vanish; one can instead choose $\mathbf{s}$ simply to *control* the sparse loss to a level of $\frac{\alpha}{\sqrt{m}}$. We expand in Appendix B how this can done.

Sensitivity Vector $\boldsymbol{\epsilon}$: Lastly, the relative sensitivity vector $\boldsymbol{\epsilon}$ represents the size of the posterior and desired level of sensitivity in layer outputs upon parameter perturbations. Since ϵ_k denotes *relative perturbation* we can simply let it be the same across all layers. i.e. $\boldsymbol{\epsilon} = \epsilon \cdot [1, \ldots, 1]$.

In summary, as we expand in Appendix B, we can compute a best in-grid generalization bound in $\mathcal{O}\left(T_{\boldsymbol{\epsilon}} \cdot \log_2\left(\prod_{k=1}^{K} d_k\right) \cdot \log_2(T_{\boldsymbol{\eta}}) \cdot (\sum_{k=1}^{K} d_k d_{k-1})\right)$.

4. Numerical Experiments

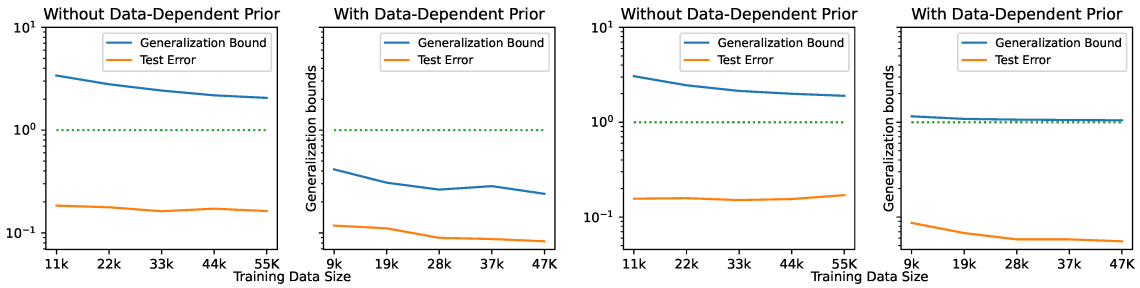
In this last section we intend to demonstrate the derived bounds on a series of feed-forward networks, of varying width and depth, on MNIST. As we now show, the resulting bounds are controlled and sometimes non-vacuous upon the optimization over a discrete grid for hyper-parameters, as explained above.

Experimental Setup: We train feed-forward networks h with weights $\{\mathbf{W}_k\}_{k=1}^{K+1}$ where $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ using the cross-entropy loss with stochastic gradient descent (SGD) for 5,000 steps with a batch size of 100 and learning rate of 0.01. The MNIST training set is randomly split into train and validation data (55,000 : 5,000). The models are optimized on the training data and the resulting measures are computed on validation data. To evaluate scaling with the number of samples, m, we train networks on randomly sampled subsets of the training data of increasing sizes from 20% to 100% of the training set. Because of the chosen architectures, all of these models are over-parametrized (i.e. having more parameters than training samples).

Recall that the bound on generalization error in Theorem 11 depends on the KL divergence between a posterior centered at trained predictor h, $\mathcal{N}(h, \boldsymbol{\sigma}^2_{\text{sparse}})$, and the prior $P = \mathcal{N}(h_{\text{prior}}, \boldsymbol{\sigma}^2_{\text{sparse}})$. Thus, each model is encouraged to be close to its initialization via a regularization term. In this way, we minimize the following regularized empirical risk based on the cross-entropy loss as well as a regularization term with penalty λ (set as $\lambda = 1.0$ for all experiments for simplicity),

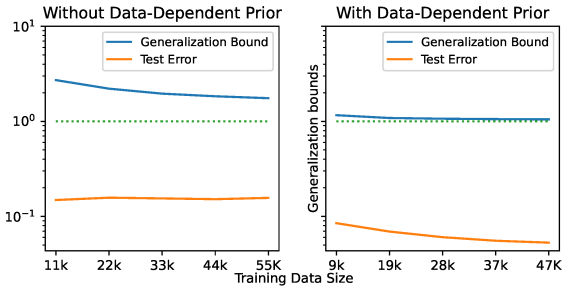
$$\min_{\{\mathbf{W}_k\}_{k=1}^{K+1}} \frac{1}{m}\sum_{i=1}^{m} \ell_{\text{cross-ent}}\Big(h, (\mathbf{x}_i, y_i)\Big) + \frac{\lambda}{K+1}\sum_{k=1}^{K+1} \|\mathbf{W}_k - \mathbf{W}_{\text{prior},k}\|_F^2 \,.$$

Choice Of Prior: As with any PAC-Bayes bound, choosing a prior distribution with an appropriate inductive bias is important. For example, optimizing the choice of prior by instantiating multiple priors simultaneously was shown to be an effective procedure to obtain good generalization bounds (Langford and Caruana, 2001; Dziugaite and Roy, 2017). In this work, we evaluate our bounds for two choices of the prior: *a)* a data-independent prior, $P_0 := \mathcal{N}(h_{\mathbf{0}}, \boldsymbol{\sigma}^2_{\text{sparse}})$ centered at a model with zero weights, $h_{\mathbf{0}}$; and *b)* a data-dependent prior $P_{\text{data}} := \mathcal{N}(h_{\text{init}}, \boldsymbol{\sigma}^2_{\text{sparse}})$ centered at a model



(a) Model width: 100.

(b) Model width: 500.



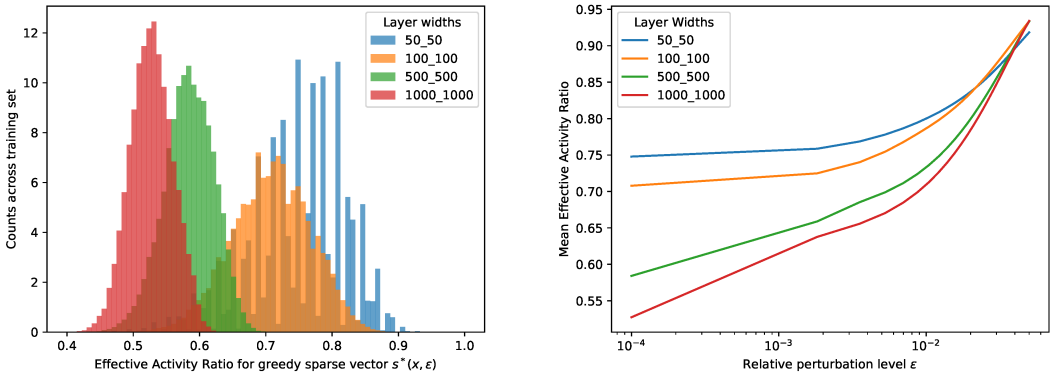
(c) Model width: 1,000.

Figure 1: Generalization error of a 2-layer model of different widths trained on MNIST.

h_{init} obtained by training on a small fraction of the training data (5% of all training data). Note that this choice is valid, as the base hyper-parameter $(\mathbf{s}, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\epsilon})$ are chosen independent of data, and the empirical risk terms in the bound are not evaluated on the small subset of data h_{init} is trained on.

Generalization Bounds Across Width: We first train a 2-layer (1 hidden layer) fully connected neural network with increasing widths, from 100 to 1,000 neurons. Note that in all cases these models are over-parametrized. In Figures 1a to 1c we plot the true risk (orange curve) and the generalization bounds (blue curve) from Theorem 11 across different sizes of training data and for the two choices of priors mentioned above. We observe that our analysis, when coupled with data-dependent prior P_{data}, generates non-vacuous bounds for a network with width of 100. Even for the naïve choice of the prior P_0, the bound is controlled and close to 1. Furthermore, note that our bounds remain controlled for larger widths. In Appendix E, we include complementary results depicting our generalization bounds for 3-layer networks.

Effective Activity Ratio: Lastly, we intend to illustrate the degree of sparsity achieved in the obtained models that allow for the bounds presented in Figure 1. For each data point $\mathbf{x}$ and relative perturbation level ϵ, we define the Effective Activity ratio $\kappa(\mathbf{x}, \epsilon) := \frac{\sum_k (d_k - s_k)(d_{k-1} - s_{k-1})}{\sum_k d_k d_{k-1}}$ where $\mathbf{s} = s^*(\mathbf{x}, \epsilon)$, the greedy sparsity vector chosen such that the sparse loss in Theorem 11 is zero. In this way, $\kappa(\mathbf{x}, \epsilon)$ measures the reduced local dimensionality of the model at input $\mathbf{x}$ under perturbations of relative size ϵ. When $\kappa(\mathbf{x}, \epsilon) = 1$, there are no sparse activation patterns that are stable under perturbations, and the full model is considered at that point. On the other hand, when $0 < \kappa(\mathbf{x}, \epsilon) \ll 1$, the size of stable sparse activation patterns $s^*(\mathbf{x}, \epsilon)_k$ at each layer is close to



(a) Histogram of Effective Activity Ratio at $\epsilon = 10^{-4}$ (b) Average Effective Activity Ratio

Figure 2: Effective activity ratio $\kappa(\mathbf{x}, \epsilon)$ based on greedy sparsity vector $s^*(\mathbf{x}, \epsilon)$ for 3-layer networks (smaller implies sparser stable activations).

the layer dimension d_k. Theorem 11 enables a theory of generalization that accounts for this local reduced dimensionality.

We present the effective activity rations for a trained 3-layer model in Figure 2, and include the corresponding results for the 2-layer model in Appendix E for completeness. The central observation from these results is that trained networks with larger width have *smaller* effective activity ratios across the training data. In Figure 2a (as well as in Figure 5a for the 2-layer model), the distribution of effective activity ratio across the training data at $\epsilon = 10^{-4}$ shows that smaller width networks have less stable sparsity. In turn, Figure 2b and Figure 5b demonstrate that this effect is stronger for smaller relative perturbation levels. This observation is likely the central reason of why our generalization bounds do not increase drastically with model size.

5. Conclusion

This work makes explicit use of the degree of sparsity that is achieved by ReLU feed-forward networks, reflecting the level of structure present in data-driven models, but without making any strong distributional assumptions on the data. Sparse activations imply that only a subset of the network is active at a given point. By studying the stability of these local sub-networks, and employing tools of derandomized PAC-Bayes analysis, we are able to provide bounds that exploit this effective reduced dimensionality of the predictors, as well as avoiding exponential dependence on the sensitivity of the function and of depth. Our empirical validation on MNIST illustrates our results, which are always controlled and sometimes result in non-vacuous bounds on the test error. Note that our strategy to instantiate our bound for practical models relied on a discretization of the space of hyper-parameters and a greedy selection of these values. This is likely suboptimal, and the grid of hyper-parameters could be further tuned for each model. Moreover, in light of the works in (Dziugaite and Roy, 2017, 2018; Zhou et al., 2019), we envision optimizing our bounds directly, leading to even tighter solutions.

Acknowledgments

We kindly thank Vaishnavh Nagarajan for helpful conversations that motivated the use of de-randomized PAC-Bayesian analysis. This work was supported by NSF grant CCF 2007649.

References

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *ArXiv*, abs/2110.11216, 2021.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *ICML*, 2018.

Arindam Banerjee, Tiancong Chen, and Yingxue Zhou. De-randomized pac-bayes margin bounds: Applications to non-convex and non-smooth predictors. *ArXiv*, abs/2002.09956, 2020.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20: 63:1–63:17, 2019.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *ArXiv*, abs/1703.11008, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Entropy-sgd optimizes the prior of a pac-bayes bound: Data-dependent pac-bayes priors via differential privacy. In *NeurIPS*, 2018.

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

Tomer Galanti, Mengjia Xu, Liane Galanti, and Tomaso Poggio. Norm-based generalization bounds for compositionally sparse neural networks. *arXiv preprint arXiv:2301.12033*, 2023.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019. URL https://arxiv.org/abs/1901.05353.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Yiding Jiang*, Behnam Neyshabur*, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

John Langford and Rich Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems*, 14, 2001.

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis D. Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *ArXiv*, abs/1806.05159, 2018.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

Ramchandran Muthukumar and Jeremias Sulam. Adversarial robustness of sparse local lipschitz predictors, 2022. URL https://arxiv.org/abs/2202.13216.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019a.

Vaishnavh Nagarajan and Zico Kolter. Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=Hygn2o0qKX.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *NIPS*, 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BygfghAcYX.

Guillermo Valle Pérez and Ard A. Louis. Generalization bounds for deep learning. *ArXiv*, abs/2012.04115, 2020.

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

John Shawe-Taylor and Robert C Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

Jeremias Sulam, Ramchandran Muthukumar, and Raman Arora. Adversarial robustness of supervised sparse coding. *Advances in neural information processing systems*, 33:2110–2121, 2020.

Joel A Tropp, Anna C Gilbert, Sambavi Muthukrishnan, and Martin J Strauss. Improved sparse approximation over quasiincoherent dictionaries. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–37. IEEE, 2003.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *NeurIPS*, 2019.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *ICLR*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

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 *ICLR*, 2019.

Appendix A. Missing Proofs

A.1. Sparsity In Feed-Forward Maps

In this subsection we provide explicit proofs for all theorems corresponding to stability of index sets and reduced size (such as Lemma 4 and proposition 9) and sensitivity of outputs (such as Proposition 10)

A.1.1. Stability Of Index Sets In A Single Layer Feed-Forward Map

Proof (For Lemma 4) To prove the first statement, note that for all $i \in I(\mathbf{W}, \mathbf{x}, s_1)$,

$$\max\left\{0, -\frac{\mathbf{w}[i]\cdot\mathbf{x}}{\xi_1\cdot\zeta_0}\right\} \geq r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) \tag{5}$$

Hence if $r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) > 0$, then $I(\mathbf{W}, \mathbf{x}, s_1)$ is inactive. Now consider any perturbed weight $\hat{\mathbf{W}}$ and any perturbed input $\hat{\mathbf{x}}$ such that $\|\hat{\mathbf{x}} - \mathbf{x}\| \leq d_0 - s_0$. The absolute difference between the normalized pre-activation values at each index can be bounded,

$$\begin{aligned}
&|\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}} - \mathbf{w}[i]\cdot\mathbf{x}| \\
&= |\ \mathbf{w}[i]\cdot(\hat{\mathbf{x}}-\mathbf{x}) + (\hat{\mathbf{w}}[i]-\mathbf{w}[i])\cdot\mathbf{x}\ + (\hat{\mathbf{w}}[i]-\mathbf{w}[i])\cdot(\hat{\mathbf{x}}-\mathbf{x})\ | \\
&\leq \max_{|J_0|=d_0-s_0} \left(\|\mathcal{P}_{J_0}(\mathbf{w}[i])\|_2\cdot\|\hat{\mathbf{x}}-\mathbf{x}\|_2 + \|\mathcal{P}_{J_0}\left(\hat{\mathbf{w}}[i]-\mathbf{w}[i]\right)\|_2\,\|\mathbf{x}\|_2 + \|\mathcal{P}_{J_0}\left(\hat{\mathbf{w}}[i]-\mathbf{w}[i]\right)\|_2\,\|\hat{\mathbf{x}}-\mathbf{x}\|_2\right). \\
&\leq \xi_1\cdot\|\hat{\mathbf{x}}-\mathbf{x}\|_2 + \left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}\cdot\zeta_0 + \left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}\cdot\|\hat{\mathbf{x}}-\mathbf{x}\|_2 \\
&= \xi_1\cdot\zeta_0\cdot\left(-1+\left(1+\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1+\frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right)\right)
\end{aligned}$$

The above inequalities show that,

$$\left|\frac{\mathbf{w}[i]\cdot\mathbf{x}-\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\cdot\zeta_0}\right| \leq -1+\left(1+\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1+\frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right) \tag{6}$$

This proves the second statement by plugging in the bounds on the relative perturbation terms above and using Equation (5) to note that,

$$\begin{aligned}
r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) &> \left(-1+\left(1+\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1+\frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right)\right) \\
\implies \forall\, i \in I(\mathbf{W}, \mathbf{x}, s_1),\quad \frac{-\mathbf{w}[i]\cdot\mathbf{x}}{\xi_1\zeta_0} &> \left(-1+\left(1+\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1+\frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right)\right) \\
\implies \forall\, i \in I(\mathbf{W}, \mathbf{x}, s_1),\quad \frac{-\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0} &> 0
\end{aligned}$$

Thus when the sparse local radius is large as stated, the index set $I(\mathbf{W}, \mathbf{x}, s_1)$ is inactive for $\hat{\Phi}(\hat{\mathbf{x}})$. As a special case, when $\epsilon = 0$, the same logic implies that the index set is inactive for $\Phi(\hat{\mathbf{x}})$. It is left to prove the stability of the sparse local radii.

For the final statement, recall the definitions of sparse local radius at $(\mathbf{W}, \mathbf{x})$ and $(\hat{\mathbf{W}}, \hat{\mathbf{x}})$ respectively

$$r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) := \sigma\left(\text{SORT}\left(-\left[\frac{\mathbf{w}[i]\cdot\mathbf{x}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right)\right),$$
$$r_{\text{sparse}}(\hat{\mathbf{W}}, \hat{\mathbf{x}}, s_1) := \sigma\left(\text{SORT}\left(-\left[\frac{\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right)\right).$$

Since ReLU is 1-Lipschitz, hence the distance between the radius measurements can be bounded as,

$$\begin{aligned}
&|r_{\text{sparse}}(\hat{\mathbf{W}}, \hat{\mathbf{x}}, s_1) - r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1)| \\
&\leq \left|\text{SORT}\left(-\left[\frac{\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right) - \text{SORT}\left(-\left[\frac{\mathbf{w}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right)\right|
\end{aligned}$$

Then observe that,

$$\begin{aligned}
&\text{SORT}\left(-\left[\frac{\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right) \\
&= \max_{\hat{I}\subset[d_1],|\hat{I}|=s_1} \min_{i\in\hat{I}} \frac{-\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0} \\
&\geq \min_{i\in I(\mathbf{W},\mathbf{x},s_1)} \frac{-\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0} \\
&\geq \min_{i\in I(\mathbf{W},\mathbf{x},s_1)} \frac{-\mathbf{w}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0} - \left(-1 + \left(1 + \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1 + \frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right)\right) \quad \text{by Equation (6).} \\
&= \text{SORT}\left(-\left[\frac{\mathbf{w}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right) - \left(-1 + \left(1 + \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1 + \frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right)\right)
\end{aligned}$$

By repeating the same arguments, one can establish that,

$$\begin{aligned}
&\left|\text{SORT}\left(-\left[\frac{\hat{\mathbf{w}}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right) - \text{SORT}\left(-\left[\frac{\mathbf{w}[i]\cdot\hat{\mathbf{x}}}{\xi_1\zeta_0}\right]_{i=1}^{d_1}, s_1\right)\right| \\
&\leq \left(-1 + \left(1 + \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\zeta_0}\right)\left(1 + \frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|_{(d_1-1,s_0)}}{\xi_1}\right)\right)
\end{aligned}$$

Hence the difference between the sparse local radii are bounded as required.

Lastly, to show the reduced sensitivity of the predictor, notice the following. Let I_0 be an inactive index set in the input of size s_0 and let $J_0 := (I_0)^c$ be its complement. When $r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) > 0$, the index set $I(\mathbf{W}, \mathbf{x}, s_1)$ is inactive. Let $J_1 := (I(\mathbf{W}, \mathbf{x}, s_1))^c$ be its complement index set. Then,

$$\Phi(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x}) = \sigma\left(\mathcal{P}_{J_1,J_0}(\mathbf{W})\mathcal{P}_{J_0}(\mathbf{x})\right)$$

Hence, $\|\Phi(\mathbf{x})\|_2 \leq \|\mathcal{P}_{J_1,J_0}(\mathbf{W})\|_2 \|\mathbf{x}\|_2 \leq \|\mathbf{W}\|_{(s_1,s_0)} \zeta_0 \leq \xi_1\sqrt{1+\eta_1}\zeta_0$. Thus proving the reduced size of the outputs. When $r_{\text{sparse}}(\mathbf{W}, \mathbf{x}, s_1) > -1 + (1+\epsilon_0)(1+\epsilon_1)$, for perturbed inputs and weights as described, the index set $I(\mathbf{W}, \mathbf{x}, s_0)$ is inactive for $\Phi(\hat{\mathbf{x}})$ and $\hat{\Phi}(\hat{\mathbf{x}})$. Again let J_1, J_0 be the complement sets,

$$\begin{aligned}
\left\|\hat{\Phi}(\hat{\mathbf{x}}) - \Phi(\mathbf{x})\right\|_2 &= \left\|\sigma\left(\hat{\mathbf{W}}\hat{\mathbf{x}}\right) - \sigma(\mathbf{W}\mathbf{x})\right\|_2 \\
&= \left\|\sigma\left(\mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}})\mathcal{P}_{J_0}(\hat{\mathbf{x}})\right) - \sigma\left(\mathcal{P}_{J_1,J_0}(\mathbf{W})\mathcal{P}_{J_0}(\mathbf{x})\right)\right\|_2 \\
&\leq \left\|\mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}})\mathcal{P}_{J_0}(\hat{\mathbf{x}}) - \mathcal{P}_{J_1,J_0}(\mathbf{W})\mathcal{P}_{J_0}(\mathbf{x})\right\|_2 \\
&\leq \left\|\mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}}) \cdot \mathcal{P}_{J_0}(\hat{\mathbf{x}} - \mathbf{x}) + \mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}} - \mathbf{W}) \cdot \mathcal{P}_{J_0}(\mathbf{x})\right\|_2 \\
&\leq \left\|\mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}}) \cdot \mathcal{P}_{J_0}(\hat{\mathbf{x}} - \mathbf{x})\right\|_2 + \left\|\mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}} - \mathbf{W}) \cdot \mathcal{P}_{J_0}(\mathbf{x})\right\|_2 \\
&\leq (\|\mathcal{P}_{J_1,J_0}(\mathbf{W})\|_2 + \left\|\mathcal{P}_{J_1,J_0}(\hat{\mathbf{W}} - \mathbf{W})\right\|_2) \cdot \epsilon_0\zeta_0 + \epsilon_1\xi_1\sqrt{1+\eta_1} \cdot \zeta_0. \\
&\leq (\epsilon_0 + \epsilon_0\epsilon_1 + \epsilon_1) \cdot \xi_1\sqrt{1+\eta_1}\zeta_0 \\
&= (-1 + (1+\epsilon_0)(1+\epsilon_1)) \cdot \xi_1\sqrt{1+\eta_1}\zeta_0
\end{aligned}$$

■

A.1.2. Reduced Size Of Layer Outputs In Multilayer Networks

Consider the layer-wise input domains, $\mathcal{X}_k := \{\mathbf{t} \in \mathbb{R}^{d_k} \mid \|\mathbf{t}\|_2 \leq \zeta_k,\ \|\mathbf{t}\|_0 \leq d_k - s_k\}$.

Proof (For Proposition 9)

From Lemma 4, $r_1(h, \mathbf{x}_0) > 0$ guarantees existence of inactive index set $I_1(h, \mathbf{x}_0)$ and a reduced size of the output such that $\|\mathbf{x}_1\| \leq \mathsf{M}_{\mathcal{X}} \|\mathbf{W}\|_{(s_1,0)}$. From Lemma 18 and the definition of the hyper-parameters ξ_1 and η_1,

$$\|\mathbf{W}\|_{(s_1,0)} \leq \|\mathbf{W}\|_{d_1-1,0} \sqrt{1 + \mu_{s_1,0}(\mathbf{W})} \leq \xi_1\sqrt{1+\eta_1}.$$

Hence $\|\mathbf{x}\|_1 \leq \zeta_1$. Thus the statement of the theorem is true for $k = 1$.

Assume that the statement is true for all layers $1 \leq n \leq k$. Hence when $r_n(h, \mathbf{x}_0) > 0$ for all layers $1 \leq n \leq k$, there exists index sets $I_1(h, \mathbf{x}_0), \ldots, I_k(h, \mathbf{x}_0)$ such that $\mathcal{P}_{I_n(h,\mathbf{x}_0)}(\mathbf{x}_n) = \mathbf{0}$ and $\|\mathbf{x}_n\| \leq \zeta_n$ for all $1 \leq n \leq k$. Thus $\mathbf{x}_n \in \mathcal{X}_n$ for all $1 \leq n \leq k$.

If additionally $r_{k+1}(h, \mathbf{x}_0) > 0$, then by invoking Lemma 4 for input $\mathbf{x}_k \in \mathcal{X}_k$ and weight $\mathbf{W}_{k+1} \in \mathcal{W}_{k+1}$, we see that $I_{k+1}(h, \mathbf{x}_0)$ is inactive for $\mathbf{x}_{k+1}$ and further Lemma 4 shows that $\|\mathbf{x}_{k+1}\| \leq \xi_{k+1}\sqrt{1+\eta_{k+1}} \cdot \zeta_k = \zeta_{k+1}$ as desired. Hence the theorem is true for all $1 \leq k \leq K$.

For the final layer we note that since $I_K(h, \mathbf{x}_0)$ of size s_K is inactive for $\mathbf{x}_K$,

$$\|h(\mathbf{x}_0)\|_\infty \leq \left\|\mathcal{P}_{[C],J_K}(\mathbf{W}_{K+1})\right\|_{2\to\infty} \|\mathbf{x}_K\|_2 \leq \|\mathbf{W}_{K+1}\|_{C-1,s_K} \|\mathbf{x}_K\| \leq \xi_{K+1} \cdot \zeta_K = \zeta_{K+1}.$$

In the above inequality, we have used the fact that for any matrix the reduced $2 \to \infty$ norm,

$$\left\|\mathcal{P}_{[C],J_K}(\mathbf{W}_{K+1})\right\|_{2\to\infty} \leq \max_{|J|=d_K-s_K} \max_{j\in C} \|\mathcal{P}_J(\mathbf{w}_{K+1}[j])\|_2 = \|\mathbf{W}_{K+1}\|_{C-1,s_K}.$$

■

A.1.3. Reduced Sensitivity Of Layer Outputs In Multilayer Networks

Proof (For Proposition 10) From Lemma 4, $r_1(h,\mathbf{x}_0) > \gamma_1$ guarantees existence of inactive index set $I_1(h,\mathbf{x}_0)$ such that $\mathcal{P}_{I_1(h,\mathbf{x}_0)}(\hat{\mathbf{x}}_1) = \mathcal{P}_{I_1(h,\mathbf{x}_0)}(\mathbf{x}_1) = \mathbf{0}$. Further from Lemma 4 (with input perturbation $\epsilon_0 = 0$), the distance between the first layer representations are bounded as $\|\hat{\mathbf{x}}_1 - \mathbf{x}_1\| \leq (-1+(1+\epsilon_0)(1+\epsilon_1))\|\mathbf{W}\|_{(s_1,0)}\mathsf{M}_{\mathcal{X}} \leq \epsilon_1 \cdot \xi_1\sqrt{1+\eta_1}\zeta_0 = \zeta_1\gamma_1$. Thus the statement of the theorem is true for $k=1$.

Assume that the statement is true for all layers $1 \leq n \leq k$. Thus there exists index sets $I_1(h,\mathbf{x}_0),\ldots,I_k(h,\mathbf{x}_0)$ such that $\mathcal{P}_{I_n(h,\mathbf{x}_0)}(\hat{\mathbf{x}}_n) = \mathcal{P}_{I_n(h,\mathbf{x}_0)}(\mathbf{x}_n) = \mathbf{0}$ and the distance between the layer representations are bounded $\|\hat{\mathbf{x}}_n - \mathbf{x}_n\| \leq \zeta_n \cdot \gamma_n$ for all $1 \leq n \leq k$.

From Proposition 9, due to the reduced size, $\mathbf{x}_k \in \mathcal{X}_k$ and the sparse local radius $r_{k+1}(h,\mathbf{x}_0) \in [0,1]$. For the perturbed input to layer $k+1$, $\hat{\mathbf{x}}_k$ we note that $\|\hat{\mathbf{x}}_k - \mathbf{x}_k\|_2 \leq \zeta_k \cdot \gamma_k$ and $\|\hat{\mathbf{x}}_k - \mathbf{x}_k\|_0 \leq d_k - s_k$. The perturbed weight $\hat{\mathbf{W}}_{k+1}$ is such that $\left\|\hat{\mathbf{W}}_{k+1} - \mathbf{W}_{k+1}\right\|_{d_{k+1}-1,s_k} \leq \epsilon_{k+1}\xi_{k+1}$. Hence applying Lemma 4 on inputs $\mathbf{x}_k \in \mathcal{X}_k$ and weight $\hat{\mathbf{W}}_{k+1} \in \mathcal{W}_{k+1}$ shows that the set $I_{k+1}(h,\mathbf{x}_0)$ is inactive for $\hat{\mathbf{x}}_{k+1}$ and further,

$$\|\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}_k\|_2 \leq \underbrace{(-1+(1+\gamma_k)(1+\epsilon_{k+1}))}_{=:\gamma_{k+1}} \underbrace{\xi_{k+1}\sqrt{1+\eta_{k+1}}\cdot\zeta_k}_{=:\zeta_{k+1}}.$$

Hence the conclusion follows for all layers $1 \leq k \leq K$. For the last layer, under the assumption that $r_k(h,\mathbf{x}_0) > \gamma_k$ for all k, we know that $I_K(h,\mathbf{x}_0)$ of size s_K is inactive for both $\mathbf{x}_K$ and $\hat{\mathbf{x}}_K$. Let J_K be the complement of $I_K(h,\mathbf{x}_0)$. We can bound the distance between the network outputs as follows,

$$\begin{aligned}
\left\|\hat{h}(\mathbf{x}_0) - h(\mathbf{x}_0)\right\|_\infty &= \left\|\hat{\mathbf{W}}_{K+1}\hat{\mathbf{x}}_K - \mathbf{W}_{K+1}\mathbf{x}_K\right\|_\infty \\
&= \left\|\mathcal{P}_{[C],J_K}(\hat{\mathbf{W}}_{K+1})\mathcal{P}_{J_K}(\hat{\mathbf{x}}_K) - \mathcal{P}_{[C],J_K}(\mathbf{W}_{K+1})\mathcal{P}_{J_K}(\mathbf{x}_K)\right\|_\infty \\
&\leq \left\|\mathcal{P}_{[C],J_K}(\hat{\mathbf{W}})\cdot\mathcal{P}_{J_K}(\hat{\mathbf{x}}_K - \mathbf{x}_K) + \mathcal{P}_{[C],J_K}(\hat{\mathbf{W}} - \mathbf{W})\cdot\mathcal{P}_{J_K}(\mathbf{x}_K)\right\|_\infty \\
&\leq \left\|\mathcal{P}_{[C],J_K}(\hat{\mathbf{W}})\cdot\mathcal{P}_{J_K}(\hat{\mathbf{x}}_K - \mathbf{x}_K)\right\|_\infty + \left\|\mathcal{P}_{[C],J_K}(\hat{\mathbf{W}} - \mathbf{W})\cdot\mathcal{P}_{J_K}(\mathbf{x}_K)\right\|_\infty \\
&\leq \left(\left\|\mathcal{P}_{[C],J_K}(\mathbf{W})\right\|_{2\to\infty} + \left\|\mathcal{P}_{[C],J_K}(\hat{\mathbf{W}} - \mathbf{W})\right\|_{2\to\infty}\right)\cdot\|\mathcal{P}_{J_K}(\hat{\mathbf{x}} - \mathbf{x})\|_2 \\
&\quad + \left\|\mathcal{P}_{[C],J_K}(\hat{\mathbf{W}} - \mathbf{W})\right\|_{2\to\infty}\cdot\|\mathcal{P}_{J_K}(\mathbf{x}_K)\|_2 \\
&\leq (1+\epsilon_{K+1})\underbrace{\xi_{K+1}\cdot\zeta_K}_{=\zeta_{K+1}}\gamma_K + \epsilon_{K+1}\underbrace{\xi_{K+1}\zeta_K}_{=\zeta_{K+1}} \\
&= \zeta_{K+1}\cdot\underbrace{((1+\epsilon_{K+1})\gamma_K + \epsilon_{K+1})}_{=\gamma_{K+1}} \\
&= \zeta_{K+1}\cdot\gamma_{K+1}.
\end{aligned}$$

■

A.2. Gaussian Sensitivity Analysis

In this subsection we seek to understand the probability that $\hat{h} \in \mathcal{B}(h, \boldsymbol{\epsilon})$ when the perturbed layer weights are randomly sampled from $\mathcal{N}(h, \boldsymbol{\sigma}^2)$. For any failure probability $\delta > 0$, we define layer-wise normalization functions $\alpha^k, \beta^k : [0,1] \to \mathbb{R}^{>0}$ that are dimension-dependent (but data/weights independent),

$$\alpha^k(s_k, s_{k-1}, \delta) := \left(\sqrt{d_k - s_k} + \sqrt{d_{k-1} - s_{k-1}}) + \sqrt{2 \log \binom{d_k}{s_k} + 2 \log \binom{d_{k-1}}{s_{k-1}} + 2 \log \left(\frac{1}{\delta}\right)} \right) \tag{7}$$

We can now bound the probability that $\hat{h} \in \mathcal{B}(h, \mathbf{s}, \boldsymbol{\epsilon})$ when constructed using Gaussian perturbations.

Lemma 12 *Define layer-wise variance parameter $\boldsymbol{\sigma}(\delta)$ as*

$$\forall\, 1 \le k \le K, \;\; \sigma^k(\delta) := \epsilon^k \min \left\{ \frac{\xi^k \sqrt{1+\eta^k}}{\alpha^k(s_k, s_{k-1}, \frac{\delta}{K+1})}, \frac{\xi^k}{\alpha^k(d_k - 1, s_{k-1}, \frac{\delta}{K+1})} \right\}.$$

and let $\sigma_{K+1}(\delta) := \epsilon_{K+1} \cdot \frac{\epsilon_{K+1}}{\alpha^k(C-1, s_K, \frac{\delta}{K+1})}$, For any $h \in \mathcal{H}$, with probability at least $(1-\delta)$, a Gaussian perturbed network sampled from $\mathcal{N}(h, \sigma^2(\delta))$ is in the local neighbourhood $\mathcal{B}(h, \boldsymbol{\epsilon})$.

Proof As per Lemma 20, with probability at least $(1-\delta)$, a perturbed network $\hat{h}$ sampled from $\mathcal{N}(h, \sigma^2(\delta))$ satisfies the following inequalities simultaneously at every layer,

$$\left\| \hat{\mathbf{W}}_k - \mathbf{W}_k \right\|_{(s_k, s_{k-1})} \le \sigma^k(\delta) \cdot \alpha^k(s_k, s_{k-1}, \delta), \quad \left\| \hat{\mathbf{W}}_k - \mathbf{W}_k \right\|_{(d_k - 1, s_{k-1})} \le \sigma^k(\delta) \cdot \alpha^k(d_k - 1, s_{k-1}, \delta).$$

Clearly by the choice of variance parameter this implies that $\hat{h} \in \mathcal{B}(h, \boldsymbol{\epsilon})$ with probabilty at least $(1-\delta)$ since,

$$\max \left\{ \frac{\left\| \hat{\mathbf{W}}_k - \mathbf{W}_k \right\|_{(s_k, s_{k-1})}}{\xi_k \sqrt{1+\eta_k}}, \frac{\left\| \hat{\mathbf{W}}_k - \mathbf{W}_k \right\|_{(d_k - 1, s_{k-1})}}{\xi_k} \right\} \le \epsilon_k.$$

■

A.3. Sparsity-Aware Generalization Theory: Expanded Result

In this section we prove a stronger version of the simplified result in Theorem 11.

Theorem 13 *Let $\mathcal{P} := \prod_{k=1}^{K} \mathcal{P}_k$ be any (factored) prior distribution over depth-$(K+1)$ feed-forward network chosen independently of data. Let $h \in \mathcal{H}$ be any feed-forward network (possibly*

trained on sample data). With probability at least $(1-\delta)$ *over the choice of i.i.d training sample* S_T *of size* m*, the generalization error of* h *is bounded as follows,*

$$R_0(h) \leq \hat{R}_{4\zeta_{K+1}\cdot\gamma_{K+1}}(h) + \frac{4(K+1)}{\sqrt{m}-1} + \sqrt{\frac{4\mathrm{KL}\left(\mathcal{N}\left(h, \boldsymbol{\sigma}^2_{\text{sparse}}\right) \;||\; P\right) + 2\log\left(\frac{2m(K+1)}{\delta}\right)}{m-1}}$$
$$+ \sum_{k\in[K]} \frac{1}{m} \sum_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\in S_T} \mathbf{1}\left\{\exists 1 \leq n \leq k,\; r_n(h,\mathbf{x}^{(i)}) < \gamma_n\right\}$$
$$+ \sum_{k\in[K]} \frac{1}{m} \sum_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\in S_T} \cdot\mathbf{1}\left\{\exists 1 \leq n \leq k,\; r_n(h,\mathbf{x}^{(i)}) < 3\gamma_n\right\}$$
$$+ \sum_{k\in[K]} \sqrt{\frac{4\cdot\sum_{n=1}^{k}\mathrm{KL}\left(\mathcal{N}\left(\mathbf{W}_n, \sigma_n^2\right) \;||\; \mathcal{P}_n\right) + 2\log\left(\frac{2m(K+1)}{\delta}\right)}{m-1}}$$

with layer-wise variance parameter $\boldsymbol{\sigma}_{\text{sparse}} = \{\sigma_1, \ldots, \sigma_K\}$ *is defined as,*

$$\sigma_k := \epsilon^k \min\left\{\frac{\xi^k\sqrt{1+\eta^k}}{\alpha^k(s_k, s_{k-1}, \frac{1}{(K+1)\sqrt{m}})}, \frac{\xi^k}{\alpha^k(d_k-1, s_{k-1}, \frac{1}{(K+1)\sqrt{m}})}\right\}. \tag{8}$$

Proof We note that $R_\gamma(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}(\ell_\gamma(h,\mathbf{z})) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathbf{1}\left\{\rho(h,\mathbf{z}) < \gamma\right\}\right]$. For the margin property $\rho(h,\mathbf{z}) := h(\mathbf{x})_y - \max_{j\neq y} h(\mathbf{x})_j$ with margin threshold γ, Lemma 16 shows that with probability at least $(1-\frac{\delta}{K+1})$ over the choice of i.i.d training sample S_T of size m, for any predictor $h\in\mathcal{H}$, the generalization error is bounded by

$$\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\rho(h,\mathbf{z}) < 0\right] \tag{9}$$
$$\leq \frac{1}{m}\sum_{\mathbf{z}^{(i)}\in S_T}\mathbf{1}\left[\rho(h,\mathbf{z}^{(i)}) < 4\zeta_{K+1}\gamma_{K+1}\right] + \frac{2}{\sqrt{m}-1} + +\sqrt{\frac{4\mathrm{KL}\left(\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)\;||\;\mathcal{P}\right) + 2\log(\frac{2m(K+1)}{\delta})}{2(m-1)}}.$$
$$+ \mu_{S_T}\left(h,(\rho, 4\zeta_{K+1}\gamma_{K+1})\right) + \mu_{\mathcal{D}}\left(h,(\rho,4\zeta_{K+1}\gamma_{K+1})\right) \tag{10}$$

It remains to bound the term $\mu_{\mathcal{D}}\left(h,(\rho,4\zeta_{K+1}\gamma_{K+1})\right)$. From Bartlett et al. (2017), the margin $\rho(\cdot,\cdot)$ is 2-Lipschitz w.r.t network outputs,

$$|\rho(\hat{h},\mathbf{z}) - \rho(h,\mathbf{z})| \leq 2\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty.$$

Hence we can reduce the noise-resilience over the margin to the event that variation in networks outputs is bounded,

$$\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[|\rho(\hat{h},\mathbf{z}) - \rho(h,\mathbf{z})| > 2\zeta_{K+1}\gamma_{K+1}\right]$$
$$\leq \underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty > \zeta_{K+1}\gamma_{K+1}\right]$$
$$\therefore \underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty > \zeta_{K+1}\gamma_{K+1}\right] \leq \frac{1}{\sqrt{m}}$$
$$\implies h \text{ is noise-resilient w.r.t } \rho \text{ at } \mathbf{z}.$$

Therefore h is not noise-resilient w.r.t ρ at $\mathbf{z} = (\mathbf{x}_0, y)$ implies

$$\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty > \zeta_{K+1}\gamma_{K+1}\right] > \frac{1}{\sqrt{m}}$$

Thus the probability (over inputs) that a predictor is not noise-resilient can be bounded using the event that the change in network output is large,

$$\begin{aligned}
\mu_{\mathcal{D}}\left(h, (\rho, \zeta_{K+1}\gamma_{K+1}\right) &= \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[|\rho(\hat{h},\mathbf{z}) - \rho(h,\mathbf{z})| > 2\zeta_{K+1}\gamma_{K+1}\right] > \frac{1}{\sqrt{m}}.\right] \\
&\leq \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty > \zeta_{K+1}\gamma_{K+1}\right] > \frac{1}{\sqrt{m}}.\right] \\
&\leq \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty \leq \zeta_{K+1}\gamma_{K+1}\right] < 1 - \frac{1}{\sqrt{m}}.\right]
\end{aligned} \tag{11}$$

We now make two observations that together helps us upper bound the above probability,

1. From Proposition 10, if the layer-wise sparse local radius at each layer k is sufficiently large,

   $$\forall\, 1 \leq k \leq K,\;\; r_k(h, \mathbf{x}_0) \geq \gamma_k$$

   then for any perturbed network $\hat{h}$ is in $\mathcal{B}(h, \boldsymbol{\epsilon})$ and the distance between the network-output,

   $$\left\|\hat{h}(\mathbf{x}_0) - h(\mathbf{x}_0)\right\|_\infty \leq \zeta_{K+1}\cdot\gamma_{K+1}$$

2. The choice of variance in the theorem statement, $\boldsymbol{\sigma}^2_{\text{sparse}} = \boldsymbol{\sigma}^2(\frac{1}{\sqrt{m}})$, the variance described in Lemma 12 for $\delta = \frac{1}{\sqrt{m}}$. Thus by Lemma 12, with probability at least $(1 - \frac{1}{\sqrt{m}})$ a randomly perturbed network $\hat{h} \sim \mathcal{N}\left(h, \boldsymbol{\sigma}^2\right)$ is within the neighbourhood $\mathcal{B}(h, \boldsymbol{\epsilon})$.

We can combine the above two observations to infer that at any input $\mathbf{z} = (\mathbf{x}, y) \sim \mathcal{D}$,

$$\begin{aligned}
&\forall\, 1 \leq k \leq K,\;\; r_k(h, \mathbf{x}_0) \geq \gamma_k \\
&\qquad\implies \underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty \leq \zeta_{K+1}\gamma_{K+1}\right] \geq 1 - \frac{1}{\sqrt{m}}.
\end{aligned}$$

Thus to ensure that the probability that $\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty \leq \zeta_{K+1}\gamma_{K+1}$ is less than $1 - \frac{1}{\sqrt{m}}$, one necessarily needs that the sparse local radius is insufficient at some layer k,

$$\begin{aligned}
&\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x}) - h(\mathbf{x})\right\|_\infty \leq \zeta_{K+1}\gamma_{K+1}\right] < 1 - \frac{1}{\sqrt{m}} \\
&\qquad\implies \exists\, 1 \leq k \leq K,\;\; r_k(h, \mathbf{x}_0) < \gamma_k
\end{aligned}$$

Plugging the above logic into Equation (11) we get a condition on the sparse local radius,

$$\begin{aligned}\mu_{\mathcal{D}}\left(h,(\rho,4\zeta_{K+1}\gamma_{K+1}\right) &= \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[|\rho(\hat{h},\mathbf{z})-\rho(h,\mathbf{z})|>2\zeta_{K+1}\gamma_{K+1}\right]>\frac{1}{\sqrt{m}}.\right]\\ &\leq \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[\left\|\hat{h}(\mathbf{x})-h(\mathbf{x})\right\|_{\infty}>\zeta_{K+1}\gamma_{K+1}\right]>\frac{1}{\sqrt{m}}.\right]\\ &\leq \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{x}_0)<\gamma_k\right)\end{aligned}$$

Similarly we can reduce the noise-resilience condition on training sample,

$$\begin{aligned}\mu_{\mathtt{S}_T}\left(h,(\rho,4\zeta_{K+1}\gamma_{K+1}\right) &= \underset{\mathbf{z}\sim\mathfrak{U}(\mathtt{S}_T)}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)}{\mathbf{Prob}}\left[|\rho(\hat{h},\mathbf{z})-\rho(h,\mathbf{z})|>2\zeta_{K+1}\gamma_{K+1}\right]>\frac{1}{\sqrt{m}}.\right]\\ &\leq \frac{1}{m}\sum_{\mathbf{z}^{(i)}\in\mathtt{S}_T}\mathbf{1}\left\{\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{x}_0)<\gamma_k\right\}\end{aligned}$$

To summarize we have the following generalization bound that holds with probability at least $(1-\frac{\delta}{K+1})$,

$$\begin{aligned}&\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\rho(h,\mathbf{z})<0\right]\\ &\quad\leq\frac{1}{m}\sum_{\mathbf{z}^{(i)}\in\mathtt{S}_T}\mathbf{1}\left[\rho(h,\mathbf{z}^{(i)})<4\zeta_{K+1}\gamma_{K+1}\right]+\frac{2}{\sqrt{m}-1}\\ &\quad+\frac{1}{m}\sum_{\mathbf{z}^{(i)}\in\mathtt{S}_T}\mathbf{1}\left\{\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{x}_0)<\gamma_k\right\}\\ &\quad+\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{x}_0)<\gamma_k\right)\\ &\quad+\sqrt{\frac{4\mathrm{KL}\left(\mathcal{N}\left(h,\boldsymbol{\sigma}^2_{\text{sparse}}\right)\ ||\ \mathcal{P}\right)+2\log\left(\frac{(K+1)m}{\delta}\right)}{m-1}}.\end{aligned}$$

We still need to bound the probability that the sparse local radii aren't sufficiently large,

$$\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{x}_0)<\gamma_k\right).$$

Consider the set of properties $\{r_k(h,\mathbf{x}_0)-\gamma_k\}_{k=1}^K$ and margin thresholds $\{2\gamma_k\}_{k=1}^K$. Lemma 16 shows that with probability at least $(1-\frac{\delta}{(K+1)})$ over the choice of i.i.d training sample $\mathtt{S}_T$ of size m, the generalization error is bounded by

$$\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{z})<\gamma_k\right). \tag{12}$$

$$\begin{aligned}&\leq\frac{1}{m}\sum_{\mathbf{z}^{(i)}\in\mathtt{S}_T}\mathbf{1}\left\{\exists\ 1\leq k\leq K:\ r_k(h,\mathbf{z}^{(i)})\leq 3\gamma^k\right\}+\frac{2}{\sqrt{m}-1}\\ &+\mu_{\mathtt{S}_T}\left(h,\left\{r_k-\gamma_k,2\gamma^k\right\}_{k=1}^K\right)+\mu_{\mathcal{D}}\left(h,\left\{r_k-\gamma_k,2\gamma^k\right\}_{k=1}^K\right)\end{aligned} \tag{13}$$

$$+\sqrt{\frac{4\cdot\sum_{k=1}^K\mathrm{KL}\left(\mathcal{N}\left(\mathbf{W}_k,\sigma_k^2\right)\ ||\ \mathcal{P}_k\right)+\log(\frac{2m(K+1)}{\delta})}{m-1}}.$$

To bound $\mu_{\mathcal{D}}\left(h, \{r_k - \gamma_k, 2\gamma^k\}_{k=1}^K\right)$, we can instantiate a recursive procedure. By the choice of variance definition and Lemma 4, we note that at any input $\mathbf{z}$, for all $2 \leq k \leq K$,

$$\begin{aligned}
&\forall\, 1 \leq n \leq k-1 :\ r_n(h, \mathbf{z}) \geq \gamma_k \\
&\implies \text{w. p. at least } (1 - \frac{1}{\sqrt{m}}),\ \forall\, 2 \leq n \leq k, \quad \left|r_n(\hat{h}, \mathbf{z}) - r_n(h, \mathbf{z})\right| \leq \gamma_n \\
&\implies h \text{ is noise-resilient at } z \text{ w.r.t properties } \{r_n - \gamma_n\}_{n=1}^k \text{at thresholds} \{2\gamma_n\}_{n=1}^k.
\end{aligned}$$

In the above, we have also used the fact that the sparse local radius in the first layer is always noise-resilient at the specified γ_1 and choice of variance. Hence, we have the following inequality for all $2 \leq k \leq K$,

$$\begin{aligned}
\mu_{\mathcal{D}}\left(h, \{r_n - \gamma_n, 2\gamma_n\}_{n=1}^k\right) &\leq \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\, 1 \leq n \leq k-1 :\ r_n(h, \mathbf{z}) < \gamma_n\right) \\
\mu_{S_T}\left(h, \{r_n - \gamma_n, 2\gamma_n\}_{n=1}^k\right) &\leq \frac{1}{m} \sum_{\mathbf{z}^{(i)} \in S_T} \mathbf{1}\left(\exists\, 1 \leq n \leq k-1 :\ r_n(h, \mathbf{z}^{(i)}) < \gamma_n\right)
\end{aligned}$$

We can now use this recursively bound the probabilty that the sparse local radii aren't sufficiently large, starting from from K,

$$\begin{aligned}
&\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\, 1 \leq n \leq k :\ r_n(h, \mathbf{z}) < \gamma_n\right). \\
&\leq \frac{1}{m} \sum_{\mathbf{z}^{(i)} \in S_T} \mathbf{1}\left\{\exists\, 1 \leq n \leq k :\ r_n(h, \mathbf{z}^{(i)}) \leq 3\gamma_n\right\} + \frac{2}{\sqrt{m} - 1} \\
&+ \frac{1}{m} \sum_{\mathbf{z}^{(i)} \in S_T} \mathbf{1}\left\{\exists\, 1 \leq n \leq k-1 :\ r_n(h, \mathbf{z}^{(i)}) < \gamma_n\right\} \\
&+ \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left(\exists\, 1 \leq n \leq k-1 :\ \rho^n(h, \mathbf{z}) < \gamma_n\right) \\
&\sqrt{\frac{4 \cdot \sum_{n=1}^k \mathrm{KL}\left(\mathcal{N}\left(\mathbf{W}_n, \sigma_n^2\right) \,||\, \mathcal{P}\right) + \log(\frac{2m(K+1)}{\delta})}{2(m-1)}}.
\end{aligned}$$

The conclusion follows by plugging in these bounds recursively. ■

To prove Theorem 11 we note that the variance is strictly lesser than the variance in Theorem 13 and that the loss terms have been collapsed into the worst-case over layers resulting in a worse generalization bound.

Appendix B. Hyper-Parameter Search

We search for good base hyper-parameters $(\mathbf{s}, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\epsilon})$ as described in Section 3.3. We base our search on the stronger bound in Theorem 13 rather than the simplified result in Theorem 11. For any choice of sensitivity vector $\boldsymbol{\epsilon}$ and sparse risk control α, we choose the sparsity vector to ensure that,

$$\sum_{k\in[K]} \frac{1}{m} \sum_{z^{(i)} \in S_T} \underbrace{\mathbf{1}\left\{\exists\, 1 \leq n \leq k\ :\ r_n(h, \mathbf{z}^{(i)}) \leq 3\gamma_n\right\}}_{=:\ell_{\mathbf{s}}(h, \mathbf{z}^{(i)})} \leq \alpha$$

Doing so automatically controls the other relaxed sparse loss term in Theorem 13,

$$\sum_{k\in[K]}\frac{1}{m}\sum_{z^{(i)}\in\mathtt{S}_T}\mathbf{1}\left\{\exists\, 1\le n\le k\ :\ r_n(h,\mathbf{z}^{(i)})\le\gamma_n\right\}$$

For each $(\boldsymbol{\epsilon},\alpha)$ and input $\mathbf{x}^{(i)}$, the sparsity vector vector $s^*(\mathbf{x}^{(i)},\boldsymbol{\epsilon})=\{s_1^{(i)},\ldots,s_K^{(i)}\}$ is decided in layer-wise fashion. At each layer k, having fixed the sparsity levels $\{s_1^{(i)},\ldots,s_{k-1}^{(i)}\}$[12], one can fix the next sparsity level as,

$$s_k^{(i)}:=\max_{s\in[d_k]}\ s\quad\text{such that } r_k(h,\mathbf{z}^{(i)})>3\gamma_n.$$

$$\Leftrightarrow s_k^{(i)}:=\max_{s\in[d_k]}\ s\quad\text{such that }\sigma\left(\text{sort}\left(-\left[\frac{\mathbf{w}_k[j]\cdot\mathbf{x}_{k-1}^{(i)}}{\xi_k\cdot\zeta_{k-1}}\right]_{j=1}^{d_k},\ s\right)\right)>3\left(-1+\prod_{n=1}^{k}(1+\epsilon_k)\right)$$

where for each feasible s, ξ_k is the closest bound to the relevant sparse norm $\|\mathbf{W}_k\|_{(d_k-1,s_{k-1}^{(i)})}$ and ζ_{k-1} is the bound on the scale of the layer input $\mathbf{x}_{k-1}^{(i)}$ dependent on the previously fixed sparsity levels, i.e. $\zeta_{k-1}:=\mathsf{M}_{\mathcal{X}_0}\prod_{n=1}^{k-1}\xi_n\sqrt{1+\eta_n}$ where for each $1\le n\le k-1$, $\|\mathbf{W}_n\|_{(d_n-1,s_{n-1}^{(i)})}\le\xi_n$ and $\mu_{(s_n^{(i)},s_{n-1}^{(i)})}(\mathbf{W}_n)\le\eta_n$. Under the choice of the $s^*(\mathbf{x}^{(i)},\boldsymbol{\epsilon})$, the sparse loss $\ell_{s^*(\mathbf{x}^{(i)},\boldsymbol{\epsilon})}(h,\mathbf{z}^{(i)})=0$. Further under the choice of the sample-wide minimum sparsity vector $\bar{s}(\boldsymbol{\epsilon})$, based on $s^*(\mathbf{x}^{(i)},\boldsymbol{\epsilon})$, i.e. $\bar{s}_k:=\min_{i\in[m]}s_k^{(i)}$, the average sparse loss $\ell_{\mathbf{s}}$ is zero.

To control the sparse loss, it is sufficent to consider the quantiles of the distribution of $s^*(\mathbf{x},\boldsymbol{\epsilon})$ by the training samples. We denote by $\hat{s}(\boldsymbol{\epsilon},\alpha)$ with layer-wise sparsity levels

$$\hat{s}_k:=\texttt{quantile}\left(\{s^*(\mathbf{x}^{(i)},\boldsymbol{\epsilon})\}_{i=1}^m,\frac{2\alpha}{K(K-1)}\right).$$

Under such a choice,

$$\frac{1}{m}\sum_{z^{(i)}\in\mathtt{S}_T}\mathbf{1}\left\{r_k(h,\mathbf{z}^{(i)})\le 3\gamma_k\right\}\le\frac{2\alpha}{K(K-1)}$$

Hence we can see that,

$$\begin{aligned}
&\sum_{k\in[K]}\frac{1}{m}\sum_{z^{(i)}\in\mathtt{S}_T}\mathbf{1}\left\{\exists\, 1\le n\le k\ :\ r_n(h,\mathbf{z}^{(i)})\le 3\gamma_n\right\}\\
&\le\frac{1}{m}\sum_{z^{(i)}\in\mathtt{S}_T}\sum_{k\in[K]}\sum_{n\in[k-1]}\mathbf{1}\left\{r_n(h,\mathbf{z}^{(i)})\le 3\gamma_n\right\}\\
&\le\frac{1}{m}\sum_{z^{(i)}\in\mathtt{S}_T}\sum_{k\in[K]}\sum_{n\in[k-1]}\frac{2\alpha}{K(K-1)}\\
&\le\frac{1}{m}\sum_{z^{(i)}\in\mathtt{S}_T}\alpha\le\alpha.
\end{aligned}$$

12. We fix $s_0^{(i)}=s_{K+1}^{(i)}=0$ for all i.

Thus we have seen how to control the average sparse loss $\ell_{\mathbf{s}}$ for a fixed sensitivity vector $\boldsymbol{\epsilon}$ and control threshold α. As a final simplification, we note that by the nature of the sensitivity analysis, it is more important that the sparse local radius at lower layers is sufficiently large as compared to later layers (for eg, the last layer only factors into one of the loss terms). Hence in our experiments, we let $\epsilon_k = \frac{\bar{\epsilon}}{K+1-k}$ at all layers for some fixed $\bar{\epsilon} \in [0,1]$. We now search for the best-in-grid generalization bound in the search space $[0,1] \times [0,1]$ to find the find the best-in-grid choice of hyper-parameters $(\bar{\epsilon}, \alpha)$.

Appendix C. PAC-Bayes Tools

In this section, we discuss results from PAC-Bayesian analysis old and new. For the sake of completeness, we first state the classical PAC-Bayes generalization theorem from McAllester (1998); Shalev-Shwartz and Ben-David (2014). Unlike Rademacher analysis, PAC-Bayes provides generalization bounds on a stochastic network. We then quote a useful de-randomization argument from Nagarajan and Kolter (2019b) that provides generalization bounds on the mean network.

C.1. Standard PAC-Bayes Theorems

Theorem 14 *(McAllester (1998); Shalev-Shwartz and Ben-David (2014)) Let $\mathcal{D}$ be an arbitrary distribution over data $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ be a hypothesis class and let $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$ be a loss function. Let $\mathcal{P}$ be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Let $\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be a set of i.i.d training samples from $\mathcal{D}$. Then, with probability of at least $(1-\delta)$ over the choice of training sample $\mathcal{S}$, for all distributions $\mathcal{Q}$ over $\mathcal{H}$ (even such that depend on $\mathcal{S}$) we have*

$$\mathop{\mathbb{E}}_{\hat{h}\sim\mathcal{Q}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_{\mathcal{Z}}} \left[\ell(\hat{h},(\mathbf{x},\mathbf{y}))\right] \leq \mathop{\mathbb{E}}_{\hat{h}\sim\mathcal{Q}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim \mathrm{Unif}(\mathcal{S}_T)} \left[\ell(\hat{h},(\mathbf{x},\mathbf{y}))\right] + \sqrt{\frac{\mathrm{KL}\left(\mathcal{Q}||\mathcal{P}\right) + \log(\frac{m}{\delta})}{2(m-1)}}. \tag{14}$$

Note that Theorem 14 bounds the generalization error of a stochastic predictor $\hat{h} \sim \mathcal{Q}$.

C.2. De-Randomized PAC-Bayes Theorems

Let $\mathcal{D}$ be an arbitrary distribution over data $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let $\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be a set of i.i.d training samples from $\mathcal{D}$. Let $\mathcal{H}$ be a hypothesis class and let $\mathcal{P}$ be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Consider a fixed predictor $h \in \mathcal{H}$ (possibly trained on data). Let $\mathcal{Q}(h,\Sigma)$ be any distribution over $\mathcal{H}$ (possibly data dependent) with mean h and covariance Σ. Let $\rho_n(h,\mathbf{z})$ for $1 \leq n \leq N$ be certain properties of the predictor h on data point $\mathbf{z}$ and let $\gamma_n > 0$ be its associated margin.

Definition 15 *(Noise-resilience, Nagarajan and Kolter (2019b)) A predictor h is said to be noise-resilient at a given data point $\mathbf{z}$ w.r.t properties ρ_n if,*

$$\mathop{\mathbf{Prob}}_{\hat{h}\sim\mathcal{Q}(h,\Sigma)} \left[\exists n : |\rho_n(\hat{h},\mathbf{z}) - \rho_n(h,\mathbf{z})| > \frac{\gamma_n}{2}\right] \leq \frac{1}{\sqrt{m}}. \tag{15}$$

Let $\mu_{\mathcal{D}}\left(h, \{(\rho_n, \gamma_n)\}_{n=1}^N\right)$ *denote the probability that* h *is not noise-resilient at a randomly drawn* $\mathbf{z} \sim \mathcal{D}$*,*

$$\mu_{\mathcal{D}}\left(h, \{(\rho_n, \gamma_n)\}_{n=1}^N\right) := \underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{Q}(h,\Sigma)}{\mathbf{Prob}}\left[\exists n : |\rho_n(\hat{h}, \mathbf{z}) - \rho_n(h, \mathbf{z})| > \frac{\gamma_n}{2}\right] > \frac{1}{\sqrt{m}}.\right]$$

Similarly let $\mu_S\left(h, \{(\rho_n, \gamma_n)\}_{n=1}^N\right)$ *denote the probability that* h *is not noise-resilient at a randomly drawn training sample* $\mathbf{z} \sim \mathfrak{U}(S)$*,*

$$\mu_{S_T}\left(h, \{(\rho_n, \gamma_n)\}_{n=1}^N\right) := \underset{\mathbf{z}\sim\mathfrak{U}(S_T)}{\mathbf{Prob}}\left[\underset{\hat{h}\sim\mathcal{Q}(h,\Sigma)}{\mathbf{Prob}}\left[\exists n : |\rho_n(\hat{h}, \mathbf{z}) - \rho_n(h, \mathbf{z})| > \frac{\gamma_n}{2}\right] > \frac{1}{\sqrt{m}}.\right]$$

Lemma 16 *(Nagarajan and Kolter, 2019b, Theorem C.1) For some fixed margin hyper-parameters* $\{\gamma_n\}_{n=1}^N$*, with probability at least* $(1-\delta)$ *over the draw of training sample* S*, for any predictor* h *and any , we have,*

$$\begin{aligned}
&\underset{\mathbf{z}\sim\mathcal{D}}{\mathbf{Prob}}\left[\exists n : \rho_n(h, \mathbf{z}) < 0\right] \qquad (16)\\
&\leq \frac{1}{m}\sum_{\mathbf{z}^{(i)}\in S_T} \mathbf{1}\left[\exists n : \rho_n(h, \mathbf{z}^{(i)}) < \gamma_n\right] + \mu_{S_T}\left(h, \{(\rho_n, \gamma_n)\}_{n=1}^N\right) + \mu_{\mathcal{D}}\left(h, \{(\rho_n, \gamma_n)\}_{n=1}^N\right)\\
&+ 2\sqrt{\frac{2\mathrm{KL}\left(\mathcal{Q}(h,\Sigma) \,||\, \mathcal{P}\right) + \log(\frac{2m}{\delta})}{2(m-1)}} + \frac{2}{\sqrt{m}-1}.
\end{aligned}$$

In the above lemma, the loss function $\ell(h, \mathbf{z}) := \mathbf{1}\left\{\exists\ n,\ \rho(h, \mathbf{z}) < 0\right\}$. Lemma 16 directly bounds the generalization of a predictor h rather than a stochastic predictor $\hat{h} \sim \mathcal{Q}(h, \Sigma)$.

Appendix D. Sparse Norm, Reduced Babel Function And Gaussian Concentration

Lemma 17 *For any matrix* $\mathbf{W} \in \mathbb{R}^{d_2\times d_1}$ *and any two sparsity levels such that* $(s_2, s_1) \preceq (\hat{s}_2, \hat{s}_1)$*,* $\|\mathbf{W}\|_{(\hat{s}_2,\hat{s}_1)} \leq \|\mathbf{W}\|_{(s_2,s_1)}$

Proof Let $\hat{J}_1 \subseteq [d_1], |J_1| = d_1 - \hat{s}_1$ and $\hat{J}_2 \subseteq [d_2], |\hat{J}_2| = d_2 - \hat{s}_2$ be two index sets such that $\|\mathbf{W}\|_{(\hat{s}_2,\hat{s}_1)} = \left\|\mathcal{P}_{\hat{J}_2,\hat{J}_1}(\mathbf{W})\right\|_2$. Consider any two extended index sets $J_1 \subseteq [d_1], |J_1| = d_1 - s_1$ and $J_2 \subset [d_2], |J_2| = d_2 - s_2$ such that $\hat{J}_1 \subseteq J_1$ and $\hat{J}_2 \subseteq J_2$. Then,

$$\begin{aligned}
\|\mathbf{W}\|_{(\hat{s}_2,\hat{s}_1)} &= \left\|\mathcal{P}_{\hat{J}_2,\hat{J}_1}(\mathbf{W})\right\|_2\\
&\leq \|\mathcal{P}_{J_2,J_1}(\mathbf{W})\|_2\\
&\leq \max_{J_2\subset[d_2],|J_2|=d_2-s_2}\ \max_{J_1\subseteq[d_1],|J_1|=d_1-s_1} \|\mathcal{P}_{J_2,J_1}(\mathbf{W})\|_2\\
&=: \|\mathbf{W}\|_{(s_2,s_1)}.
\end{aligned}$$

■

Recall the definition of reduced babel function from 2,

$$\mu_{s_2,s_1}(\mathbf{W}) := \frac{1}{\|\mathbf{W}\|^2_{(d_2-1,s_1)}} \max_{\substack{J_2\subset[d_2],\\ |J_2|=d_2-s_2}} \max_{j\in J_2} \left[\sum_{\substack{i\in J_2,\\ i\neq j}} \max_{\substack{J_1\subseteq[d_1]\\ |J_1|=d_1-s_1}} \left|\mathcal{P}_{J_1}(\mathbf{w}[i])\mathcal{P}_{J_1}(\mathbf{w}[j])^T\right| \right],$$

To compute this we note, that maximum reduced inner product $\max_{\substack{J_1\subseteq[d_1]\\ |J_1|=d_1-s_1}} \left|\mathcal{P}_{J_1}(\mathbf{w}[i])\mathcal{P}_{J_1}(\mathbf{w}[j])^T\right|$ can be computed taking the sum of the top-k column indices in an element-wise product of rows $\mathbf{w}[i]$ and $\mathbf{w}[j]$. The full algorithm to compute the babel function is described in Algorithm 1. One can note that it has a computational complexity of $\mathcal{O}(d_2^2 d_1)$ for each $\mu_{s_2,s_1}(\mathbf{W})$. Further optimizations that leverage PyTorch broadcasting are possible. The reduced babel function is useful as it provides a bound on the sparse norm.

Algorithm 1 Computing the Reduced Babel function

Require: Weight matrix $W \in \mathbb{R}^{d_2\times d_1}$, sparsity levels $s_2 \in d_2 - 1$ and $0 \leq s_1 \leq d_1 - 1$.
Ensure: The reduced babel function at specified sparsity, $\mu_{s_2,s_1}(\mathbf{W})$.
Initialize a vector of Gerschgorin disk radii, $\mathbf{r} = \mathbf{0}_{d_0}$
Initialize a matrix of maximum reduced inner products, $\mathbf{A} = \mathbf{0}_{(d_2\times d_2)}$.
Initialize top-k elementwise squares $\mathbf{t} = \mathbf{0}_{d_2}$
$\mathbf{t}[i] = \text{SUM}(\text{TOP-K}(\mathbf{w}[i] \circ \mathbf{w}[i], d_1 - s_1))$
$\|\mathbf{W}\|_{d_2-1,s_1} = \sqrt{\max_i \mathbf{t}[i]}$
Compute maximum reduced inner product for each (i, j)
for $1 \leq i \leq d_2$ **do**
  for $1 \leq j \leq d_2,\ j \neq i$ **do**
    positive = $\text{TOP-K}(\mathbf{w}[i] \circ \mathbf{w}[j], d_1 - s_1)$
    negative = $\text{TOP-K}(\mathbf{w}[i] \circ \mathbf{w}[j], d_1 - s_1)$
    $A[i, j] = \max\{\text{SUM}(\text{positive}),\ \text{SUM}(\text{negative})\}$.
  end for
end for
Compute gerschgorin radii.
for $1 \leq i \leq d_2 - 1$ **do**
  $r[i] = \text{SUM}(\text{TOP-K}(\mathbf{a}[i], d_1 - s_1))$
end for
return $\mu_{s_2,s_1}(\mathbf{W}) = \frac{\max_i \mathbf{r}[i]}{\|\mathbf{W}\|_{d_2-1,s_1}}$.
=0

Lemma 18 *(Muthukumar and Sulam, 2022, Lemma 3) For any matrix $\mathbf{W} \in \mathbb{R}^{d_2\times d_1}$, the sparse norm can be bounded as*

$$\|\mathbf{W}\|_{(s_2,s_1)} \leq \|\mathbf{W}\|_{(d_2-1,s_1)} \sqrt{1 + \mu_{s_2,s_1}(\mathbf{W})}.$$

Proof Despite the slight modifications to the reduced babel function definition and the novel sparse norm definition $\|\mathbf{W}\|_{(d_2-1,s_1)}$, the proof follows an identical series of steps as in Muthukumar and Sulam (2022). ■

While Lemma 18 presents a useful deterministic bound for the sparse norm. We can also present a high probability bound on the sparse norm of a Gaussian matrix. To start off we present some well-known lemmas,

Lemma 19 *(Concentration of norm of a sub-Gaussian sub-Matrix) The operator norm of a sub-matrix indexed by sets $J_2 \subseteq [d_2]$ of size $(d_2 - s_1)$ and $J_1 \subseteq [d_1]$ of size $(d_1 - s_1)$ is bounded in high probability,*

$$\textbf{Prob}\left(\|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2 \geq \sigma(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1} + t) \right) \leq e^{-\frac{t^2}{2}}, \quad \forall\, t \geq 0.$$

Proof This is a straightforward application of a classical result on the concentration of norm of Gaussian Matrix (Wainwright, 2019, Theorem 6.1) instantiated for the submatrix $\mathcal{P}_{J_2,J_1}(\mathbf{A})$. ■

Lemma 20 *(Concentration of sparse norm) For sparsity level $0 \leq s_2 \leq d_2 - 1$ and $0 \leq s_1 \leq d_1 - 1$, the operator norm of any sub-matrix indexed of size $(d_2 - s_2) \times (d_1 - s_1)$ is bounded in high probability,*

$$\textbf{Prob}\left(\|\mathbf{A}\|_{(s_2,s_1)} \geq \sigma(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1} + t) \right) \leq \binom{d_2}{s_2}\binom{d_1}{s_1} e^{-\frac{t^2}{2}}, \quad \forall\, t \geq 0.$$

Hence w.p. at least $(1 - \delta)$,

$$\|\mathbf{A}\|_{(s_2,s_1)} \leq \sigma \left(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1}) + \sqrt{2\log\binom{d_2}{s_2} + 2\log\binom{d_1}{s_1} + 2\log\left(\frac{1}{\delta}\right)} \right).$$

Proof Recall that $\max_{\substack{|J_2|=d_2-s_2,\\ |J_1|=d_1-s_1}} \|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2$. For each S_2, S_1, by Lemma 19, we have that,

$$\textbf{Prob}\left(\|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2 \geq \sigma(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1} + t) \right) \leq e^{-\frac{t^2}{2}}, \quad \forall\, t \geq 0.$$

Thus,

$$\begin{aligned}
&\textbf{Prob}\left(\max_{\substack{|J_2|=d_2-s_2\\ |J_1|=d_1-s_1}} \|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2 \geq \sigma(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1} + t) \right) \\
&\leq \textbf{Prob}\left(\exists J_2, J_1,\ \ \|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2 \geq \sigma(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1} + t) \right) \\
&\leq \sum_{\substack{|J_2|=d_2-s_2\\ |J_1|=d_1-s_1}} \textbf{Prob}\left(|\, \|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2 \,| \geq \sigma(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1} + t) \right) \\
&\leq \binom{d_2}{d_2 - s_2}\binom{d_1}{d_1 - s_1} e^{-\frac{t^2}{2}} = \binom{d_2}{s_2}\binom{d_1}{s_1} e^{-\frac{t^2}{2}}.
\end{aligned}$$

Hence w.p. at least $(1 - \delta)$,

$$\max_{\substack{|J_2|=d_2-s_2\\ |J_1|=d_1-s_1}} \|\mathcal{P}_{J_2,J_1}(\mathbf{A})\|_2 \leq \sigma \left(\sqrt{d_2 - s_2} + \sqrt{d_1 - s_1}) + \sqrt{2\log\binom{d_2}{s_2} + 2\log\binom{d_1}{s_1} + 2\log\left(\frac{1}{\delta}\right)} \right).$$

■

Appendix E. Experiment

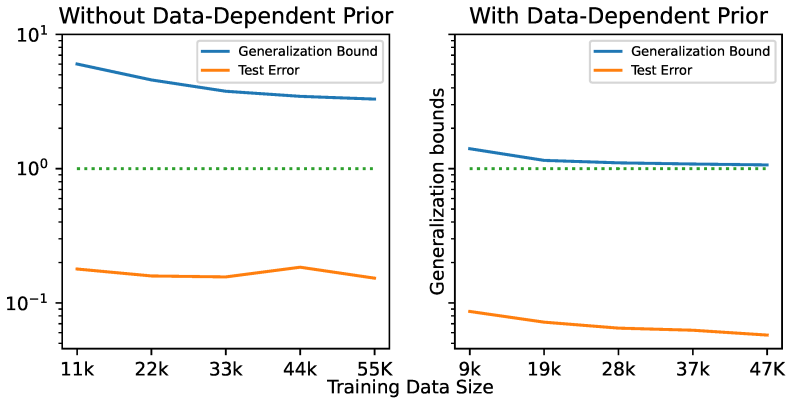


Figure 3: Generalization error of a 2-hidden layer model of width [100, 100] trained on MNIST

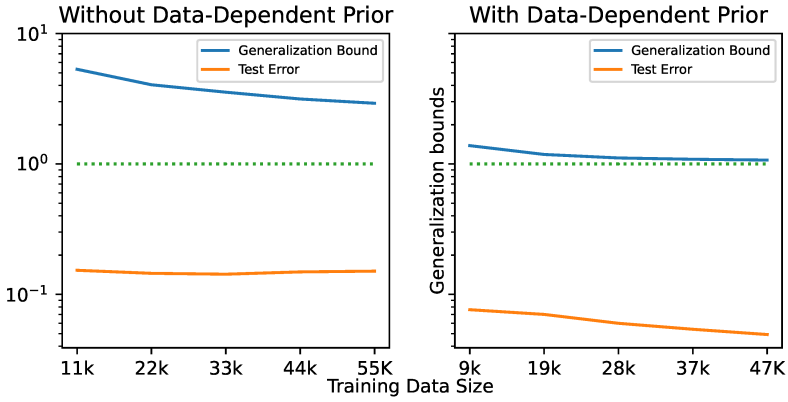
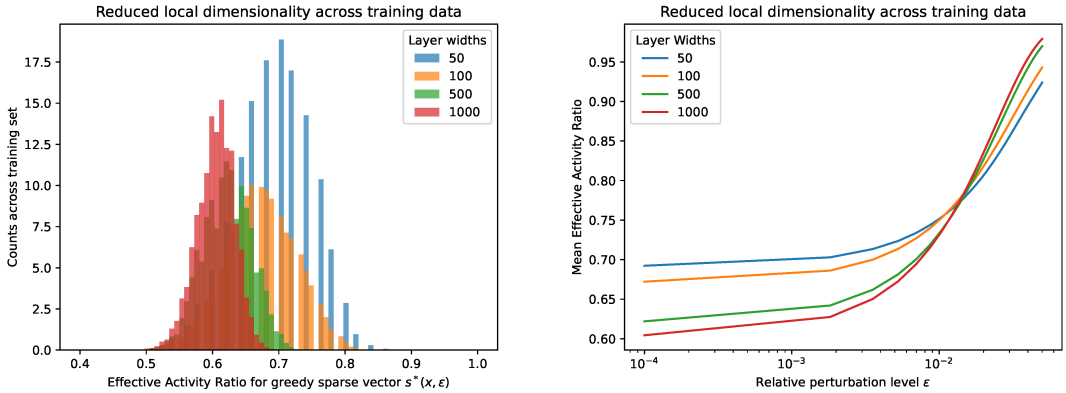


Figure 4: Generalization error of a 2-hidden layer model of width [500, 500] trained on MNIST



(a) Histogram of Effective Activity Ratio at $\epsilon = 10^{-4}$.

(b) Average Effective Activity Ratio.

Figure 5: Effective Activity ratio $\kappa(\mathbf{x}, \epsilon)$ based on greedy sparsity vector $s^*(\mathbf{x}, \epsilon)$ for 2-layer networks (smaller implies sparser stable activations).