Generalization in ReLU Networks via Restricted Isometry and Norm Concentration

Abstract

Regression tasks, while aiming to model relationships across the entire input space, are often constrained by limited training data. Nevertheless, if the hypothesis functions can be represented effectively by the data, there is potential for identifying a model that generalizes well. This paper introduces the Neural Restricted Isometry Property (NeuRIPs), which acts as a uniform concentration event that ensures all shallow ReLU networks are sketched with comparable quality. To determine the sample complexity necessary to achieve NeuRIPs, we bound the covering numbers of the networks using the Sub-Gaussian metric and apply chaining techniques. Assuming the NeuRIPs event, we then provide bounds on the expected risk, applicable to networks within any sublevel set of the empirical risk. Our results show that all networks with sufficiently small empirical risk achieve uniform generalization.

1 Introduction

A fundamental requirement of any scientific model is a clear evaluation of its limitations. In recent years, supervised machine learning has seen the development of tools for automated model discovery from training data. However, these methods often lack a robust theoretical framework to estimate model limitations. Statistical learning theory quantifies the limitation of a trained model by the generalization error. This theory uses concepts such as the VC-dimension and Rademacher complexity to analyze generalization error bounds for classification problems. While these traditional complexity notions have been successful in classification problems, they do not apply to generic regression problems with unbounded risk functions, which are the focus of this study. Moreover, traditional tools in statistical learning theory have not been able to provide a fully satisfying generalization theory for neural networks.

Understanding the risk surface during neural network training is crucial for establishing a strong theoretical foundation for neural network-based machine learning, particularly for understanding generalization. Recent studies on neural networks suggest intriguing properties of the risk surface. In large networks, local minima of the risk form a small bond at the global minimum. Surprisingly, global minima exist in each connected component of the risk's sublevel set and are path-connected. In this work, we contribute to a generalization theory for shallow ReLU networks, by giving uniform generalization error bounds within the empirical risk's sublevel set. We use methods from the analysis of convex linear regression, where generalization bounds for empirical risk minimizers are derived from recent advancements in stochastic processes' chaining theory. Empirical risk minimization for non-convex hypothesis functions cannot generally be solved efficiently. However, under certain assumptions, it is still possible to derive generalization error bounds, as we demonstrate in this paper for shallow ReLU networks. Existing works have applied methods from compressed sensing to bound generalization errors for arbitrary hypothesis functions. However, they do not capture the risk's stochastic nature through the more advanced chaining theory.

This paper is organized as follows. We begin in Section II by outlining our assumptions about the parameters of shallow ReLU networks and the data distribution to be interpolated. The expected and empirical risk are introduced in Section III, where we define the Neural Restricted Isometry Property

.

(NeuRIPs) as a uniform norm concentration event. We present a bound on the sample complexity for achieving NeuRIPs in Theorem 1, which depends on both the network architecture and parameter assumptions. We provide upper bounds on the generalization error that are uniformly applicable across the sublevel sets of the empirical risk in Section IV. We prove this property in a network recovery setting in Theorem 2, and also an agnostic learning setting in Theorem 3. These results ensure a small generalization error, when any optimization algorithm finds a network with a small empirical risk. We develop the key proof techniques for deriving the sample complexity of achieving NeuRIPs in Section V, by using the chaining theory of stochastic processes. The derived results are summarized in Section VI, where we also explore potential future research directions.

2 Notation and Assumptions

In this section, we will define the key notations and assumptions for the neural networks examined in this study. A Rectified Linear Unit (ReLU) function $\phi : \mathbb{R} \to \mathbb{R}$ is given by $\phi(x) := \max(x, 0)$. Given a weight vector $w \in \mathbb{R}^d$, a bias $b \in \mathbb{R}$, and a sign $\kappa \in \{\pm 1\}$, a ReLU neuron is a function $\phi(w, b, \kappa) : \mathbb{R}^d \to \mathbb{R}$ defined as

$$\phi(w, b, \kappa)(x) = \kappa\phi(w^T x + b).$$

Shallow neural networks are constructed as weighted sums of neurons. Typically they are represented by a graph with n neurons in a single hidden layer. When using the ReLU activation function, we can apply a symmetry procedure to represent these as sums:

$$\bar{\phi}_{\bar{p}}(x) = \sum_{i=0}^{n} \phi_{p_i}(x),$$

where $\bar{p}$ is the tuple $(p_1, \ldots, p_n)$.

Assumption 1. The parameters $\bar{p}$, which index shallow ReLU networks, are drawn from a set

$$\bar{P} \subseteq (\mathbb{R}^d \times \mathbb{R} \times \{\pm 1\})^n.$$

For $\bar{P}$, we assume there exist constants $c_w \geq 0$ and $c_b \in [1, 3]$, such that for all parameter tuples $\bar{p} = \{(w_1, b_1, \kappa_1), \ldots, (w_n, b_n, \kappa_n)\} \in \bar{P}$, we have

$$\|w_i\| \leq c_w \quad \text{and} \quad |b_i| \leq c_b.$$

We denote the set of shallow networks indexed by a parameter set $\bar{P}$ by

$$\Phi_{\bar{P}} := \{\phi_{\bar{p}} : \bar{p} \in \bar{P}\}.$$

We now equip the input space $\mathbb{R}^d$ of the networks with a probability distribution. This distribution reflects the sampling process and makes each neural network a random variable. Additionally, a random label y takes its values in the output space $\mathbb{R}$, for which we assume the following.

Assumption 2. The random sample $x \in \mathbb{R}^d$ and label $y \in \mathbb{R}$ follow a joint distribution μ such that the marginal distribution μ_x of sample x is standard Gaussian with density

$$\frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|x\|^2}{2}\right).$$

As available data, we assume independent copies $\{(x_j, y_j)\}_{j=1}^m$ of the random pair (x, y), each distributed by μ.

3 Concentration of the Empirical Norm

Supervised learning algorithms interpolate labels y for samples x, both distributed jointly by μ on $\mathcal{X} \times \mathcal{Y}$. This task is often solved under limited data accessibility. The training data, respecting Assumption 2, consists of m independent copies of the random pair (x, y). During training, the interpolation quality of a hypothesis function $f : \mathcal{X} \to \mathcal{Y}$ can only be assessed at the given random samples $\{x_j\}_{j=1}^m$. Any algorithm therefore accesses each function f through its sketch samples

$$S[f] = (f(x_1), \ldots, f(x_m)),$$

where S is the sample operator. After training, the quality of a resulting model is often measured by its generalization to new data not used during training. With $\mathbb{R}^d \times \mathbb{R}$ as the input and output space, we quantify a function f's generalization error with its expected risk:

$$E_\mu[f] := E_\mu|y - f(x)|^2.$$

The functional $||\cdot||_\mu$, also gives the norm of the space $L^2(\mathbb{R}^d, \mu_x)$, which consists of functions $f : \mathbb{R}^d \to \mathbb{R}$ with

$$\|f\|_\mu^2 := E_{\mu_x}[|f(x)|^2].$$

If the label y depends deterministically on the associated sample x, we can treat y as an element of $L^2(\mathbb{R}^d, \mu_x)$, and the expected risk of any function f is the function's distance to y. By sketching any hypothesis function f with the sample operator S, we perform a Monte-Carlo approximation of the expected risk, which is termed the empirical risk:

$$\|f\|_m^2 := \frac{1}{m}\sum_{j=1}^m \left(f(x_j) - y_j\right)^2 = \left\| \frac{1}{\sqrt{m}}(y_1, \dots, y_m)^T - S[f] \right\|_2^2.$$

The random functional $||\cdot||_m$ also defines a seminorm on $L^2(\mathbb{R}^d, \mu_x)$, referred to as the empirical norm. Under mild assumptions, $||\cdot||_m$ fails to be a norm.

In order to obtain a well generalizing model, the goal is to identify a function f with a low expected risk. However, with limited data, we are restricted to optimizing the empirical risk. Our strategy for deriving generalization guarantees is based on the stochastic relation between both risks. If $\{x_j\}_{j=1}^m$ are independently distributed by μ_x, the law of large numbers implies that for any $f \in L^2(\mathbb{R}^d, \mu_x)$ the convergence

$$\lim_{m\to\infty} \|f\|_m = \|f\|_\mu.$$

While this establishes the asymptotic convergence of the empirical norm to the function norm for a single function f, we have to consider two issues to formulate our concept of norm concentration: First, we need non-asymptotic results, that is bounds on the distance $|\|f\|_m - \|f\|_\mu|$ for a fixed number of samples m. Second, the bounds on the distance need to be uniformly valid for all functions f in a given set.

Sample operators which have uniform concentration properties have been studied as restricted isometries in the area of compressed sensing. For shallow ReLU networks of the form (1), we define the restricted isometry property of the sampling operator S as follows.

Definition 1. Let $s \in (0, 1)$ be a constant and $\bar{P}$ be a parameter set. We say that the Neural Restricted Isometry Property (NeuRIPs($\bar{P}$)) is satisfied if, for all $\bar{p} \in \bar{P}$ it holds that

$$(1 - s)\|\phi_{\bar{p}}\|_\mu \leq \|\phi_{\bar{p}}\|_m \leq (1 + s)\|\phi_{\bar{p}}\|_\mu.$$

In the following Theorem, we provide a bound on the number m of samples, which is sufficient for the operator S to satisfy NeuRIPs($\bar{P}$).

Theorem 1. There exist universal constants C_1, $C_2 \in \mathbb{R}$ such that the following holds: For any sample operator S, constructed from random samples $\{x_j\}$, respecting Assumption 2, let $\bar{P} \subset (\mathbb{R}^d \times \mathbb{R} \times \{\pm 1\})^n$ be any parameter set satisfying Assumption 1 and $||\phi_{\bar{p}}||_\mu > 1$ for all $\bar{p} \in \bar{P}$. Then, for any $u > 2$ and $s \in (0, 1)$, NeuRIPs($\bar{P}$) is satisfied with probability at least $1 - 17\exp(-u/4)$ provided that

$$m \geq \frac{n^3 c_w^2}{(1-s)^2} \max\left(C_1 \frac{(8c_b + d + \ln(2))}{u}, C_2 \frac{n^2 c_w^2}{(u/s)^2} \right).$$

One should notice that, in Theorem 1, there is a tradeoff between the parameter s, which limits the deviation $|\,\|\cdot\|_m - \|\cdot\|_\mu|$, and the confidence parameter u. The lower bound on the corresponding sample size m is split into two scaling regimes when understanding the quotient u of $|\|\cdot\|_m - \|\cdot\|_\mu|/s$ as a precision parameter. While in the regime of low deviations and high probabilities the sample size m must scale quadratically with u/s, in the regime of less precise statements one observes a linear scaling.

4 Uniform Generalization of Sublevel Sets of the Empirical Risk

When the NeuRIPs event occurs, the function norm $||\cdot||_\mu$, which is related to the expected risk, is close to $||\cdot||_m$, which corresponds to the empirical risk. Motivated by this property, we aim to find a shallow ReLU network $\phi_{\bar{p}}$ with small expected risk by solving the empirical risk minimization problem:

$$\min_{\bar{p}\in\bar{P}} \|\phi_{\bar{p}} - y\|_m^2.$$

Since the set $\Phi_{\bar{P}}$ of shallow ReLU networks is non-convex, this minimization cannot be solved with efficient convex optimizers. Therefore, instead of analyzing only the solution $\phi_{\bar{p}}^*$ of the optimization problem, we introduce a tolerance $\epsilon > 0$ for the empirical risk and provide bounds on the generalization error, which hold uniformly on the sublevel set

$$\bar{Q}_{y,\epsilon} := \left\{\bar{p} \in \bar{P} : \|\phi_{\bar{p}} - y\|_m^2 \leq \epsilon\right\}.$$

Before considering generic regression problems, we will initially assume the label y to be a neural network itself, parameterized by a tuple p^* within the hypothesis set P. For all (x, y) in the support of μ, we have $y = \phi_{p^*}(x)$ and the expected risk's minimum on P is zero. Using the sufficient condition for NeuRIPs from Theorem 1, we can provide generalization bounds for $\phi_{\bar{p}} \in \bar{Q}_{y,\epsilon}$ for any $\epsilon > 0$.

Theorem 2. Let $\bar{P}$ be a parameter set that satisfies Assumption 1 and let $u \geq 2$ and $t \geq \epsilon > 0$ be constants. Furthermore, let the number m of samples satisfy

$$m \geq 8n^3 c_w^2 \left(8c_b + d + \ln(2)\right) \max\left(C_1 \frac{u}{(t-\epsilon)^2}, C_2 \frac{n^2 c_w^2 u}{(t-\epsilon)^2}\right),$$

where C_1 and C_2 are universal constants. Let $\{(x_j, y_j)\}_{j=1}^m$ be a dataset respecting Assumption 2 and let there exist a $\bar{p}^* \in \bar{P}$ such that $y_j = \phi_{\bar{p}^*}(x_j)$ holds for all $j \in [m]$. Then, with probability at least $1 - 17\exp(-u/4)$, we have for all $\bar{q} \in \bar{Q}_{y,\epsilon}$ that

$$\|\phi_{\bar{q}} - \phi_{\bar{p}^*}\|_\mu^2 \leq t.$$

Proof. We notice that $\bar{Q}_{y,\epsilon}$ is a set of shallow neural networks with $2n$ neurons. We normalize such networks with a function norm greater than t and parameterize them by

$$\bar{R}_t := \{\phi_{\bar{p}} - \phi_{\bar{p}^*} : \bar{p} \in \bar{P}, \|\phi_{\bar{p}} - \phi_{\bar{p}^*}\|_\mu > t\}.$$

We assume that NeuRIPs($\bar{R}_t$) holds for $s = (t-\epsilon)^2/t^2$. In this case, for all $\bar{q} \in \bar{Q}_{y,\epsilon}$, we have that $\|\phi_{\bar{q}} - \phi_{\bar{p}^*}\|_m \geq t$ and thus $\bar{q} \notin \bar{Q}_{\phi_{\bar{p}^*},\epsilon}$, which implies that $\|\phi_{\bar{q}} - \phi_{\bar{p}^*}\|_\mu \leq t$.

We also note that $\bar{R}_t$ satisfies Assumption 1 with a rescaled constant c_w/t and normalization-invariant c_b, if $\bar{P}$ satisfies it for c_w and c_b. Theorem 1 gives a lower bound on the sample complexity for NeuRIPs($\bar{R}_t$), completing the proof.

At any network where an optimization method terminates, the concentration of the empirical risk at the expected risk can be achieved with less data than needed to achieve an analogous NeuRIPs event. However, in the chosen stochastic setting, we cannot assume that the termination of an optimization and the norm concentration at that network are independent events. We overcome this by not specifying the outcome of an optimization method and instead stating uniform bounds on the norm concentration. The only assumption on an algorithm is therefore the identification of a network that permits an upper bound ϵ on its empirical risk. The event NeuRIPs($\bar{R}_t$) then restricts the expected risk to be below the corresponding level t.

We now discuss the empirical risk surface for generic distributions μ that satisfy Assumption 2, where y does not necessarily have to be a neural network.

Theorem 3. There exist constants C_0, C_1, C_2, C_3, C_4, and C_5 such that the following holds: Let $\bar{P}$ satisfy Assumption 1 for some constants c_w, c_b, and let $\bar{p}^* \in \bar{P}$ be such that for some $c_{\bar{p}^*} \geq 0$ we have

$$E_\mu\left[\exp\left(\frac{(y - \phi_{\bar{p}^*}(x))^2}{c_{\bar{p}^*}^2}\right)\right] \leq 2.$$

We assume, for any $s \in (0, 1)$ and confidence parameter $u > 0$, that the number of samples m is large enough such that

$$m \geq \frac{8}{(1-s)^2} \max\left(C_1\left(\frac{n^3 c_w^2 (8c_b + d + \ln(2))}{u}\right), C_2 n^2 c_w^2 \left(\frac{u}{s}\right)\right).$$

We further select confidence parameters $v_1, v_2 > C_0$, and define for some $\omega \geq 0$ the parameter

$$\eta := 2(1-s)\|\phi_{\bar{p}^*} - y\|_\mu + C_3 v_1 v_2 c_{\bar{p}^*} \frac{1}{(1-s)^{1/4}} + \omega\sqrt{1-s}.$$

If we set $\epsilon = \|\phi_{\bar{p}^*} - y\|_m^2 + \omega^2$ as the tolerance for the empirical risk, then the probability that all $\bar{q} \in \bar{Q}_{y,\epsilon}$ satisfy

$$\|\phi_{\bar{q}} - y\|_\mu \leq \eta$$

is at least

$$1 - 17\exp\left(-\frac{u}{4}\right) - C_5 v_2 \exp\left(-\frac{C_4 m v_2^2}{2}\right).$$

Proof sketch. (Complete proof in Appendix E) We first define and decompose the excess risk by

$$\mathcal{E}(\bar{q}, \bar{p}^*) := \|\phi_{\bar{q}} - y\|_\mu^2 - \|\phi_{\bar{p}^*} - y\|_\mu^2 = \|\phi_{\bar{q}} - \phi_{\bar{p}^*}\|_\mu^2 - \frac{2}{m}\sum_{j=1}^{m}(\phi_{\bar{p}^*}(x_j) - y_j)(\phi_{\bar{q}}(x_j) - \phi_{\bar{p}^*}(x_j)).$$

It suffices to show, that within the stated confidence level we have $\|\phi_{\bar{q}} - y\|_\mu > \eta$. This implies the claim since $\|\phi_{\bar{q}} - y\|_m \leq \epsilon$ implies $\|\phi_{\bar{q}} - y\|_\mu \leq \eta$. We have $E[\mathcal{E}(\bar{q}, \bar{p}^*)] > 0$. It now only remains to strengthen the condition on $\eta > 3\|\phi_{\bar{p}^*} - y\|_\mu$ to achieve $\mathcal{E}(\bar{q}, \bar{p}^*) > \omega^2$. We apply Theorem 1 to derive a bound on the fluctuation of the first term. The concentration rate of the second term is derived similar to Theorem 1 by using chaining techniques. Finally in Appendix E, Theorem 12 gives a general bound to achieve

$$\mathcal{E}(\bar{q}, \bar{p}^*) > \omega^2$$

uniformly for all $\bar{q}$ with $\|\phi_{\bar{q}} - \phi_{\bar{p}^*}\|_\mu > \eta$. Theorem 3 then follows as a simplification.

It is important to notice that, in Theorem 3, as the data size m approaches infinity, one can select an asymptotically small deviation constant s. In this limit, the bound η on the generalization error converges to $3\|\phi_{\bar{p}^*} - y\|_\mu + \omega$. This reflects a lower limit of the generalization bound, which is the sum of the theoretically achievable minimum of the expected risk and the additional tolerance ω. The latter is an upper bound on the empirical risk, which real-world optimization algorithms can be expected to achieve.

5 Size Control of Stochastic Processes on Shallow Networks

In this section, we introduce the key techniques for deriving concentration statements for the empirical norm, uniformly valid for sets of shallow ReLU networks. We begin by rewriting the event NeuRIPs($\bar{P}$) by treating μ as a stochastic process, indexed by the parameter set $\bar{P}$. The event NeuRIPs($\bar{P}$) holds if and only if we have

$$\sup_{\bar{p}\in\bar{P}} |\|\phi_{\bar{p}}\|_m - \|\phi_{\bar{p}}\|_\mu| \leq s \sup_{\bar{p}\in\bar{P}} \|\phi_{\bar{p}}\|_\mu.$$

The supremum of stochastic processes has been studied in terms of their size. To determine the size of a process, it is essential to determine the correlation between its variables. To this end, we define the Sub-Gaussian metric for any parameter tuples $\bar{p}, \bar{q} \in \bar{P}$ as

$$d_{\psi^2}(\phi_{\bar{p}}, \phi_{\bar{q}}) := \inf\left\{C_{\psi^2} \geq 0 : E\left[\exp\left(\frac{|\phi_{\bar{p}}(x) - \phi_{\bar{q}}(x)|^2}{C_{\psi^2}^2}\right)\right] \leq 2\right\}.$$

A small Sub-Gaussian metric between random variables indicates that their values are likely to be close. To capture the Sub-Gaussian structure of a process, we introduce ϵ-nets in the Sub-Gaussian metric. For a given $\epsilon > 0$, these are subsets $\bar{Q} \subseteq \bar{P}$ such that for every $\bar{p} \in \bar{P}$, there is a $\bar{q} \in \bar{Q}$ satisfying

$$d_{\psi^2}(\phi_{\bar{p}}, \phi_{\bar{q}}) \leq \epsilon.$$

The smallest cardinality of such an ϵ-net $\bar{Q}$ is known as the Sub-Gaussian covering number $N(\Phi_{\bar{P}}, d_{\psi^2}, \epsilon)$. The next Lemma offers a bound for such covering numbers specific to shallow ReLU networks.

Lemma 1. Let $\bar{P}$ be a parameter set satisfying Assumption 1. Then there exists a set $\hat{P}$ with $\bar{P} \subseteq \hat{P}$ such that

$$N(\Phi_{\hat{P}}, d_{\psi^2}, \epsilon) \leq 2^n \cdot \left(\frac{16nc_bc_w}{\epsilon} + 1\right)^n \cdot \left(\frac{32nc_bc_w}{\epsilon} + 1\right)^n \cdot \left(\frac{1}{\epsilon}\sin\left(\frac{1}{16nc_w}\right) + 1\right)^d.$$

The proof of this Lemma is based on the theory of stochastic processes and can be seen in Theorem 8 of Appendix C.

To obtain bounds of the form (6) on the size of a process, we use the generic chaining method. This method offers bounds in terms of the Talagrand-functional of the process in the Sub-Gaussian metric. We define it as follows. A sequence $T = (T_k)_{k\in\mathbb{N}_0}$ in a set T is admissible if $T_0 = 1$ and $T_k \leq 2^{(2^k)}$. The Talagrand-functional of the metric space is then defined as

$$\gamma_2(T, d) := \inf_{(T_k)} \sup_{t\in T} \sum_{k=0}^{\infty} 2^k d(t, T_k),$$

where the infimum is taken across all admissible sequences.

With the bounds on the Sub-Gaussian covering number from Lemma 1, we provide a bound on the Talagrand-functional for shallow ReLU networks in the following Lemma. This bound is expected to be of independent interest.

Lemma 2. Let $\bar{P}$ satisfy Assumption 1. Then we have

$$\gamma_2(\Phi_{\bar{P}}, d_{\psi^2}) \leq \sqrt{\frac{2}{\pi}}\left(\frac{8n^{3/2}c_w(8c_b + d + 1)}{\ln(2)}\sqrt{2\ln(2)}\right).$$

The key ideas to show this bound are similar to the ones used to prove Theorem 9 in Appendix C.

To provide bounds for the empirical process, we use the following Lemma, which we prove in Appendix D.

Lemma 3. Let Φ be a set of real functions, indexed by a parameter set $\bar{P}$ and define

$$N(\Phi) := \int_0^{\infty} \sqrt{\ln N(\Phi, d_{\psi^2}, \epsilon)}d\epsilon \quad \text{and} \quad \Delta(\Phi) := \sup_{\phi\in\Phi} \|\phi\|_{\psi^2}.$$

Then, for any $u \geq 2$, we have with probability at least $1 - 17\exp(-u/4)$ that

$$\sup_{\phi\in\Phi} |\|\phi\|_m - \|\phi\|_\mu| \leq \frac{u}{\sqrt{m}}\left[N(\Phi) + \frac{10}{3}\Delta(\Phi)\right].$$

The bounds on the sample complexity for achieving the NeuRIPs event, from Theorem 1, are proven by applying these Lemmata.

Proof of Theorem 1. Since we assume $||\phi_{\bar{p}}||_\mu > 1$ for all $\bar{p} \in \bar{P}$, we have

$$\sup_{\bar{p}\in\bar{P}} |\|\phi_{\bar{p}}\|_m - \|\phi_{\bar{p}}\|_\mu| \leq \sup_{\bar{p}\in\bar{P}} |\|\phi_{\bar{p}}\|_m - \|\phi_{\bar{p}}\|_\mu|/\|\phi_{\bar{p}}\|_\mu.$$

Applying Lemma 3, and further applying the bounds on the covering numbers and the Talagrand-functional for shallow ReLU networks, the NeuRIPs($\bar{P}$) event holds in case of $s > 3$. The sample complexities that are provided in Theorem 1 follow from a refinement of this condition.

6 Uniform Generalization of Sublevel Sets of the Empirical Risk

In case of the NeuRIPs event, the function norm $||\cdot||_\mu$ corresponding to the expected risk is close to $||\cdot||_m$, which corresponds to the empirical risk. With the previous results, we can now derive uniform generalization error bounds in the sublevel set of the empirical risk.

We use similar techniques and we define the following sets.

$$\|f\|_p = \sup_{1\leq q\leq p} \|f\|_q$$

$$\Lambda_{k_0,u} = \inf_{(T_k)} \sup_{f\in F} \sum_{k_0}^{\infty} 2^k \|f - T_k(f)\|_{u2^k}$$

and we need the following lemma:

Lemma 9. For any set F of functions and $u \geq 1$, we have

$$\Lambda_{0,u}(F) \leq 2\sqrt{e}(\gamma_2(F, d_{\psi^2}) + \Delta(F)).$$

Theorem 10. Let P be a parameter set satisfying Assumption 1. Then, for any $u \geq 1$, we have with probability at least $1 - 17\exp(-u/4)$ that

$$\sup_{\bar{p}\in P} \|\phi_{\bar{p}}\|_m - \|\phi_{\bar{p}}\|_\mu \leq \frac{u}{\sqrt{m}} \left(16n^{3/2}c_w(8c_b + d + 1) + 2nc_w\right).$$

Proof. To this end we have to bound the Talagrand functional, where we can use Dudley's inequality (Lemma 6). To finish the proof, we apply the bounds on the covering numbers provided by Theorem 6.

Theorem 11. Let $\bar{P} \subseteq (\mathbb{R}^d \times \mathbb{R} \times \pm 1)^n$ satisfy Assumption 1. Then there exist universal constants C_1, C_2 such that

$$\sup_{\bar{p}\in P} \|\phi_{\bar{p}}\|_m - \|\phi_{\bar{p}}\|_\mu \leq \sqrt{\frac{2}{\pi}} \left(\frac{8n^{3/2}c_w(8c_b + d + 1)}{\ln(2)} \sqrt{2\ln(2)}\right).$$

7 Conclusion

In this study, we investigated the empirical risk surface of shallow ReLU networks in terms of uniform concentration events for the empirical norm. We defined the Neural Restricted Isometry Property (NeuRIPs) and determined the sample complexity required to achieve NeuRIPs, which depends on realistic parameter bounds and the network architecture. We applied our findings to derive upper bounds on the expected risk, which are valid uniformly across sublevel sets of the empirical risk. If a network optimization algorithm can identify a network with a small empirical risk, our results guarantee that this network will generalize well. By deriving uniform concentration statements, we have resolved the problem of independence between the termination of an optimization algorithm at a certain network and the empirical risk concentration at that network. Future studies may focus on performing uniform empirical norm concentration on the critical points of the empirical risk, which could lead to even tighter bounds for the sample complexity.

We also plan to apply our methods to input distributions more general than the Gaussian distribution. If generic Gaussian distributions can be handled, one could then derive bounds for the Sub-Gaussian covering number for deep ReLU networks by induction across layers. We also expect that our results on the covering numbers could be extended to more generic Lipschitz continuous activation functions other than ReLU. This proposition is based on the concentration of measure phenomenon, which provides bounds on the Sub-Gaussian norm of functions on normal concentrating input spaces. Because these bounds scale with the Lipschitz constant of the function, they can be used to find ϵ-nets for neurons that have identical activation patterns.

Broader Impact

Supervised machine learning now affects both personal and public lives significantly. Generalization is critical to the reliability and safety of empirically trained models. Our analysis aims to achieve a deeper understanding of the relationships between generalization, architectural design, and available data. We have discussed the concepts and demonstrated the effectiveness of using uniform concentration events for generalization guarantees of common supervised machine learning algorithms.