Université Paris-Dauphine – PSL
Department of MIDO

LEARNING THEORY:
EXAM – TWO–LAYER NEURAL NETWORKS

Arthur Danjou

Course Supervisor:

Katia MEZIANI

Master 2 ISF (Initial Track)
ACADEMIC YEAR 2025/2026

Abstract

This document examines the mathematical foundations of generalization bounds for two-layer neural networks with ReLU activation. It first derives a naive generalization bound based on *empirical Rademacher complexity* and highlights why this bound fails to explain the behavior of overparameterized models. To address this limitation, we establish a symmetrization inequality tailored to ReLU networks. Exploiting the *positive homogeneity of the ReLU activation*, we then introduce a scale-invariant complexity measure. This reparameterization yields a tighter, width-independent generalization bound and therefore provides a more faithful assessment of the network's true capacity.

Contents

Introduction

This document presents a detailed resolution of the Learning Theory evaluation on two-layer neural networks. The central theme of this exercise is to demonstrate how scale-invariant parameterizations can yield width-independent generalization bounds, overcoming the limitations of naive weight-norm constraints in overparameterized regimes. The study is structured in three main parts. Part A explores a standard but suboptimal width-dependent bound using *empirical Rademacher complexity*. Part B establishes a specific symmetrization inequality for ReLU networks. Finally, Part C leverages the *positive homogeneity of the ReLU activation* to introduce a scale-invariant complexity measure, culminating in a tighter generalization bound that better reflects the true capacity of modern neural networks.

Throughout this document, we assume that for some constant $C > 0$, the input vectors satisfy $\|X\|_2^2 \leq C^2$ almost surely.

Part A: A naive width-dependent bound

Question 1.

By definition, the *empirical Rademacher complexity* of the hypothesis class $\mathcal{H}$ on the sample S_n is given by:

$$\begin{aligned}\mathcal{R}_{S_n}(\mathcal{H}) &= \mathbb{E}_\sigma \left[\sup_{f_\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_\theta(X_i) \right] \\ &= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\substack{\|w\|_2 \leq B_w \\ \|u_j\|_2 \leq B_u \, \forall j}} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{m} w_j \phi(\langle u_j, X_i \rangle) \right]\end{aligned}$$

By linearity, we can swap the sums to isolate the outer weights w_j:

$$\mathcal{R}_{S_n}(\mathcal{H}) = \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\substack{\|w\|_2 \leq B_w, \\ \|u_j\|_2 \leq B_u \, \forall j}} \sum_{j=1}^{m} w_j \left(\sum_{i=1}^{n} \sigma_i \phi(\langle u_j, X_i \rangle) \right) \right]$$

Let $Z \in \mathbb{R}^m$ be the vector with components $Z_j = \sum_{i=1}^{n} \sigma_i \phi(\langle u_j, X_i \rangle)$. The inner product between w and Z can be bounded using the *Cauchy-Schwarz inequality*:

$$\sum_{j=1}^{m} w_j Z_j \leq \|w\|_2 \sqrt{\sum_{j=1}^{m} Z_j^2}$$

Taking the supremum over w subject to the constraint $\|w\|_2 \leq B_w$, the maximum is achieved when the vector w is collinear with Z:

$$\sup_{\|w\|_2 \leq B_w} \sum_{j=1}^{m} w_j Z_j \leq B_w \sqrt{\sum_{j=1}^{m} \left(\sum_{i=1}^{n} \sigma_i \phi(\langle u_j, X_i \rangle) \right)^2}$$

Substituting this back into the expression for the Rademacher complexity, we obtain:

$$\mathcal{R}_{S_n}(\mathcal{H}) \leq \frac{B_w}{n} \mathbb{E}_\sigma \left[\sup_{\|u_j\|_2 \leq B_u \, \forall j} \sqrt{\sum_{j=1}^{m} \left(\sum_{i=1}^{n} \sigma_i \phi(\langle u_j, X_i \rangle) \right)^2} \right]$$

The original supremum is taken over the hidden-layer weight matrix $U \in \mathbb{R}^{m\times d}$. Because the constraint $\|u_j\|_2 \le B_u$ applies independently to each row u_j of U, the feasible set for the matrix is the Cartesian product of the feasible sets for its individual rows. Therefore, maximizing the sum over the matrix U is strictly equivalent to maximizing each term independently over the vectors $u_j \in \mathbb{R}^d$. The square root function being strictly increasing, we can move the supremum inside:

$$\begin{aligned}\mathcal{R}_{S_n}(\mathcal{H}) &\le \frac{B_w}{n}\mathbb{E}_\sigma\left[\sqrt{\sum_{j=1}^{m}\sup_{\|u_j\|_2\le B_u}\left(\sum_{i=1}^{n}\sigma_i\phi(\langle u_j, X_i\rangle)\right)^2}\right] \\ &= \frac{B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u}\left|\sum_{i=1}^{n}\sigma_i\phi(\langle u, X_i\rangle)\right|\right]\end{aligned}$$

We observe that for any real-valued functional A, we have $\sup_u |A(u)| \le \sup_u A(u) + \sup_u(-A(u))$. Taking the expectation over σ and leveraging the fact that the Rademacher vector σ and its negation $-\sigma$ are identically distributed, we obtain $\mathbb{E}_\sigma[\sup_u(-A(u))] = \mathbb{E}_\sigma[\sup_u A(u)]$. Summing these two identical expectations yields a factor of 2:

$$\mathcal{R}_{S_n}(\mathcal{H}) \le \frac{2B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u}\sum_{i=1}^{n}\sigma_i\phi(\langle u, X_i\rangle)\right]$$

The ReLU activation function $\phi(z) = \max(0, z)$ is 1-Lipschitz and satisfies $\phi(0) = 0$. By *Talagrand's contraction lemma*, we can remove the activation function without increasing the complexity beyond its Lipschitz constant:

$$\mathcal{R}_{S_n}(\mathcal{H}) \le 1 \times \frac{2B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u}\sum_{i=1}^{n}\sigma_i\langle u, X_i\rangle\right]$$

By linearity of the inner product and applying *Cauchy-Schwarz inequality* once more:

$$\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u}\sum_{i=1}^{n}\sigma_i\langle u, X_i\rangle\right] = \mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u}\left\langle u, \sum_{i=1}^{n}\sigma_i X_i\right\rangle\right] = B_u\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i X_i\right\|_2\right]$$

To bound this expectation, we apply *Jensen's inequality* using the concavity of the square root function:

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i X_i\right\|_2\right] \le \sqrt{\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i X_i\right\|_2^2\right]}$$

Expanding the squared L_2 norm and using the fact that the Rademacher variables are independent, centered, and have unit variance ($\mathbb{E}[\sigma_i\sigma_k] = \delta_{ik}$):

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i X_i\right\|_2^2\right] = \sum_{i=1}^{n}\sum_{k=1}^{n}\mathbb{E}_\sigma[\sigma_i\sigma_k]\langle X_i, X_k\rangle = \sum_{i=1}^{n}\|X_i\|_2^2$$

Given the assumption that $\|X_i\|_2^2 \le C^2$ almost surely, we obtain:

$$\sqrt{\sum_{i=1}^{n}\|X_i\|_2^2} \le \sqrt{nC^2} = C\sqrt{n}$$

Combining all the partial bounds together yields the final result:

$$\mathcal{R}_{S_n}(\mathcal{H}) \leq \frac{2B_w\sqrt{m}}{n} B_u C\sqrt{n}$$
$$\mathcal{R}_{S_n}(\mathcal{H}) \leq 2B_w B_u C\sqrt{\frac{m}{n}}$$

This demonstrates that the inequality holds with the absolute constant $C_0 = 2$. □

Question 2.

The generalization bound derived in the previous question scales with the factor $\sqrt{m}$, where m represents the width of the hidden layer. In the regime of modern deep learning, neural networks are heavily overparameterized, meaning that the number of parameters vastly exceeds the number of available training samples ($m \gg n$).

Under this overparameterized regime, as the width m grows, the theoretical upper bound on the Rademacher complexity diverges and becomes vacuous. A bound that is arbitrarily large provides no meaningful mathematical guarantee regarding the generalization gap.

However, empirical evidence consistently demonstrates that highly overparameterized neural networks generalize exceptionally well to unseen data, contrary to what this $\sqrt{m}$ dependence suggests. This fundamental discrepancy highlights that the capacity of a neural network is not merely a function of its size or a naive parameter count. Therefore, the hypothesis class defined strictly by independent norm constraints B_w and B_u fails to capture the true inductive bias of the model. Explaining this generalization requires a more refined, width-independent complexity measure.

Part B: A symmetrization inequality for ReLU networks

To simplify the notation, let us denote the inner product for a fixed σ and θ as $Z(\sigma, \theta) = \langle \sigma, f_\theta(\mathcal{S}_n)\rangle$. We aim to show that $\mathbb{E}_\sigma\left[\sup_\theta |Z(\sigma,\theta)|\right] \leq 2\,\mathbb{E}_\sigma\left[\sup_\theta Z(\sigma,\theta)\right]$.

Using the provided hint, the absolute value of any real number z can be decomposed as:

$$|z| = \phi(z) + \phi(-z),$$

where $\phi(z) = \max(0, z)$ is the *ReLU activation function*. Applying this identity to $Z(\sigma, \theta)$:

$$|Z(\sigma,\theta)| = \phi(Z(\sigma,\theta)) + \phi(-Z(\sigma,\theta)).$$

Taking the supremum over θ on both sides and using the sub-additivity of the supremum ($\sup(A + B) \leq \sup A + \sup B$):

$$\sup_\theta |Z(\sigma,\theta)| \leq \sup_\theta \phi(Z(\sigma,\theta)) + \sup_\theta \phi(-Z(\sigma,\theta)).$$

Taking the expectation with respect to σ:

$$\mathbb{E}_\sigma\left[\sup_\theta |Z(\sigma,\theta)|\right] \leq \mathbb{E}_\sigma\left[\sup_\theta \phi(Z(\sigma,\theta))\right] + \mathbb{E}_\sigma\left[\sup_\theta \phi(-Z(\sigma,\theta))\right].$$

Notice that $-Z(\sigma,\theta) = -\langle\sigma, f_\theta(\mathcal{S}_n)\rangle = \langle -\sigma, f_\theta(\mathcal{S}_n)\rangle = Z(-\sigma,\theta)$. Since the Rademacher vector σ and its negation $-\sigma$ are identically distributed, the two expectations on the right-hand side are

equal:

$$\mathbb{E}_\sigma\left[\sup_\theta \phi(-Z(\sigma,\theta))\right] = \mathbb{E}_\sigma\left[\sup_\theta \phi(Z(-\sigma,\theta))\right] = \mathbb{E}_\sigma\left[\sup_\theta \phi(Z(\sigma,\theta))\right].$$

Summing these two identical expectations gives:

$$\mathbb{E}_\sigma\left[\sup_\theta |Z(\sigma,\theta)|\right] \leq 2\,\mathbb{E}_\sigma\left[\sup_\theta \phi(Z(\sigma,\theta))\right].$$

It remains to show that $\mathbb{E}_\sigma[\sup_\theta \phi(Z(\sigma,\theta))] \leq \mathbb{E}_\sigma[\sup_\theta Z(\sigma,\theta)]$. Since ϕ is non-decreasing, for any fixed θ we have $Z(\sigma,\theta) \leq \sup_{\theta'} Z(\sigma,\theta')$, and therefore:

$$\phi(Z(\sigma,\theta)) \leq \phi\left(\sup_{\theta'} Z(\sigma,\theta')\right).$$

Taking the supremum over θ on the left-hand side:

$$\sup_\theta\ \phi(Z(\sigma,\theta)) \leq \phi\left(\sup_\theta Z(\sigma,\theta)\right).$$

By the assumption of the problem, $\sup_\theta Z(\sigma,\theta) \geq 0$ for all $\sigma \in \{-1,+1\}^n$. On any non-negative argument, the ReLU acts as the identity, so:

$$\phi\left(\sup_\theta Z(\sigma,\theta)\right) = \sup_\theta Z(\sigma,\theta).$$

Combining these two steps:

$$\sup_\theta\ \phi(Z(\sigma,\theta)) \leq \sup_\theta\ Z(\sigma,\theta).$$

Taking the expectation over σ and substituting into our earlier bound, we conclude:

$$\mathbb{E}_\sigma\left[\sup_\theta |\langle\sigma, f_\theta(\mathcal{S}_n)\rangle|\right] \leq 2\,\mathbb{E}_\sigma\left[\sup_\theta \langle\sigma, f_\theta(\mathcal{S}_n)\rangle\right].$$

This completes the proof of the symmetrization inequality. □

Part C: A scale-invariant complexity measure

Question 1.

Let $X \in \mathbb{R}^d$ be an arbitrary input vector. The output of the neural network parameterized by the new set of weights $\theta' = \{(\lambda_j w_j, u_j/\lambda_j)\}_{j=1}^m$ is given by:

$$f_{\theta'}(X) = \sum_{j=1}^m (\lambda_j w_j)\phi\left(\left\langle \frac{u_j}{\lambda_j}, X\right\rangle\right)$$

By the linearity of the inner product, we can extract the scalar $\frac{1}{\lambda_j}$:

$$\left\langle \frac{u_j}{\lambda_j}, X\right\rangle = \frac{1}{\lambda_j}\langle u_j, X\rangle$$

Since we are given that $\lambda_j > 0$ for all $j = 1, \ldots, m$, it follows that $\frac{1}{\lambda_j} > 0$. We can therefore apply the positive homogeneity property of the *ReLU activation function*, $\phi(\alpha z) = \alpha\phi(z)$ for any $\alpha > 0$:

$$\phi\left(\frac{1}{\lambda_j}\langle u_j, X\rangle\right) = \frac{1}{\lambda_j}\phi(\langle u_j, X\rangle)$$

Substituting this back into the expression for the network output, we obtain:

$$\begin{aligned} f_{\theta'}(X) &= \sum_{j=1}^{m} \lambda_j w_j \left(\frac{1}{\lambda_j}\phi(\langle u_j, X\rangle)\right) \\ &= \sum_{j=1}^{m} \left(\lambda_j \frac{1}{\lambda_j}\right) w_j \phi(\langle u_j, X\rangle) \\ &= \sum_{j=1}^{m} w_j \phi(\langle u_j, X\rangle) \end{aligned}$$

This final expression is exactly the definition of the network output under the original parameterization θ. Thus, we have shown that $f_{\theta'}(X) = f_\theta(X)$ for any input X, meaning the function computed by the network remains perfectly unchanged under this reparameterization.

Question 2.

We evaluate the complexity measure $\mathcal{C}$ on the reparameterized weights $\theta' = \{(\lambda_j w_j, u_j/\lambda_j)\}_{j=1}^m$. By definition, we have:

$$\mathcal{C}(\theta') = \sum_{j=1}^{m} |\lambda_j w_j| \left\| \frac{u_j}{\lambda_j} \right\|_2$$

Using the properties of the absolute value for the product of scalars and the property of the L_2 norm regarding scalar multiplication ($\|\alpha v\|_2 = |\alpha|\|v\|_2$ for any scalar α and vector v), we can rewrite the terms inside the sum as:

$$\mathcal{C}(\theta') = \sum_{j=1}^{m} |\lambda_j||w_j| \frac{1}{|\lambda_j|} \|u_j\|_2$$

Since we are given that $\lambda_j > 0$ for all $j = 1, \ldots, m$, we have $|\lambda_j| = \lambda_j$. The expression simplifies to:

$$\begin{aligned} \mathcal{C}(\theta') &= \sum_{j=1}^{m} \lambda_j |w_j| \frac{1}{\lambda_j} \|u_j\|_2 \\ &= \sum_{j=1}^{m} \left(\lambda_j \frac{1}{\lambda_j}\right) |w_j| \|u_j\|_2 \\ &= \sum_{j=1}^{m} |w_j| \|u_j\|_2 \end{aligned}$$

This final expression is exactly the definition of the complexity measure for the original parameters θ. We have thus demonstrated that $\mathcal{C}(\theta') = \mathcal{C}(\theta)$, proving that the complexity measure is strictly scale-invariant.

Question 3.

We want to bound the empirical Rademacher complexity of the hypothesis class $\mathcal{H}'$. By definition of *empirical Rademacher complexity*, we have:

$$\mathcal{R}_{S_n}(\mathcal{H}') = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\mathcal{C}(\theta)\leq B} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{m} w_j \phi(\langle u_j, X_i \rangle) \right]$$

Using the scale-invariance properties established in the previous questions, we can reparameterize the network without changing its output or its complexity measure. For $u_j \neq 0$, define $\alpha_j = w_j \|u_j\|_2 \in \mathbb{R}$ and $v_j = \frac{u_j}{\|u_j\|_2} \in \mathbb{R}^d$. If $u_j = 0$, set $\alpha_j = 0$ and $v_j = 0$. By the positive homogeneity of *ReLU activation function*, the network output can be rewritten as:

$$f_\theta(X_i) = \sum_{j=1}^{m} w_j \|u_j\|_2 \phi\left(\left\langle \frac{u_j}{\|u_j\|_2}, X_i \right\rangle \right) = \sum_{j=1}^{m} \alpha_j \phi(\langle v_j, X_i \rangle)$$

The complexity measure constraint becomes $\sum_{j=1}^{m} |\alpha_j| = \sum_{j=1}^{m} |w_j| \|u_j\|_2 = \mathcal{C}(\theta) \leq B$, and by definition, the new weight vectors satisfy $\|v_j\|_2 \leq 1$. Substituting this into the Rademacher complexity and swapping the sums yields:

$$\mathcal{R}_{S_n}(\mathcal{H}') = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\substack{\sum_{j=1}^{m} |\alpha_j| \leq B, \\ \|v_j\|_2 \leq 1}} \sum_{j=1}^{m} \alpha_j \left(\sum_{i=1}^{n} \sigma_i \phi(\langle v_j, X_i \rangle) \right) \right]$$

By applying *Hölder's inequality*, the inner product over the m hidden units is bounded by the L_1 norm of α multiplied by the maximum absolute value of the inner sums:

$$\sum_{j=1}^{m} \alpha_j \left(\sum_{i=1}^{n} \sigma_i \phi(\langle v_j, X_i \rangle) \right) \leq \left(\sum_{j=1}^{m} |\alpha_j| \right) \sup_{\|v\|_2 \leq 1} \left| \sum_{i=1}^{n} \sigma_i \phi(\langle v, X_i \rangle) \right|$$

Using the constraint $\sum_{j=1}^{m} |\alpha_j| \leq B$, we can bound the Rademacher complexity by taking the scalar B out of the supremum:

$$\mathcal{R}_{S_n}(\mathcal{H}') \leq \frac{B}{n}\mathbb{E}_\sigma \left[\sup_{\|v\|_2 \leq 1} \left| \sum_{i=1}^{n} \sigma_i \phi(\langle v, X_i \rangle) \right| \right]$$

Now, we must apply the symmetrization inequality derived in Part B. It allows us to remove the absolute value inside the supremum at the cost of a factor of 2:

$$\mathcal{R}_{S_n}(\mathcal{H}') \leq \frac{2B}{n}\mathbb{E}_\sigma \left[\sup_{\|v\|_2 \leq 1} \sum_{i=1}^{n} \sigma_i \phi(\langle v, X_i \rangle) \right]$$

From this point, the mathematical steps perfectly mirror Part A. By applying *Talagrand's contraction lemma* for the 1-Lipschitz ReLU function, we can remove the activation without increasing the bound:

$$\mathcal{R}_{S_n}(\mathcal{H}') \leq \frac{2B}{n}\mathbb{E}_\sigma \left[\sup_{\|v\|_2 \leq 1} \left\langle v, \sum_{i=1}^{n} \sigma_i X_i \right\rangle \right]$$

Applying the *Cauchy-Schwarz inequality*, the supremum over v subject to $\|v\|_2 \leq 1$ is simply the L_2 norm of the Rademacher sum:

$$\mathcal{R}_{S_n}(\mathcal{H}') \leq \frac{2B}{n}\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^{n} \sigma_i X_i \right\|_2 \right]$$

Finally, using *Jensen's inequality* and the independence of the Rademacher variables as done in Part A, we bound the expected L_2 norm using the explicit assumption $\|X_i\|_2 \leq C$:

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^{n} \sigma_i X_i \right\|_2 \right] \leq \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2} \leq \sqrt{nC^2} = C\sqrt{n}$$

Injecting this back gives the final scale-invariant generalization bound:

$$\mathcal{R}_{S_n}(\mathcal{H}') \leq \frac{2B}{n} C\sqrt{n}$$
$$\mathcal{R}_{S_n}(\mathcal{H}') \leq \frac{2BC}{\sqrt{n}}$$

This concludes the proof, demonstrating that the inequality holds with the absolute constant $C_1 = 2$. □

Question 4.

The generalization bound obtained in Part A is strictly width-dependent, scaling as $\mathcal{O}(\sqrt{\frac{m}{n}})$. As discussed previously, this bound becomes vacuous in the overparameterized regime where the number of hidden units m is extremely large, thereby failing to explain why very wide neural networks can still generalize well in practice.

Conversely, the bound derived in Part C using the complexity measure $\mathcal{C}(\theta)$ scales as $\mathcal{O}(\frac{1}{\sqrt{n}})$. This new bound is entirely independent of the network's width m, effectively decoupling the statistical guarantee from the raw number of parameters.

The measure $\mathcal{C}(\theta)$ provides a more appropriate notion of capacity because it accounts for the algebraic symmetries of the architecture, specifically the positive homogeneity of the ReLU activation function. In Part A, the hypothesis class was defined by independent constraints on the norms of the weights. However, due to scale-invariance, one could artificially inflate the norm of the hidden layer while proportionally deflating the output layer. This transformation results in the exact same mathematical function but drastically inflates the naive complexity bound.

The scale-invariant measure $\mathcal{C}(\theta)$ resolves this ambiguity by coupling the norms of the incoming and outgoing weights for each hidden unit. Consequently, it bounds the true functional capacity of the network rather than the arbitrary scale of its parameters. This proves that overparameterized neural networks can generalize perfectly well, provided that their scale-invariant complexity remains bounded.

Conclusion

This exercise highlights a foundational concept in the modern statistical learning theory of neural networks: the necessity of scale-invariant capacity measures.

In Part A, we demonstrated that a naive approach relying strictly on weight norms yields a generalization bound that scales with $\sqrt{m}$. In the context of modern deep learning, where networks are massively overparameterized ($m \gg n$), this width-dependent bound becomes vacuous and fails to explain the strong empirical generalization of such models.

By exploiting the positive homogeneity of the ReLU activation function in Parts B and C, we established that the network's function remains invariant under specific rescaling operations. This allowed us to define a new complexity measure, $\mathcal{C}(\theta)$, which captures the true capacity of the two-layer network independently of its width m. Consequently, the resulting generalization bound depends solely on the sample size $\mathcal{O}(\frac{1}{\sqrt{n}})$ and the scale-invariant complexity.

Ultimately, this illustrates that overparameterized neural networks can generalize effectively, provided that their scale-invariant complexity is bounded, a property often achieved in practice through explicit regularization or the implicit regularization of gradient-based optimization methods.

Appendix: Mathematical Tools

This appendix outlines the fundamental mathematical definitions and theorems used in the resolution of this exercise. As these are standard results in empirical process theory and probability, they are stated here without proof and assumed to be true.

Definition 1 (Empirical Rademacher Complexity)**.** Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, and let $S_n = (X_1, \ldots, X_n)$ be a fixed sample of size n with elements in $\mathcal{X}$. The empirical Rademacher complexity of $\mathcal{H}$ with respect to the sample S_n is defined as:

$$\mathcal{R}_{S_n}(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right]$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables taking values in $\{-1, +1\}$ with equal probability.

Definition 2 (ReLU Activation and Positive Homogeneity)**.** The Rectified Linear Unit (ReLU) activation function is defined as $\phi(z) = \max(0, z)$. It satisfies the positive homogeneity property, meaning that for any strictly positive scalar $\alpha > 0$ and any real number z:

$$\phi(\alpha z) = \alpha \phi(z)$$

Furthermore, the absolute value of any real number can be decomposed using the ReLU function as $|z| = \phi(z) + \phi(-z)$.

Theorem 3 (Cauchy-Schwarz Inequality)**.** *For all vectors u, v in an inner product space over $\mathbb{R}$, the Cauchy-Schwarz inequality states that:*

$$|\langle u, v \rangle| \leq \|u\|_2 \|v\|_2$$

where $\langle \cdot, \cdot \rangle$ is the inner product and $\| \cdot \|_2$ is the induced L_2 norm.

Theorem 4 (Hölder's Inequality)**.** *For any vectors $u, v \in \mathbb{R}^m$, the absolute value of their inner product is bounded by the product of their L_1 and L_∞ norms:*

$$|\langle u, v \rangle| \leq \|u\|_1 \|v\|_\infty$$

where $\|u\|_1 = \sum_{j=1}^{m} |u_j|$ and $\|v\|_\infty = \max_{1 \leq j \leq m} |v_j|$.

Theorem 5 (Jensen's Inequality)**.** *Let X be a random variable and φ be a measurable function. If φ is convex, then:*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Conversely, if φ is a concave function, the inequality is reversed:

$$\varphi(\mathbb{E}[X]) \geq \mathbb{E}[\varphi(X)]$$

Lemma 6 (Talagrand's Contraction Lemma)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz function such that $\phi(0) = 0$. For any hypothesis class $\mathcal{H}$ of real-valued functions and any sample $S_n = (X_1, \ldots, X_n)$, the empirical Rademacher complexity of the composition satisfies:*

$$\frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i \phi(f(X_i)) \right] \leq \frac{L}{n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i f(X_i) \right]$$