# A Proofs

## A.1 Equivalence Classes for Neural Networks

In this section, we show that two neural networks are equivalent only via sign-flips and rotations of the weights. To do this, we prove that the first-layer units must be linearly independent. The proof is uses Lemma 20.6 in Anthony & Bartlett (2009), which is based on Albertini et al. (1993). We will only prove the case where there is a single input feature.

**Lemma 1.** *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be the standard sigmoid function. Choose $k \in \mathbb{N}$ and $\theta_1, \theta_{1,0}, \theta_2, \theta_{2,0}, ..., \theta_k, \theta_{k,0} \in \mathbb{R}$ such that $\theta_j \neq 0$ and $(\theta_j, \theta_{j,0}) \neq \pm(\theta_{j'}, \theta_{j',0})$ for all $j$ and $j' \neq j$ and $j, j' \in \{1...k\}$. Then the set*

$$\{x \mapsto \sigma(\theta_j x + \theta_{j,0}) : 1 \leq j \leq k\} \cup \{x \mapsto 1\}$$

*of real functions are linearly independent over a set $\mathcal{X}$ containing some ball of radius $\delta > 0$.*

*Proof.* Suppose that there exists some $\alpha_0, ..., \alpha_{k+1} \in \mathbb{R}$ not all zero such that for all $x \in \mathcal{X}$, we have $g(x) = \alpha_0 + \sum_{j=1}^{k} \alpha_j \sigma(\theta_j x + \theta_{j,0}) = 0$. Therefore any convergent sequence $\{x_i\} \in \mathcal{X}$ will have limit $\lim_{i \to \infty} g(x_i) = 0$. In addition, we know that the sigmoid function is analytic in the strip $\{z \in \mathbb{C} : |\mathcal{I}(z)| < \pi\}$ where $\mathcal{I}(z)$ is the imaginary part of $z$, so $g$ is analytic in this strip too. By the Principle of Permanence, this means that $g(x)$ must be zero for all $x \in \mathbb{R}$. This contradicts Lemma 20.6 in Anthony & Bartlett (2009). □

## A.2 Proof of Theorem 1

The proof for Theorem 1 is composed of two main steps. First, we show that the excess loss of any $f_{\boldsymbol{\eta}}$ is lower bounded by a quadratic function of the distance from $\boldsymbol{\eta}$ to $EQ_0$. Using this, we analyze the definition of $\hat{\boldsymbol{\eta}}$ to derive the result.

**Lemma 2** (Bounded third derivatives)**.** *If the first, second, and third derivatives of the sigmoidal function are bounded and $\mathcal{X}$ is bounded, then the third derivative of the expected loss function is bounded uniformly over $\Theta$ by some constant $G > 0$. That is, we have*

$$\sup_{\boldsymbol{\eta} \in \Theta} \max_{j_1 j_2 j_3} \left| \frac{\partial^3}{\partial \eta_{j_1} \partial \eta_{j_2} \partial \eta_{j_3}} \mathbb{P} \ell_\eta \right| \leq G. \tag{1}$$

To prove this, simply compute the third derivative and bound each term. The proof is straightforward so we omit it.

Using the local strong convexity assumption in Conditions 1 and 2 and the two lemmas above, we can lower bound the excess loss by a quadratic function.

**Lemma 3** (Quadratic lower bound)**.** *Suppose Conditions 1 and 2 hold. Consider constants $R > 0$ and $c_1, c_2 \geq 1/2$. Suppose $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}, b)$ satisfies $d_0(\boldsymbol{\eta}) \leq R$ and the following condition:*

$$\|\boldsymbol{\theta}_{S^c}\|_1 \leq c_1 \left\| (\boldsymbol{t}, \boldsymbol{\beta}, b) - \left( \boldsymbol{t}_0^{(\eta)}, \boldsymbol{\beta}_0^{(\eta)}, b_0^{(\eta)} \right) \right\|_2 + c_2 \|\boldsymbol{\theta}_S - \boldsymbol{\theta}_0^{(\eta)}\|_1. \tag{2}$$

*Then* $\mathcal{E}(\boldsymbol{\eta}) \geq \frac{1}{C_0^2}\|\boldsymbol{\eta} - \boldsymbol{\eta}_0^{(\eta)}\|_2^2$ *where* $C_0^2 = \max\left[\frac{1}{\epsilon_0}, \frac{R^2}{\alpha_{\epsilon_0}}\right]$, $\epsilon_0 = \frac{3h_{min}}{2C}$, *and* $C = \frac{1}{6}G\left((c_1+1)\sqrt{2m+1} + (c_2+1)\sqrt{ms}\right)^3$, *and* $G$ *is from Lemma 2.*

*Proof.* By Taylor expansion, we have

$$\mathcal{E}(\boldsymbol{\eta}) = \frac{1}{2}\left(\boldsymbol{\eta} - \boldsymbol{\eta}_0^{(\eta)}\right)^\top I(\boldsymbol{\eta}_0)\left(\boldsymbol{\eta} - \boldsymbol{\eta}_0^{(\eta)}\right) + r_\eta \tag{3}$$

By Lemma 2, we have $|r_\eta| \leq \frac{G}{6}\|\boldsymbol{\eta} - \boldsymbol{\eta}_0^{(\eta)}\|_1^3$. Moreover, if (2) holds, then

$$\begin{aligned}
\|\boldsymbol{\eta} - \boldsymbol{\eta}_0^{(\eta)}\|_1 &\leq c_1\left\|(\boldsymbol{t},\boldsymbol{\beta},b) - \left(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)}\right)\right\|_2 + (c_2+1)\|\boldsymbol{\theta}_S - \boldsymbol{\theta}_0^{(\eta)}\|_1 + \left\|(\boldsymbol{t},\boldsymbol{\beta},b) - \left(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)}\right)\right\|_1 \\
&\leq (c_1+1)\sqrt{2m+1}\left\|(\boldsymbol{t},\boldsymbol{\beta},b) - \left(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)}\right)\right\|_2 + (c_2+1)\sqrt{ms}\|\boldsymbol{\theta}_S - \boldsymbol{\theta}_0^{(\eta)}\|_2 \\
&\leq \left((c_1+1)\sqrt{2m+1} + (c_2+1)\sqrt{ms}\right) d_0(\boldsymbol{\eta}).
\end{aligned}$$

Thus, if Condition 1 holds and (2) holds, we have

$$\mathcal{E}(\boldsymbol{\eta}) \geq \frac{1}{2}h_{min}d_0^2(\boldsymbol{\eta}) - \frac{G}{6}\left((c_1+1)\sqrt{2m+1} + (c_2+1)\sqrt{ms}\right)^3 d_0^3(\boldsymbol{\eta}).$$

Now we simply apply the Auxiliary Lemma in Städler et al. (2010) but with $d_0$ as the norm. □

Finally, we are ready to prove Theorem 1. We use $\mathbb{P}_n$ to denote the expectation over the empirical distribution.

*Proof of Theorem 1.* By definition,

$$\mathbb{P}_n\ell_{\hat{\eta}} + \lambda\|\hat{\boldsymbol{\theta}}\|_1 \leq \mathbb{P}_n\ell_0 + \lambda\|\boldsymbol{\theta}_0^{(\eta)}\|_1.$$

For simplicity, we'll denote $\boldsymbol{\eta}_0^{(\hat{\eta})}$, the closest point in $EQ_0$ to $\hat{\eta}$, by $\boldsymbol{\eta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)$. Rearranging, we get

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda\|\hat{\boldsymbol{\theta}}\|_1 \leq \left|(\mathbb{P}_n - \mathbb{P})\left(\ell_{\eta_0} - \ell_{\hat{\eta}}\right)\right| + \lambda\|\boldsymbol{\theta}_0^{(\eta)}\|_1$$

Over the set $\mathcal{T}_{\tilde{\lambda}}$, we have that

$$\begin{aligned}
\mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda\|\hat{\boldsymbol{\theta}}\|_1 \leq T\tilde{\lambda}\left(\tilde{\lambda} \vee \left\|(\hat{\boldsymbol{t}},\hat{\boldsymbol{\beta}},\hat{b}) - (\boldsymbol{t}_0,\boldsymbol{\beta}_0,b_0)\right\|_2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1\right) \\
+ \lambda\|\boldsymbol{\theta}_0\|_1
\end{aligned} \tag{4}$$

Now we consider three possible cases.

**Case 1:** $\tilde{\lambda} \geq \left\|(\hat{\boldsymbol{t}},\hat{\boldsymbol{\beta}},\hat{b}) - (\boldsymbol{t}_0,\boldsymbol{\beta}_0,b_0)\right\|_2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$

Then (4) reduces to

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda\|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq T\tilde{\lambda}^2 + \lambda\tilde{\lambda} \tag{5}$$

**Case 2:** $\tilde{\lambda} < \left\|(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}}, \hat{b}) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\right\|_2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$

In this case, (4) reduces to the following, after some rearranging of terms:

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + (\lambda - T\tilde{\lambda})\|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq T\tilde{\lambda}\left\|(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}}, \hat{b}) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\right\|_2 + (\lambda + T\tilde{\lambda})\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_1 \tag{6}$$

We use Lemma 3 later, so we check the conditions are satisfied.

Given the constraints of $\lambda$, the (2) in Lemma 3 is satisfied with $c_1 = \frac{1}{2}$ and $c_2 = \frac{1+a}{2}$.

In addition, it is easy to show that

$$\|\hat{\boldsymbol{\theta}}\|_1 \leq K + \frac{1+a}{2}\|\boldsymbol{\theta}_{0,S}\|_1.$$

Thus, we can apply Lemma 3 with $R = 2K + (3+a)\max_{(\boldsymbol{\theta}_{0,\cdot})\in EQ_0}\|\boldsymbol{\theta}_0\|_1$.

Now we split Case 2 into two sub-cases.

**Case 2a:** $\left\|(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}}, \hat{b}) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\right\|_2 \geq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1.$

In this case, (4) reduces to

$$\begin{aligned}
\mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda\|\hat{\boldsymbol{\theta}}_{S^c}\|_1 &\leq \left(2T\tilde{\lambda} + \lambda\right)\left\|(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}}, \hat{b}) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\right\|_2 \\
&\leq \frac{1}{2}\left\{\left(2T\tilde{\lambda} + \lambda\right)^2 C_0^2 + \frac{1}{C_0^2}\left\|(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}}, \hat{b}) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\right\|_2^2\right\} \\
&\leq \frac{1}{2}\left\{\left(2T\tilde{\lambda} + \lambda\right)^2 C_0^2 + \mathcal{E}(\hat{\boldsymbol{\eta}})\right\},
\end{aligned}$$

where we apply Lemma 3 in the last inequality. Rearranging, we get

$$\frac{1}{2}\mathcal{E}(\hat{\boldsymbol{\eta}}) + \lambda\|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq \left(2T\tilde{\lambda} + \lambda\right)^2 \frac{C_0^2}{2}.$$

**Case 2b:** $\left\|(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}}, \hat{b}) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\right\|_2 < \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1.$

In this case, (4) reduces to

$$\begin{aligned}
\mathcal{E}(\hat{\boldsymbol{\eta}}) + \left(\lambda - 2T\tilde{\lambda}\right)\|\hat{\boldsymbol{\theta}}_{S^c}\|_1 &\leq \left(2T\tilde{\lambda} + \lambda\right)\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_1 \\
&\leq \left(2T\tilde{\lambda} + \lambda\right)\sqrt{m|S|}\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_2 \\
&\leq \frac{1}{2}\left\{\left(2T\tilde{\lambda} + \lambda\right)^2 m|S|C_0^2 + \frac{1}{C_0^2}\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_2^2\right\} \\
&\leq \frac{1}{2}\left\{\left(2T\tilde{\lambda} + \lambda\right)^2 2m|S|C_0^2 + \frac{1}{2}\mathcal{E}(\hat{\boldsymbol{\eta}})\right\},
\end{aligned}$$

where we apply Lemma 3 in the last inequality. Rearranging, we get

$$\mathcal{E}(\hat{\boldsymbol{\eta}}) + \left(\lambda - 2T\tilde{\lambda}\right)\|\hat{\boldsymbol{\theta}}_{S^c}\|_1 \leq \left(2T\tilde{\lambda} + \lambda\right)^2 \frac{m|S|C_0^2}{2}.$$

Combining all three cases, we have proven the desired result. □

## A.3 Proof of Theorem 2

Next we use empirical process techniques to prove Theorem 2. To control the behavior of the functions, we use metric entropy, a classic measure of complexity. Let the $u$-entropy of a function class $\mathcal{G}$ with respect to the norm $\|\cdot\|$ be denoted $H(u, \mathcal{G}, \|\cdot\|)$, which is equal to the log of its $u$-covering number $N(u, \mathcal{G}, \|\cdot\|)$. For more details on metric entropy, refer to van der Vaart & Wellner (1996).

Define the empirical process term

$$v_n(\boldsymbol{\eta}) = (\mathbb{P}_n - \mathbb{P}_{XY})\ell_\eta(x, y). \tag{7}$$

The basic idea of the proof is to split $v_n(\boldsymbol{\eta})$ into a truncated component $v_n^{trunc}(\boldsymbol{\eta})$ and a remainder term $v_n - v_n^{trunc}(\boldsymbol{\eta})$, where

$$v_n^{trunc}(\boldsymbol{\eta}) = (\mathbb{P}_n - \mathbb{P}_{XY})\left[\ell_\eta(x, y)1\left\{G(y) \leq M_n\right\}\right] \tag{8}$$

for some truncation function $G(\cdot)$ and constant $M_n > 0$. We then control the truncated empirical process and the remainder separately.

We begin with defining our truncation function $G(\cdot)$. We need to consider the derivative of the loss function with respect to the values of the hidden layer nodes. To do this, let $\phi(x; \boldsymbol{\eta}) = (g_{\theta,t}(x), \boldsymbol{\beta}, b)$ where $g_{\theta,t}(x) = \boldsymbol{\theta}^\top x + t$. Then the loss function with respect to the hidden layer nodes $\phi = (g_{\theta,t}, \boldsymbol{\beta}, b)$ is $\tilde{\ell}(y, \phi) = \left(y - \tilde{f}(\phi)\right)^2$ where $\tilde{f}(\phi) = \sum_{i=1}^m \beta_i \psi\left(\{g_{\theta,t}\}_i\right) + b$. The following lemma shows that the derivative of the loss function with respect to the hidden layer nodes is bounded by a function $G(\cdot)$.

For the rest of this document, $c_i$ will denote some positive constant that can depend on $X_{max}$.

**Lemma 4** (Bounded derivative of the loss function)**.** *Suppose the first, second and third derivatives of $\psi$ are bounded. Then we have*

$$\sup_{\phi:\|(\beta,b)\|_2 \leq K} \left\|\nabla_\phi \tilde{\ell}(y, \phi)\right\|_\infty \leq G(y) = \left(\frac{K}{4} \vee 1\right)\left|K\sqrt{m+1} + |y|\right|.$$

The proof is straightforward so we omit here. The proof just bounds each component in the derivative.

In order to control the truncated empirical process, we will bound the entropy of the function class

$$\mathcal{G}_r = \left\{\left\{\ell_\eta(x, y) - \ell_{\eta_0}(x, y)\right\} 1\left\{G(y) \leq M_n\right\} : \boldsymbol{\eta} \in \Theta_r\right\}$$

where

$$\Theta_r = \left\{\eta = (\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}, b) : \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0^{(\eta)}\right\|_1 + \|(\boldsymbol{t}, \boldsymbol{\beta}, b) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0^{(\eta)}, b_0^{(\eta)})\|_2 \leq r\right\} \tag{9}$$

Throughout, we will be taking the entropy with respect to the empirical norm

$$\|g\|_{P_n} = \left(\frac{1}{n}\sum_{i=1}^n g^2(x_i, y_i)\right)^{1/2}$$

for some given set of $n$ observations $x^n, y^n$.

The following lemma bounds the entropy of $\mathcal{G}_r$.

**Lemma 5.** *For any $r, M_n > 0$, the following holds for all $u > 0$*

$$H\left(u, \mathcal{G}_r, \|\cdot\|_{P_n}\right) \leq c_0 \left(m + \frac{\left(rX_{max}M_n m\right)^2}{u^2}\right) \log\left(\frac{rpM_n m}{u}\right) \tag{10}$$

*for some absolute constant $c_0 > 0$.*

*Proof.* Let $\boldsymbol{\eta}_0 \in EQ_0$, $\Theta_{r,\eta_0} = \left\{\boldsymbol{\eta} \in \Theta_r : \boldsymbol{\eta}_0^{(\eta)} = \boldsymbol{\eta}_0\right\}$, and

$$\mathcal{G}_{r,\eta_0} = \left\{\left\{\ell_\eta(x,y) - \ell_0(x,y)\right\} 1\left\{G(y) \leq M_n\right\} : \boldsymbol{\eta} \in \Theta_{r,\eta_0}\right\}.$$

Then $\Theta_r = \cup_{\eta_0 \in EQ_0} \Theta_{r,\eta_0}$ and $\mathcal{G}_r = \cup_{\eta_0 \in EQ_0} \mathcal{G}_{r,\eta_0}$. Thus we can bound the covering number of $\Theta_r$ by

$$N\left(u, \mathcal{G}_r, \|\cdot\|_{P_n}\right) \leq \sum_{\eta_0 \in EQ_0} N\left(u, \mathcal{G}_{r,\eta_0}, \|\cdot\|_{P_n}\right) \tag{11}$$

Thus we will focus on bounding $N\left(u, \mathcal{G}_{r,\eta_0}, \|\cdot\|_{P_n}\right)$.

Let $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}, b)$ and $\boldsymbol{\eta}' = \left(\boldsymbol{\theta}', \boldsymbol{t}', \boldsymbol{\beta}', b'\right)$ and suppose $\eta, \eta' \in \Theta_{r,\eta_0}$. By the Mean Value Theorem and Lemma 4, we have

$$\begin{aligned}
&\left|\ell_\eta(x,y) - \ell_{\eta'}(x,y)\right| 1\left\{G(y) \leq M_n\right\} \\
&\leq M_n \left(\sum_{k=1}^m \left|\left(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k'\right)^\top x\right| + \|\boldsymbol{t} - \boldsymbol{t}'\|_1 + \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1 + |b - b'|\right)
\end{aligned}$$

Squaring both sides and applying Cauchy Schwarz, we get

$$\begin{aligned}
&\left|\ell_\eta(x,y) - \ell_{\eta'}(x,y)\right|^2 1\left\{G(y) \leq M_n\right\} \\
&\leq M_n^2(3m+1) \left(\sum_{k=1}^m \left|\left(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k'\right)^\top x\right|^2 + \|(\boldsymbol{t}, \boldsymbol{\beta}, b) - (\boldsymbol{t}', \boldsymbol{\beta}', b')\|_2^2\right)
\end{aligned} \tag{12}$$

Thus to bound the entropy of $\mathcal{G}_{r,\eta_0}$, it suffices to bound the entropy of $\{(\boldsymbol{t}, \boldsymbol{\beta}, b) : \|(\boldsymbol{t}, \boldsymbol{\beta}, b) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\|_2 \leq r\}$ and $\left\{f(x) = \left(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\right)^\top x : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \leq r\right\}$ for $k = 1, ..., m$.

The entropy of a ball with radius $r$ in $\mathbb{R}^{2m+1}$ is for all $u \geq 0$,

$$H\left(u, \left\{(\boldsymbol{t}, \boldsymbol{\beta}, b) \in \mathbb{R}^{2m+1} : \|(\boldsymbol{t}, \boldsymbol{\beta}, b) - (\boldsymbol{t}_0, \boldsymbol{\beta}_0, b_0)\|_2 \leq r\right\}, \|\cdot\|_2\right) \leq (2m+1) \log\left(\frac{5r}{u}\right).$$

By Lemma 14.29 in Bühlmann & Van De Geer (2011), the entropy of $\left(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\right)^\top x$ is for all $u \geq 0$

$$H\left(u, \left\{\left(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\right)^\top x : \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\|_1 \leq r\right\}, \|\cdot\|_2\right) \leq \left(\frac{\left(rX_{max}\right)^2}{u^2} + 1\right) \log\left(1 + p\right).$$

Putting these bounds together, we have the entropy bound for all $u \geq 0$

$$H\left(u, \mathcal{G}_{r,\eta_0}, \|\cdot\|_{P_n}\right)$$
$$\leq \left(2m + \frac{\left(rX_{max}M_n\right)^2 m\left(3m+1\right)}{u^2} + 1\right)\left(\log(1+p) + \log\left(\frac{5rM_n\sqrt{(3m+1)}}{u}\right)\right)$$

Then by (11) and the fact that $EQ_0$ has at most $2^m m!$ elements, we have our result. □

Next we need to show that the symmetrized truncated empirical process term is small with high probability. We use the Rademacher random variables $W$, which are defined to have distribution $Pr(W=1) = Pr(W=-1) = 0.5$.

**Lemma 6.** *Let $W_1, ..., W_n$ be $n$ independent Rademacher random variables. Let*

$$\delta = \frac{c_3}{\sqrt{n}} rM_n m^{3/2} \log(nmM_n)\sqrt{\log\left(p \vee nm\right)}. \tag{13}$$

*Then for all $T \geq 1$, we have*

$$Pr_W\left(\sup_{\eta\in\Theta_r}\left|\frac{1}{n}\sum_{i=1}^{n} W_i\left[\ell_\eta\left(x_i, y_i\right) - \ell_0\left(x_i, y_i\right)\right]\right| \geq c_3 T\delta\right)$$
$$\leq c_4 \exp\left(-\frac{T^2 m \log^2(nmM_n)\log(p\vee nm)\left(r^2\vee 1\right)}{c_5}\right).$$

*Proof.* We apply Lemma 3.2 in Geer (2000). First we check all the conditions are satisfied. By Taylor expansion, it is easy to show that for any $r > 0$, we have

$$\sup_{\eta\in\Theta_r}\frac{1}{n}\sum_{i=1}^{n}\left|\ell_\eta(x_i,y_i) - \ell_{\eta'}(x_i,y_i)\right|^2 1\left\{G(y_i)\leq M_n\right\} \leq c_1 M_n^2\left(X_{max}+m\right)^2\left[r^2\wedge 1\right]. \tag{14}$$

Set $R_n^2$ equal to the right hand side of (14). Then we can bound Dudley's integral in Lemma 3.2 as follows

$$\int_{r/n}^{R_n} H^{1/2}\left(u, \mathcal{G}_r, \|\cdot\|_n\right) du \leq c_2 rM_n m^{3/2}\log(nmM_n)\sqrt{\log\left(p\vee nm\right)}.$$

Let

$$\delta = \frac{c_3}{\sqrt{n}} rM_n m^{3/2} \log(nmM_n)\sqrt{\log\left(p \vee nm\right)}.$$

For any $T \geq 1$, we have

$$Pr\left(\sup_{\eta\in\Theta_r}\left|\frac{1}{n}\sum_{i=1}^{n} W_i\left[\ell_\eta\left(x_i, y_i\right) - \ell_{\eta_0}\left(x_i, y_i\right)\right]\right| \geq T\delta\right)$$
$$\leq c_4 \exp\left(-\frac{T^2 m \log^2(nmM_n)\log(p\vee nm)\left(r^2\vee 1\right)}{c_5}\right).$$

□

Using Lemma 6 above, combined with a symmetrization lemma like Corollary 3.4 in Geer (2000), we have the following corollary.

**Corollary 1.** *Let $\delta$ be defined as in* (13)*. Then for all $T \geq 1$, we have*

$$Pr\left(\sup_{\eta\in\Theta_r}\left|v_n^{trunc}(\eta)-v_n^{trunc}(\eta_0)\right|\geq c_6T\delta\right)$$
$$\leq c_7\exp\left(-\frac{T^2m\log^2(nmM_n)\log(p\vee nm)\left(r^2\vee 1\right)}{c_8}\right).$$

Now we are ready to bound the scaled truncated empirical process over the entire parameter space $\Theta$.

**Lemma 7.** *Let*

$$\tilde{\lambda}=\frac{c_9}{\sqrt{n}}M_nm^{3/2}\log(nmM_n)\sqrt{\log\left(p\vee nm\right)}.$$

*Then for any $T \geq 1$, we have*

$$Pr\left(\sup_{\eta\in\Theta}\frac{\left|v_n^{trunc}(\boldsymbol{\eta})-v_n^{trunc}(\boldsymbol{\eta}_0^{(\eta)})\right|}{\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2\right)\vee\tilde{\lambda}}\geq Tc\tilde{\lambda}\right)$$
$$\leq c_{13}\log n\exp\left(-\frac{T^2m\log^2(nmM_n)\log(p\vee nm)}{c_{14}}\right).$$

*Proof.* We use a peeling argument by partitioning $\Theta$ into

$$\Theta=\left[\cup_{j=J}^{-\infty}\Theta_j\right]\cup\Theta_{J+1}$$

where $J=\max\left\{j:2^{-j-1}\geq\tilde{\lambda}\right\}$. For $j=J,...,-\infty$, let

$$\Theta_j=\left\{\boldsymbol{\eta}\in\Theta:2^{-j-1}<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2\leq 2^{-j}\right\}$$

and

$$\Theta_{J+1}=\left\{\boldsymbol{\eta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2\leq 2^{-J-1}\right\}.$$

Then using a peeling argument, we have

$$
\begin{aligned}
&Pr\left(\sup_{\eta\in\Theta}\frac{|v_n^{trunc}(\boldsymbol{\eta})-v_n^{trunc}(\boldsymbol{\eta}_0)|}{\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2\right)\vee\tilde{\lambda}}\geq T\tilde{\lambda}\right)\\
&\leq\sum_{j=J+1}^{-\infty}Pr\left(\sup_{\eta\in\Theta_j}\frac{|v_n^{trunc}(\boldsymbol{\eta})-v_n^{trunc}(\boldsymbol{\eta}_0)|}{\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2\right)\vee\tilde{\lambda}}\geq T\tilde{\lambda}\right)\\
&\leq(J+2)c_9\exp\left(-\frac{T^2m\log^2(nmM_n)\log(p\vee nm)}{c_{10}}\right)\\
&+\sum_{j=1}^{\infty}c_{11}\exp\left(-\frac{T^2m\log^2(nmM_n)\log(p\vee nm)2^{2j}}{c_{12}}\right)\\
&\leq c_{13}\log n\exp\left(-\frac{T^2m\log^2(nmM_n)\log(p\vee nm)}{c_{14}}\right)
\end{aligned}
$$

where we used the fact that $J\leq c_{15}\log n$.

□

Finally, we control the remainder term

$$v_n^{rem}(\boldsymbol{\eta})=v_n(\boldsymbol{\eta})-v_n^{trunc}(\boldsymbol{\eta}) \tag{15}$$

using properties of sub-gaussian random variables.

**Lemma 8.** *Suppose $\epsilon$ are independent sub-gaussian random variables. If $M_n=\sqrt{c\log n}$ for some $c>0$ that only depends on the sub-gaussian parameters, we have*

$$Pr\left(\frac{\left|v_n^{rem}(\boldsymbol{\eta})-v_n^{rem}(\boldsymbol{\eta}_0^{(\eta)})\right|1\left\{G(y)>M_n\right\}}{\left\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\right\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2}\geq\frac{1}{\sqrt{n}}\right)\leq O_p\left(\frac{1}{n}\right) \tag{16}$$

*Proof.* By Taylor expansion, we have that

$$
\begin{aligned}
&|\ell_\eta(x,y)-\ell_{\eta_0}(x,y)|\,1\left\{G(y)>M_n\right\}\\
&\leq G(y)1\left\{G(y)>M_n\right\}\left(X_{max}+\sqrt{m+1}\right)\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{(\eta)}\right\|_1+\|(\boldsymbol{t},\boldsymbol{\beta},b)-(\boldsymbol{t}_0^{(\eta)},\boldsymbol{\beta}_0^{(\eta)},b_0^{(\eta)})\|_2\right).
\end{aligned} \tag{17}
$$

Let $F_{sup}^*=\sup_{x\in\mathcal{X}}|f^*(x)|$. Since

$$
\begin{aligned}
G(y)&=\left(\frac{K}{4}\vee 1\right)\left|K\sqrt{m+1}+|y|\right| &(18)\\
&\leq\left(\frac{K}{4}\vee 1\right)\left(\left|K\sqrt{m+1}+F_{sup}^*\right|+|\epsilon|\right), &(19)
\end{aligned}
$$

it suffices to bound

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} 2\left|K\sqrt{m+1}+F^{*}_{sup}\right| + |\epsilon_i|1\{G(y_i) > M_n\} + E\left[|\epsilon_i|1\{G(y_i) > M_n\}\right] \geq \frac{c}{\sqrt{n}}\right) \tag{20}$$

for any constant $c > 0$. This can be done using Markov's inequality and the fact that sub-gaussian random variables satisfy

$$E\left[|Z|\,1\left\{|Z| \geq M\right\}\right] \leq C' \exp\left(-c' M^2\right) \tag{21}$$

for constants $C', c' > 0$ that only depend on the sub-gaussian parameters. □

Finally, the proof for Theorem 2 simply combines the results in Lemmas 7 and 8.

# References


Albertini, Francesca, Sontag, Eduardo D, and Maillot, Vincent. Uniqueness of weights for neural networks. *Artificial Neural Networks for Speech and Vision*, pp. 115–125, 1993.

Anthony, Martin and Bartlett, Peter L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Bühlmann, Peter and Van De Geer, Sara. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Geer, Sara A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Städler, Nicolas, Bühlmann, Peter, and Van De Geer, Sara. 1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

van der Vaart, Aad W and Wellner, Jon A. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.