## A. Proofs of Section 2

We prove the theorem in a general manner, where we assume that each vector $w^*$ is $s$-sparse, that is, it only has $s$ non-zero coordinates. To prove Thm. 2.3 we assign $s = d$.

We start by showing that the function $x \mapsto \alpha x_i$ can be approximated by prunning a two-layer network:

**Lemma A.1.** *Let $s \in [d]$, and fix some scalar $\alpha \in [-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}]$, index $i \in [d]$, and some $\epsilon, \delta > 0$. Let $w^{(1)}, \ldots, w^{(k)} \in \mathbb{R}^d$ chosen randomly from $U([-1,1]^d)$, and $u^{(1)}, \ldots, u^{(k)} \in [-1,1]$ chosen randomly from $U([-1,1])$. Then, for $k \geq \frac{4}{\epsilon^2}\log(\frac{2}{\delta})$, w.p at least $1-\delta$ there exists a binary mask $b^{(1)}, \ldots, b^{(k)} \in \{0,1\}^d$, such that $g(x) = \sum_j u^{(j)}\sigma(\langle w^{(j)} \odot b^{(j)}, x\rangle)$ satisfies $|g(x) - \alpha x_i| \leq 2\epsilon$, for $\|x\|_\infty \leq 1$. Furthermore, we have $\sum_j \|b^{(j)}\|_0 \leq 2$ and $\max_j \|b^{(j)}\|_0 \leq 1$.*

*Proof.* If $|\alpha| \leq \epsilon$ then choosing $b^{(1)} = \cdots = b^{(k)} = (0, \ldots, 0)$ gives the required. Assume $|\alpha| \geq \epsilon$, and assume w.l.o.g that $\alpha > 0$. Fix some $j \in [k']$. Note that:

$$\mathbb{P}\left[|w_i^{(j)} - \alpha| \leq \epsilon \wedge |u^{(j)} - 1| \leq \epsilon\right] = \mathbb{P}\left[|w_i^{(j)} - \alpha| \leq \epsilon\right]\mathbb{P}\left[|u^{(j)} - 1| \leq \epsilon\right] = \frac{\epsilon}{2}\cdot\frac{\epsilon}{2} = \frac{\epsilon^2}{4}\,,$$

and similarly $\mathbb{P}\left[|w_i^{(j)} + \alpha| \leq \epsilon \wedge |u^{(j)} + 1| \leq \epsilon\right] \leq \frac{\epsilon^2}{4}$. Therefore, we have:

$$\mathbb{P}\left[\nexists j \in [k]\ s.t\ |w_i^{(j)} - \alpha| \leq \epsilon \wedge |u^{(j)} - 1| \leq \epsilon\right] = \left(1 - \frac{\epsilon^2}{4}\right)^k \leq \exp\left(-\frac{k\epsilon^2}{4}\right) \leq \frac{\delta}{2}\,,$$

where we used the assumption that $k \geq \frac{4}{\epsilon^2}\log(\frac{2}{\delta})$, and similarly:

$$\mathbb{P}\left[\nexists j \in [k']\ s.t\ |w_i^{(j)} + \alpha| \leq \epsilon \wedge |u^{(j)} + 1| \leq \epsilon\right] \leq \frac{\delta}{2}\,.$$

Therefore, using the union bound, w.p at least $1-\delta$ there exist $j, j'$ such that $|w_i^{(j)} - \alpha| \leq \epsilon, |u^{(j)} - 1| \leq \epsilon$ and $|w_i^{(j')} + \alpha| \leq \epsilon, |u^{(j')} + 1| \leq \epsilon$ and since $|\alpha| \geq \epsilon$ we get $j \neq j'$. Now, setting $b_i^{(j)} = 1, b_i^{(j')} = 1$, and the rest to zero, we get that:

$$g(x) = u^{(j)}\sigma(w_i^{(j)}x_i) + u^{(j')}\sigma(w_i^{(j')}x_i)$$

We will use the fact that $\sigma(a) - \sigma(-a) = a$ for every $a \in \mathbb{R}$. If $x_i \geq 0$, we get that $g(x) = u^{(j)}w_i^{(j)}x_i$ and therefore:

$$|g(x) - \alpha x_i| = |x_i||u^{(j)}w_i^{(j)} - \alpha| \leq |u^{(j)}w_i^{(j)} - u_j\alpha| + |u^{(j)}\alpha - \alpha| \leq |u^{(j)}||w_i^{(j)} - \alpha| + |u^{(j)} - 1||\alpha| \leq 2\epsilon$$

In a similar fashion, we get that for $x_i < 0$ we have $|g(x) - \alpha x_i| = |x_i||u^{(j')}w_i^{(j')} - \alpha| \leq 2\epsilon$, which gives the required. Since we have $\|b^{(j)}\|_0 = 1$, $\|b^{(j')}\|_0 = 1$ and $\|b^{(j'')}\|_0 = 0$ for every $j'' \neq j, j'$, the mask achieves the required. □

Using the previous result, we can show that a linear function $x \mapsto \langle w^*, x\rangle$ can be implemented by pruning a two layer network:

**Lemma A.2.** *Let $s \in [d]$, and fix some $w^* \in [-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}]^d$ with $\|w^*\|_0 \leq s$, and some $\epsilon, \delta > 0$. Let $w^{(1)}, \ldots, w^{(k)} \in \mathbb{R}^d$ chosen randomly from $U([-1,1]^d)$, and $u \in [-1,1]^k$ chosen randomly from $U([-1,1]^k)$. Then, for $k \geq s \cdot \left\lceil \frac{16s^2}{\epsilon^2}\log(\frac{2s}{\delta})\right\rceil$, w.p at least $1-\delta$ there exists a binary mask $b^{(1)}, \ldots, b^{(k)} \in \{0,1\}^d$, such that $g(x) = \sum_{i=1}^k u_i\sigma(\langle w^{(i)} \odot b^{(i)}, x\rangle)$ satisfies $|g(x) - \langle w^*, x\rangle| \leq \epsilon$, for $\|x\|_\infty \leq 1$. Furthermore, we have $\sum_i \|b^{(i)}\|_0 \leq 2s$ and $\max_i \|b^{(i)}\|_0 \leq 1$.*

*Proof.* We assume $k = s \cdot \left\lceil \frac{16s^2}{\epsilon^2}\log(\frac{2s}{\delta})\right\rceil$ (otherwise, mask excessive neurons), and let $k' := \frac{k}{s}$. With slight abuse of notation, we denote $w^{(i,j)} := w^{(j+k'i)}$, $u^{(i,j)} := u_{j+k'i}$ and $b^{(i,j)} := b^{(j+k'i)}$. Let $I := \{i \in [d] : w_i^* \neq 0\}$. By the assumption on $w^*$ we have $|I| \leq s$, and we assume w.l.o.g. that $I \subseteq [s]$. Fix some $i \in [s]$, and denote $g_i(x) = \sum_j u^{(i,j)}\sigma(\langle w^{(i,j)} \odot b^{(i,j)}, x\rangle)$. Let $\epsilon' = \frac{\epsilon}{2s}$ and $\delta' = \frac{\delta}{s}$, then from Lemma A.1, with probability at least $1-\delta'$ there exists a binary mask $b^{(i,1)}, \ldots, b^{(i,k')} \in \{0,1\}^d$ with $\sum_j \|b^{(i,j)}\|_0 \leq 2$ such that $|g_i(x) - w_i^*x_i| \leq 2\epsilon' = \frac{\epsilon}{s}$ for every

$x \in \mathbb{R}^d$ with $\|x\|_\infty \leq 1$. Now, using the union bound we get that with probability at least $1-\delta$, the above holds for all $i \in [s]$, and so:

$$|g(x) - \langle w^*, x\rangle| = |\sum_{i\in[s]} g_i(x) - \sum_{i\in[s]} w_i^* x_i| \leq \sum_{i\in[s]} |g_i(x) - w_i^* x_i| \leq \epsilon$$

Furthermore, we have $\sum_{i\in[s]} \sum_j \|b^{(i,j)}\|_0 \leq 2s$ and $\max_{i,j} \|b^{(i,j)}\|_0 \leq 1$, by the result of Lemma A.1. □

Now, we can show that a network with a single neuron can be approximated by prunning a three-layer network:

**Lemma A.3.** *Let $s \in [d]$, and fix some $w^* \in [-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}]^d$ with $\|w^*\|_0 \leq s$, some $v^* \in [-1,1]$ and some $\epsilon, \delta > 0$. Let $w^{(1)}, \ldots, w^{(k_1)} \in \mathbb{R}^d$ chosen randomly from $U([-1,1]^d)$, $u^{(1)}, \ldots, u^{(k_2)} \in [-1,1]^{k_1}$ chosen randomly from $U([-1,1]^{k_1})$, and $v \in [-1,1]^{k_2}$ chosen randomly from $U([-1,1]^{k_2})$. Then, for $k_1 \geq s \cdot \left\lceil \frac{64s^2}{\epsilon^2} \log(\frac{4s}{\delta}) \right\rceil$, $k_2 \geq \frac{2}{\epsilon}\log(\frac{2}{\delta})$, w.p at least $1-\delta$ there exists a binary mask $b^{(1)}, \ldots, b^{(k_1)} \in \{0,1\}^d$, $\hat{b} \in \{0,1\}^{k_2}$, such that $g(x) = \sum_{i=1}^{k_2} \hat{b}_i v_i \sigma(\sum_{j=1}^{k_1} u_j^{(i)} \sigma(\langle w^{(j)} \odot b^{(j)}, x\rangle))$ satisfies $|g(x) - v^*\sigma(\langle w^*, x\rangle)| \leq \epsilon$, for $\|x\|_2 \leq 1$. Furthermore, we have $\sum_j \|b^{(j)}\|_0 \leq 2s$ and $\max_j \|b^{(j)}\|_0 \leq 1$.*

*Proof.* Let $\epsilon' = \frac{\epsilon}{2}$, and note that for every $i \in [k_2]$ we have $\mathbb{P}[|v_i - v^*| \leq \epsilon'] \geq \epsilon'$. Therefore, the probability that for some $i \in [k_2]$ it holds that $|v_i - v^*| \leq \epsilon'$ is at least $1 - (1-\epsilon')^{k_2} \geq 1 - e^{-k_2\epsilon'} \geq 1 - \frac{\delta}{2}$, where we use the fact that $k_2 \geq \frac{1}{\epsilon'}\log(\frac{2}{\delta})$. Now, assume this holds for $i \in [k_2]$. Let $\hat{b}_j = \mathbb{1}\{j = i\}$, and so:

$$g(x) = v_i \sigma(\sum_{j=1}^{k_1} u_j^{(i)} \sigma(\langle w^{(j)} \circ b^{(j)}, x\rangle)$$

Then, from Lemma A.2, with probability at least $1 - \frac{\delta}{2}$ there exists $b^{(1)}, \ldots, b^{(k_1)}$ s.t. for every $\|x\|_\infty \leq 1$:

$$\left| \sum_{j=1}^{k_1} u_j^{(i)} \sigma(\langle w^{(j)} \circ b^{(j)}, x\rangle - \langle w^*, x\rangle) \right| \leq \epsilon'$$

And therefore, for every $\|x\|_2 \leq 1$:

$$\begin{aligned}
&|g(x) - v^*\sigma(\langle w^*, x\rangle)| \\
&\leq |v_i| \left| \sigma(\sum_{j=1}^{k_1} u_j^{(i)} \sigma(\langle w^{(j)} \circ b^{(j)}, x\rangle) - \sigma(\langle w^*, x\rangle)) \right| + |v_i - v^*| \, |\sigma(\langle w^*, x\rangle)| \\
&\leq |v_i| \left| \sum_{j=1}^{k_1} u_j^{(i)} \sigma(\langle w^{(j)} \circ b^{(j)}, x\rangle - \langle w^*, x\rangle) \right| + |v_i - v^*| \, \|w^*\| \|x\| \leq 2\epsilon' = \epsilon
\end{aligned}$$

□

Finally, we show that pruning a three-layer network can approximate a network with $n$ neurons, since it is only a sum of networks with 1 neuron, as analyzed in the previous lemma:

**Lemma A.4.** *Let $s \in [d]$, and fix some $w^{(1)*}, \ldots, w^{(n)*} \in [-1,1]^d$ with $\|w^{(i)*}\|_0 \leq s$, $v^* \in [-1,1]^n$ and let $f(x) = \sum_{i=1}^n v_i^* \sigma(\langle w^{(i)*}, x\rangle)$. Fix some $\epsilon, \delta > 0$. Let $w^{(1)}, \ldots, w^{(k_1)} \in \mathbb{R}^d$ chosen randomly from $U([-1,1]^d)$, $u^{(1)}, \ldots, u^{(k_2)} \in [-1,1]^{k_1}$ chosen randomly from $U([-1,1]^{k_1})$, and $v \in [-1,1]^{k_2}$ chosen randomly from $U([-1,1]^{k_2})$. Then, for $k_1 \geq ns \cdot \left\lceil \frac{64s^2n^2}{\epsilon^2} \log(\frac{4ns}{\delta}) \right\rceil$, $k_2 \geq \frac{2n}{\epsilon}\log(\frac{2n}{\delta})$, w.p at least $1-\delta$ there exists a binary mask $b^{(1)}, \ldots, b^{(k_1)} \in \{0,1\}^d$, $\tilde{b}^{(1)}, \ldots, \tilde{b}^{(k_2)} \in \{0,1\}^{k_1}$, $\hat{b} \in \{0,1\}^{k_2}$, such that $g(x) = \sum_{i=1}^{k_2} \hat{b}_i v_i \sigma(\sum_{j=1}^{k_1} \tilde{b}_j^{(i)} u_j^{(i)} \sigma(\langle w^{(j)} \odot b^{(j)}, x\rangle))$ satisfies $|g(x) - f(x)| \leq \epsilon$, for $\|x\|_2 \leq 1$. Furthermore, we have $\sum_j \|b^{(j)}\|_0 \leq 2s$ and $\max_j \|b^{(j)}\|_0 \leq 1$.*

*Proof.* Denote $k_1' = \frac{k_1}{n}, k_2' = \frac{k_2}{n}$ and assume $k_1', k_2' \in \mathbb{N}$ (otherwise mask exceeding neurons). With slight abuse of notation, we denote $w^{(i,j)} := w^{(j+k_1'i)}$, $u^{(i,j)} := \left(u_{ik_1'}^{(j+ik_2')}, \ldots, u_{(i+1)k_1'}^{(j+ik_2')}\right)$, $v^{(i,j)} := v_{j+ik_2'}$ and similarly $b^{(i,j)} := b^{(j+k_1'i)}$,

$\tilde{b}^{(i,j)} = \left(\tilde{b}_{ik_1'}^{(j+ik_2')}, \ldots, \tilde{b}_{(i+1)k_1'}^{(j+ik_2')}\right)$ and $\hat{b}^{(i,j)} = \hat{b}_{j+ik_2'}$. Define for every $i \in [n]$:

$$g_i(x) = \sum_j \hat{b}^{(i,j)} v^{(i,j)} \sigma\left(\sum_l \tilde{b}_l^{(i,j)} u_l^{(i,j)} \sigma(\langle b^{(i,l)} \circ w^{(i,l)}, x\rangle)\right)$$

Now, by setting $\tilde{b}_l^{(j+k_1'i)} = \mathbb{1}\{ik_1' \le l < (i+1)k_1'\}$ we get that $g(x) = \sum_{i=1}^n g_i(x)$. Now, from Lemma A.3 we get that with probability at least $1 - \frac{\delta}{n}$ we have $\left|g_i(x) - v_i^* \sigma(\langle w^{(i)*}, x\rangle)\right| \le \frac{\epsilon}{n}$ for every $\|x\|_2 \le 1$. Using the union bound, we get that with probability at least $1-\delta$, for $\|x\|_2 \le 1$ we have $|g(x) - f(x)| \le \sum_{i=1}^n \left|g_i(x) - v_i^* \sigma(\langle w^{(i)*}, x\rangle)\right| \le \epsilon$. □

*Proof.* of Theorem 2.3.

From Lemma A.4 with $s = d$. □

In a similar fashion, we can prove a result for deep networks. We start by showing that a single layer can be approximated by pruning:

**Lemma A.5.** *Let $s \in [d]$, and fix some $w^{(1)*}, \ldots, w^{(n)*} \in [-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}]^d$ with $\|w^{(i)*}\|_0 \le s$ and let $F : \mathbb{R}^d \to \mathbb{R}^n$ such that $F(x)_i = \sigma(\langle w^{(i)*}, x\rangle)$. Fix some $\epsilon, \delta > 0$. Let $w^{(1)}, \ldots, w^{(k)} \in \mathbb{R}^d$ chosen randomly from $U([-1,1]^d)$ and $u^{(1)}, \ldots, u^{(n)} \in [-1,1]^k$ chosen randomly from $U([-1,1]^k)$. Then, for $k \ge ns \cdot \left\lceil \frac{16s^2 n}{\epsilon^2} \log(\frac{2ns}{\delta}) \right\rceil$, w.p at least $1-\delta$ there exists a binary mask $b^{(1)}, \ldots, b^{(k)} \in \{0,1\}^d$, $\tilde{b}^{(1)}, \ldots, \tilde{b}^{(n)} \in \{0,1\}^{k_1}$, $\hat{b} \in \{0,1\}^k$, such that for $G : \mathbb{R}^d \to \mathbb{R}^n$ with $G(x)_i = \sigma(\sum_{j=1}^k \tilde{b}_j^{(i)} u_j^{(i)} \sigma(\langle w^{(j)} \odot b^{(j)}, x\rangle))$ we have $\|G(x) - F(x)\|_2 \le \epsilon$, for $\|x\|_\infty \le 1$. Furthermore, we have $\sum_j \|b^{(j)}\|_0 \le 2sn$ and $\sum_i \|\tilde{b}^{(i)}\|_0 \le 2sn$.*

*Proof.* Denote $k' = \frac{k}{n}$ and assume $k' \in \mathbb{N}$ (otherwise mask exceeding neurons). With slight abuse of notation, we denote $w^{(i,j)} := w^{(j+k'i)}$, $b^{(i,j)} := b^{(j+k'i)}$ and we denote $\tilde{u}^{(i)} := \left(u_{ik'}^{(i)}, \ldots, u_{(i+1)k'}^{(i)}\right)$. Define for every $i \in [n]$:

$$g_i(x) = \sum_j \tilde{u}_j^{(i)} \sigma(\langle b^{(i,j)} \circ w^{(i,j)}, x\rangle)$$

Now, by setting $\tilde{b}_l^{(j+k_1'i)} = \mathbb{1}\{ik_1' \le l < (i+1)k_1'\}$ we get that $G(x)_i = \sigma(g_i(x))$. Now, from Lemma A.2 with $\epsilon' = \frac{\epsilon}{\sqrt{n}}$ and $\delta' = \frac{\delta}{n}$, since $k \ge s \cdot \left\lceil \frac{16s^2}{(\epsilon')^2} \log(\frac{2s}{\delta'}) \right\rceil$ we get that with probability at least $1 - \frac{\delta}{n}$ we have $\left|g_i(x) - \langle w^{(i)*}, x\rangle\right| \le \frac{\epsilon}{\sqrt{n}}$ for every $\|x\|_\infty \le 1$. Using the union bound, we get that with probability at least $1-\delta$, for $\|x\|_\infty \le 1$ we have:

$$\|G(x) - F(x)\|_2^2 = \sum_i (\sigma(g_i(x)) - \sigma(\langle w^{(i)*}, x\rangle))^2 \le \sum_i (g_i(x) - \langle w^{(i)*}, x\rangle)^2 \le \epsilon^2$$

Notice that Lemma A.2 also gives $\sum_j \|b^{(i,j)}\|_0 \le 2s$ and so $\sum_{i=1}^n \sum_j \|b^{(i,j)}\|_0 \le 2sn$. Since we can set $\tilde{b}_j^{(i)} = 0$ for every $i, j$ with $b^{(i,j)} = 0$, we get the same bound on $\sum_i \|\tilde{b}^{(i)}\|_0$. □

Using the above, we can show that a deep network can be approximated by pruning. We show this result with the assumption that each neuron in the network has only $s$ non-zero weights. To get a similar result without this assumption, as is stated in Thm. 2.1, we can simply choose $s$ to be its maximal value - either $d$ for the first layer of $n$ for intermediate layers.

**Theorem A.6.** *(formal statement of Thm. 2.1, when $s = \max\{n, d\}$). Let $s, n \in \mathbb{N}$, and fix some $W^{(1)*}, \ldots, W^{(l)*}$ such that $W^{(1)*} \in [-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}]^{d \times n}$, $W^{(2)*}, \ldots, W^{(l-1)*} \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]^{n \times n}$ and $W^{(l)*} \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]^{n \times 1}$. Assume that for every $i \in [l]$ we have $\|W^{(i)*}\|_2 \le 1$ and $\max_j \|W_j^{(i)}\|_0 \le s$. Denote $F^{(i)}(x) = \sigma(W^{(i)*} x)$ for $i < l$ and $F^{(l)}(x) = W^{(l)*} x$, and let $F(x) := F^{(l)} \circ \cdots \circ F^{(1)}(x)$. Fix some $\epsilon, \delta \in (0,1)$. Let $W^{(1)}, \ldots, W^{(l)}, U^{(1)}, \ldots, U^{(l)}$ such that $W^{(1)}$ is chosen randomly from $U([-1,1]^{d \times k})$, $W^{(2)}, \ldots, W^{(l)}$ is chosen randomly from $U([-1,1]^{n \times k})$, $U^{(1)}, \ldots, U^{(l-1)}$ chosen from $U([-1,1]^{k \times n})$ and $U^{(l)}$ chosen from $U([-1,1]^k)$. Then, for $k \ge ns \cdot \left\lceil \frac{64 s^2 l^2 n}{\epsilon^2} \log(\frac{2nsl}{\delta}) \right\rceil$, w.p. at least $1-\delta$ there exist $B^{(i)}$ a binary mask for $W^{(i)}$ with matching dimensions, and $\tilde{B}^{(i)}$ a binary mask for $U^{(i)}$ with matching dimensions, s.t.:*

$$|G(x) - F(x)| \le \epsilon \; for \; \|x\|_2 \le 1$$

*Where we denote* $G = G^{(l)} \circ \cdots \circ G^{(1)}$*, with* $G^{(i)}(x) := \sigma(\tilde{B}^{(i)} \circ U^{(i)}\sigma(B^{(i)} \circ W^{(i)}x))$ *for every* $i < l$ *and* $G^{(l)}(x) := \tilde{B}^{(l)} \circ U^{(l)}\sigma(B^{(l)} \circ W^{(l)}x)$*. Furthermore, we have* $\|B^{(i)}\|_0 \leq 2sn$ *and* $\|\tilde{B}^{(i)}\|_0 \leq 2sn$*.*

*Proof.* Fix some $i < l$. From A.5, with probability at least $1 - \frac{\delta}{l}$ there exists a choice for $\tilde{B}^{(i)}, B^{(i)}$ such that for every $\|x\|_\infty \leq 1$ we have $\|F^{(i)}(x) - G^{(i)}(x)\|_2 \leq \frac{\epsilon}{2l}$. Note that we want to show that every layer is well approximated given the output of the previous layer, which can slightly deviate from the output of the original network. So, we need to relax the condition of Lemma A.5 to $\|x\|_\infty \leq 2$ in order to allow these small deviations from the target network.

Notice that if $\|x\|_\infty \leq 2$, from homogeneity of $G^{(i)}, F^{(i)}$ to positive scalars we get that:

$$\|G^{(i)}(x) - F^{(i)}(x)\|_2 = 2\|G^{(i)}(\frac{1}{2}x) - F^{(i)}(\frac{1}{2}x)\|_2 \leq \frac{\epsilon}{l}$$

Similarly, from Lemma A.2, with probability at least $1-\frac{\delta}{l}$ it holds that $\left|F^{(l)}(x) - G^{(l)}(x)\right| \leq \frac{\epsilon}{l}$ for every $x$ with $\|x\|_\infty \leq 2$. Assume that all the above holds, and using the union bound this happens with probability at least $1-\delta$. Notice that for every $x$ we have $\|F^{(i)}(x)\|_2 \leq \|W^{(i)*}x\|_2 \leq \|W^{(i)*}\|_2\|x\|_2 \leq \|x\|_2$, and so $\|F^{(i)} \circ \cdots \circ F^{(1)}(x)\|_2 \leq \|F^{(i-1)} \circ \cdots \circ F^{(1)}(x)\|_2 \leq \cdots \leq \|x\|_2$. Fix some $x$ with $\|x\|_2 \leq 1$ and denote $x^{(i)} = F^{(i)} \circ \cdots \circ F^{(1)}(x)$ and $\hat{x}^{(i)} = G^{(i)} \circ \cdots \circ G^{(1)}(x)$. Now, we will show that $\|x^{(i)} - \hat{x}^{(i)}\|_2 \leq \frac{i\epsilon}{l}$ for every $i \leq l$, by induction on $i$. The case $i = 0$ is trivial, and assume the above holds for $i-1$. Notice that in this case we have $\|\hat{x}^{(i-1)}\|_\infty \leq \|\hat{x}^{(i-1)}\|_2 \leq \|x^{(i-1)}\|_2 + \|x^{(i-1)} - \hat{x}^{(i-1)}\|_2 \leq 2$. Therefore:

$$\begin{aligned}
\|x^{(i)} - \hat{x}^{(i)}\|_2 &= \|G^{(i)}(\hat{x}^{(i-1)}) - F^{(i)}(x^{(i-1)})\|_2 \\
&\leq \|G^{(i)}(\hat{x}^{(i-1)}) - F^{(i)}(\hat{x}^{(i-1)})\|_2 + \|F^{(i)}(\hat{x}^{(i-1)}) - F^{(i)}(x^{(i-1)})\|_2 \\
&\leq \frac{\epsilon}{l} + \|W^{(i)*}(\hat{x}^{(i-1)} - x^{(i-1)})\|_2 \leq \frac{\epsilon}{l} + \|W^{(i)*}\|_2\|\hat{x}^{(i-1)} - x^{(i-1)}\|_2 \leq \frac{i\epsilon}{l}
\end{aligned}$$

From the above, we get that $|F(x) - G(x)| = \|x^{(l)} - \hat{x}^{(l)}\|_2 \leq \epsilon$. □

## B. Proofs of Section 3

First we will need the following lemma, which intuitively shows a generalization bound over linear predictors, where each coordinate of each sample is pruned with equal probability and independently.

**Lemma B.1.** *Let* $k > 0$*, and* $v^{(1)}, \ldots, v^{(k)} \in [-1, 1]^d$*. Let* $\hat{v}^{(j)}$ *be Bernoulli random variables such that for each* $j$*, with probability* $\epsilon$ *we have* $\hat{v}(j) = \frac{1}{\epsilon}v^{(j)}$*, and with probability* $1 - \epsilon$ *we have* $\hat{v}^{(j)} = 0$*. Then we have w.p* $> 1 - \delta$ *that:*

$$\sup_{z:\|z\|\leq L} \left| \frac{1}{k}\sum_{j=1}^{k} \langle \hat{v}^{(j)}, z\rangle - \frac{1}{k}\sum_{j=1}^{k} \langle v^{(j)}, z\rangle \right| \leq \frac{L}{\epsilon\sqrt{k}} \left(3\sqrt{d} + \log\left(\frac{1}{\delta}\right)\right)$$

*Proof.* Note that for each $j \in [k]$ we have that $\mathbb{E}\left[\hat{v}^{(j)}\right] = v^{(j)}$, thus for every vector $z \in \mathbb{R}^d$, also $\mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k}\langle \hat{v}^{(j)}, z\rangle\right] = \frac{1}{k}\sum_{j=1}^{k}\langle v^{(j)}, z\rangle$. Hence, using a standard argument about Rademacher complexity (see (29) Lemma 26.2) we have that:

$$\begin{aligned}
&\mathbb{E}_{\hat{v}^{(1)},\ldots,\hat{v}^{(k)}} \left[ \sup_{z:\|z\|\leq L} \left| \frac{1}{k}\sum_{j=1}^{k} \langle \hat{v}^{(j)}, z\rangle - \frac{1}{k}\sum_{j=1}^{k} \langle v^{(j)}, z\rangle \right| \right] \\
\leq &\frac{2}{k}\mathbb{E}_{\hat{v}^{(1)},\ldots,\hat{v}^{(k)}}\mathbb{E}_{\xi_1,\ldots,\xi_k} \left[ \sup_{z:\|z\|\leq L} \sum_{j=1}^{k} \xi_j \langle \hat{v}^{(j)} - v^{(j)}, z\rangle \right] \qquad (3)
\end{aligned}$$

where $\xi_1, \ldots, \xi_k$ are standard Rademacher random variables. Set $\tilde{v}^{(j)} = \hat{v}^{(j)} - v^{(j)}$ ,using Cauchy-Schwartz we can bound Eq. (3) by:

$$\frac{2}{k}\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\mathbb{E}_{\xi_1,\ldots,\xi_k} \left[ \sup_{z:\|z\|\leq L} \|z\| \cdot \left\| \sum_{j=1}^{k} \xi_j \tilde{v}^{(j)} \right\| \right] \leq \frac{2L}{k}\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\mathbb{E}_{\xi_1,\ldots,\xi_k} \left[ \left\| \sum_{j=1}^{k} \xi_j \tilde{v}^{(j)} \right\| \right] . \qquad (4)$$

Next, we can use Jensen's inequality on Eq. (4) to bound it

$$\frac{2L}{k}\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\mathbb{E}_{\xi_1,\ldots,\xi_k}\left[\left\|\sum_{j=1}^{k}\xi_j\tilde{v}^{(j)}\right\|\right] \leq \frac{2L}{k}\sqrt{\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\mathbb{E}_{\xi_1,\ldots,\xi_k}\left[\left\|\sum_{j=1}^{k}\xi_j\tilde{v}^{(j)}\right\|^2\right]}$$
$$\leq \frac{2L}{k}\sqrt{\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\mathbb{E}_{\xi_1,\ldots,\xi_k}\left[\sum_{i=1}^{k}\sum_{j=1}^{k}\xi_i\xi_j\tilde{v}^{(i)\top}\tilde{v}^{(j)}\right]} = \frac{2L}{k}\sqrt{\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\left[\sum_{j=1}^{k}\|\tilde{v}^{(j)}\|^2\right]}.$$

Finally, using the fact that $\|\tilde{v}^{(j)}\|^2 \leq \|\hat{v}^{(j)}\|^2 + \|v^{(j)}\|^2 \leq \frac{1}{\epsilon^2}\|v^{(j)}\|^2 + \|v^{(j)}\|^2 \leq \frac{2d}{\epsilon^2}$ we have that:

$$\frac{2L}{k}\sqrt{\mathbb{E}_{\tilde{v}^{(1)},\ldots,\tilde{v}^{(k)}}\left[\sum_{j=1}^{k}\|\tilde{v}^{(j)}\|^2\right]} \leq \frac{3L\sqrt{d}}{\epsilon\sqrt{k}}$$

In order to prove the lemma we will use McDiarmid's inequality to get guarantees with high probability. Note that for every $l \in [k]$, by taking $\tilde{\hat{v}}^{(l)}$ instead of $\hat{v}^{(l)}$ we have for every $z$ with $\|z\| \leq L$ that:

$$\left|\frac{1}{k}\sum_{j=1}^{k}\langle\hat{v}^{(j)},z\rangle - \frac{1}{k}\left(\sum_{j\neq l}\langle\hat{v}^{(j)},z\rangle - \langle\tilde{\hat{v}}^{(l)},z\rangle\right)\right| \leq \frac{1}{k}\left|\langle\hat{v}^{(l)},z\rangle - \langle\tilde{\hat{v}}^{(l)},z\rangle\right| \leq \frac{L}{\epsilon k}$$

By using Mcdiarmid's theorem we get

$$\mathbb{P}\left(\sup_{z:\|z\|\leq L}\left|\frac{1}{k}\sum_{j=1}^{k}\langle\hat{v}^{(j)},z\rangle - \frac{1}{k}\sum_{j=1}^{k}\langle v^{(j)},z\rangle\right| \geq \frac{3L\sqrt{d}}{\epsilon k} + t\right) \leq \exp\left(-\frac{-2t^2\epsilon^2 k}{L^2}\right),$$

setting the r.h.s to $\delta$, and $t = \frac{\sqrt{\log\left(\frac{1}{\delta}\right)}L}{\epsilon\sqrt{k}}$ we have w.p $> 1-\delta$ that:

$$\sup_{z:\|z\|\leq L}\left|\frac{1}{k}\sum_{j=1}^{k}\langle\hat{v}^{(j)},z\rangle - \frac{1}{k}\sum_{j=1}^{k}\langle v^{(j)},z\rangle\right| \leq \frac{L}{\epsilon\sqrt{k}}\left(3\sqrt{d} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right).$$

□

Next, we show the main argument, which states that by pruning a neurons from a large enough 2-layer neural network, it can approximate any other 2-layer neural network for which the weights in the first layer are the same, and the weights in the second layer are bounded.

**Lemma B.2.** *Let $k_1 \in \mathbb{N}$ and $\epsilon, \delta, M > 0$ and assume that $\sigma$ is L-Lipschitz with $\sigma(0) \leq L$. Let $k_2 > \frac{256\log\left(\frac{2k_1}{\delta}\right)k_1^4L^4}{\epsilon^4}$, and for every $i \in [k_1],\ j \in [k_2]$ initialize $w_i^{(j)} \sim \mathcal{D}$ for any distribution $\mathcal{D}$ with $\mathbb{P}\left(\|w_i\| \leq 1\right) = 1$ and $u_i^{(j)} \sim U([-1,1])$. Let $v^{(1)},\ldots,v^{(k_2)} \in \mathbb{R}^{k_1}$ with $\|v^{(j)}\|_\infty \leq M$ for every $j \in [k_2]$, and define $f^{(j)}(x) = \sum_{i=1}^{k_1} v_i^{(j)}\sigma\left(\langle w_i^{(j)},x\rangle\right)$. Then there exist $b^{(1)},\ldots,b^{(k_2)} \in \{0,1\}^{k_1}$ such that for the functions $\tilde{g}^{(j)}(x) = \sum_{i=1}^{k_1} b_i^{(j)}\cdot u_i^{(j)}\sigma\left(\langle w_i^{(j)},x\rangle\right)$ w.p $> 1-\delta$ we have:*

$$\sup_{x:\|x\|\leq 1}\left|\frac{c_1}{k_2}\sum_{j=1}^{k_2}\tilde{g}^{(j)}(x) - \frac{1}{k_2 M}\sum_{j=1}^{k_2}f^{(j)}(x)\right| \leq \epsilon$$

*where $c_1 = \frac{8k_1L}{\epsilon}$*

*Proof.* Denote $\epsilon' = \frac{\epsilon}{4k_1L}$, and for $j \in [k_2]$ denote $\bar{v}^{(j)} = \frac{1}{M}v^{(j)}$, so we have $\|\bar{v}^{(j)}\|_\infty \leq 1$. Let $b_i^{(j)} = \mathbb{1}\left\{\left|u_i^{(j)} - \bar{v}_i^{(j)}\right| \leq \epsilon'\right\}$, note that the $b_i^{(j)}$-s are i.i.d Bernoulli random variables with $\mathbb{P}\left[b_i^{(j)} = 1\right] = \frac{\epsilon'}{2}$.

Set the following vectors: $\hat{v}^{(j)} = \frac{2}{\epsilon'}\begin{pmatrix} b_1^{(j)}\bar{v}_1^{(j)} \\ \vdots \\ b_{k_1}^{(j)}\bar{v}_{k_1}^{(j)} \end{pmatrix}$, $\hat{u}^{(j)} = \frac{2}{\epsilon'}\begin{pmatrix} b_1^{(j)}u_1^{(j)} \\ \vdots \\ b_{k_1}^{(j)}\bar{u}_{k_1}^{(j)} \end{pmatrix}$, and denote the function $z^{(j)} : \mathbb{R}^d \to \mathbb{R}^{k_1}$ with $z_i^{(j)}(x) = \sigma\left(\langle w_i^{(j)}, x\rangle\right)$. Now, the functions $f^{(j)}(x)$ can be written as $f^{(j)}(x) = \langle v^{(j)}, z^{(j)}(x)\rangle$, we denote

$$\tilde{g}(x) = \sum_{j=1}^{k_2}\sum_{i=1}^{k_1} b_i^{(j)}u_i^{(j)}\sigma\left(\langle w_i^{(j)}, x\rangle\right) = \sum_{j=1}^{k_2}\langle b^{(j)} \odot u^{(j)}, z^{(j)}(x)\rangle$$
$$\hat{g}(x) = \frac{2}{\epsilon'}\sum_{j=1}^{k_2}\sum_{i=1}^{k_1} b_i^{(j)}u_i^{(j)}\sigma\left(\langle w_i^{(j)}, x\rangle\right) = \sum_{j=1}^{k_2}\langle \hat{u}^{(j)}, z^{(j)}(x)\rangle.$$

Our goal is to bound the following, when the supremum is taken over $\|x\| \leq 1$:

$$\begin{aligned}
&\sup_x \left|\frac{c_1}{k_2}\tilde{g}(x) - \frac{1}{k_2 M}\sum_{j=1}^{k_2} f^{(j)}(x)\right| = \sup_x \left|\frac{1}{k_2}\hat{g}(x) - \frac{1}{k_2 M}\sum_{j=1}^{k_2} f^{(j)}(x)\right| \\
&= \sup_x \left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{u}^{(j)}, z^{(j)}(x)\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \bar{v}^{(j)}, z^{(j)}(x)\rangle\right| \\
&\leq \sup_x \left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{u}^{(j)}, z^{(j)}(x)\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{v}^{(j)}, z^{(j)}(x)\rangle\right| + \sup_x \left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{v}^{(j)}, z^{(j)}(x)\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \bar{v}^{(j)}, z^{(j)}(x)\rangle\right|
\end{aligned} \tag{5}$$

where $c_1 = \frac{2}{\epsilon'} = \frac{8k_1L}{\epsilon}$. We will now bound each expression in Eq. (5) with high probability. For the first expression, we first bound:

$$\begin{aligned}
\sup_x \left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{u}^{(j)}, z^{(j)}(x)\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{v}^{(j)}, z^{(j)}(x)\rangle\right| &= \sup_x \left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{u}^{(j)} - \hat{v}^{(j)}, z^{(j)}(x)\rangle\right| \\
&\leq \frac{1}{k_2}\sum_{j=1}^{k_2}\sup_x \left|\langle \hat{u}^{(j)} - \hat{v}^{(j)}, z^{(j)}(x)\rangle\right| .
\end{aligned}$$

Fix $i \in [k_1]$ and set $X_i^{(j)} := \sup_x \left|\left(\hat{u}_i^{(j)} - \hat{v}_i^{(j)}\right) \cdot z_i^{(j)}(x)\right|$ and note that for every $x$ with $\|x\| \leq 1$ we have that $\sup_x \left|z_i^{(j)}(x)\right| \leq 2L$. For the random variables $X_i^{(j)}$ we get:

- $X_i^{(j)} \leq \left|\hat{u}_i^{(j)} - \hat{v}_i^{(j)}\right| \cdot \sup_x \left|z_i^{(j)}(x)\right| \leq 4L$
- $\mathbb{E}\left[X_i^{(j)}\right] \leq 2\epsilon' L$

We now use Hoeffding's inequality to get that:

$$\mathbb{P}\left(\frac{1}{k_2}\sum_{j=1}^{k_2} X_i^{(j)} \geq 2\epsilon' L + t\right) \leq \exp\left(-\frac{t^2 k_2}{8L^2}\right).$$

Replacing the r.h.s with $\delta_1$ and setting $t = \epsilon' L$, we get that if $k_2 \geq \frac{8\log\left(\frac{1}{\delta_1}\right)}{\epsilon'^2}$ then w.p $1-\delta_1$:

$$\frac{1}{k_2}\sup_x \left|\left(\hat{u}_i^{(j)} - \hat{v}_i^{(j)}\right)\cdot z_i^{(j)}(x)\right| \leq 3\epsilon' L.$$

Setting $\delta_1 = \frac{\delta}{2k_1}$, and applying union bound for $i = 1,\ldots,k_1$ we get that w.p $> 1-\frac{\delta}{2}$ we have:

$$\frac{1}{k_2}\sum_{j=1}^{k_2}\sup_x \left|\langle \hat{u}^{(j)} - \hat{v}^{(j)}, z^{(j)}(x)\rangle\right| \leq 3k_1\epsilon' L. \tag{6}$$

For the second expression in Eq. (5) we first note that for all $j \in [k_2]$ we have $\max_{x:\|x\|\leq 1}\|z^{(j)}(x)\| \leq 2L\sqrt{k_1}$. Hence we can bound the second expression

$$\begin{aligned}&\sup_x \left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{v}^{(j)}, z^{(j)}(x)\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \bar{v}^{(j)}, z^{(j)}(x)\rangle\right| \\ \leq &\sum_{z\in\mathbb{R}^{k_1}:\|z\|\leq 2L\sqrt{k_1}}\left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{v}^{(j)}, z\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \bar{v}^{(j)}, z\rangle\right| .\end{aligned}$$

Using Lemma B.1 on the above term, w.p $> 1-\frac{\delta}{2}$ we have that:

$$\sum_{z\in\mathbb{R}^{k_1}:\|z\|\leq 2L\sqrt{k_1}}\left|\frac{1}{k_2}\sum_{j=1}^{k_2}\langle \hat{v}^{(j)}, z\rangle - \frac{1}{k_2}\sum_{j=1}^{k_2}\langle \bar{v}^{(j)}, z\rangle\right| \leq \frac{2L\sqrt{k_1}}{\epsilon'\sqrt{k_2}}\left(3\sqrt{k_1} + \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \tag{7}$$

Combining Eq. (6) with Eq. (7), applying union bound and taking $k_2 \geq \frac{256L^4k_1^4\log\left(\frac{2}{\delta}\right)}{\epsilon^4}$, we can now use the bound in Eq. (5) to get w.p $> 1-\delta$:

$$\sup_x \left|\frac{1}{k_2}\hat{g}(x) - \frac{1}{k_2 M}\sum_{j=1}^{k_2} f^{(j)}(x)\right| \leq \epsilon .$$

□

We are now ready to prove the main theorem:

*Proof of Thm. 3.2.* Set $m = \frac{256\log\left(\frac{2n}{\delta}\right)C^4n^4L^4}{\epsilon^4}\cdot\frac{\log\left(\frac{1}{\delta}\right)}{2\delta^3}$ and initialize a 2-layer neural network with width $k := m\cdot n$ and initialization as described in the theorem, denote $g(x) = \sum_{j=1}^m\sum_{i=1}^n u_i^{(j)}\sigma(\langle w_i^{(j)}, x\rangle)$ as this network. By the assumption of the theorem, for each $j \in [m]$ w.p $> 1-\delta$ there exists a vector $v^{(j)}$ with $\|v^{(j)}\|_\infty \leq C$ such that the function $f^{(j)}(x) = \sum_{i=1}^n v_i^{(j)}\sigma(\langle w_i^{(j)}, x\rangle)$ satisfy that $L_D\left(f^{(j)}\right) \leq \epsilon$. Let $Z_j$ be the random variable such that $Z_j = 0$ if there exists a vector $v^{(j)}$ that satisfies the above, and $Z_j = 1$ otherwise. the random variables $Z_j$ are i.i.d since we initialize each $w_i^{(j)}$ i.i.d, and $\mathbb{P}(Z_j = 1) = \delta$, $\mathbb{E}[Z_j] = \delta$. Denote $Z = \sum_{j=1}^m Z_j$, then $\mathbb{E}[Z] = m\delta$. We use Hoeffding's inequality on $Z$ to get that:

$$\mathbb{P}\left(\frac{1}{m}Z \geq \delta + t\right) \leq \exp(-2mt^2) .$$

Replacing the r.h.s with $\delta$ and setting $t = \delta$ we get that if $m > \frac{\log\left(\frac{1}{\delta}\right)}{2\delta^2}$ then w.p $> 1-\delta$ we have that $Z \leq 2\delta$. In particular, there are at least $m_0 = \frac{256\log\left(\frac{2n}{\delta}\right)C^4n^4L^4}{\epsilon^4}$ indices (denote them w.l.o.g $j = 1,\ldots,m_0$) such that for every $j \in [m_0]$ there exists a vector $v^{(j)}$ with $\|v^{(j)}\|_\infty \leq C$ such that the function $f^{(j)}(x) = \sum_{i=1}^n v_i^{(j)}\sigma(\langle w_i^{(j)}, x\rangle)$ satisfy that $L_D\left(f^{(j)}\right) \leq \epsilon$.

We now use Lemma B.2 with $\delta$, $\frac{\epsilon}{C}$ and $v^{(1)},\ldots,v^{(m_0)}$ to get that w.p $>1-\delta$ that there exists a neuron-subnetwork $\tilde{g}(x)$ and constant $c'>0$ such that:

$$\sup_{x:\|x\|\leq 1}\left|c'\tilde{g}(x)-\frac{1}{m_0 C}\sum_{j=1}^{m_0}f^{(j)}(x)\right|\leq\frac{\epsilon}{C} \tag{8}$$

Set $c=C\cdot c'$, the loss of $c\tilde{g}(x)$ can be bounded by:

$$\begin{aligned}L_D(c\tilde{g}) =&\mathbb{E}_{(x,y)\sim D}\left[(c\tilde{g}(x)-y)^2\right]\\ &\leq 2\mathbb{E}_{(x,y)\sim D}\left[\left(c\tilde{g}(x)-\frac{1}{m_0}\sum_{j=1}^{m_0}f^{(j)}(x)\right)^2\right]+2\mathbb{E}_{(x,y)\sim D}\left[\left(\frac{1}{m_0}\sum_{j=1}^{m_0}f^{(j)}(x)-y\right)^2\right]\end{aligned} \tag{9}$$

We will bound each term of the above expression. Using Eq. (8) we have:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim D}\left[\left(c\tilde{g}(x)-\frac{1}{m}\sum_{j=1}^{m}f^{(j)}(x)\right)^2\right]&\leq\sup_{x:\|x\|\leq 1}\left(c\tilde{g}(x)-\frac{1}{m}\sum_{j=1}^{m}f^{(j)}(x)\right)^2\\ &\leq C\cdot\sup_{x:\|x\|\leq 1}\left(c'\tilde{g}(x)-\frac{1}{mC}\sum_{j=1}^{m}f^{(j)}(x)\right)^2\leq C\cdot\frac{\epsilon}{C}=\epsilon\end{aligned} \tag{10}$$

For the second term in Eq. (9) we have that:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim D}\left[\left(\frac{1}{m}\sum_{j=1}^{m}f^{(j)}(x)-y\right)^2\right]&\leq\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{(x,y)\sim D}\left[\left(f^{(j)}(x)-y\right)^2\right]\\ &\leq\frac{1}{m}\sum_{j=1}^{m}L_D\left(f^{(j)}\right)\leq\epsilon\end{aligned} \tag{11}$$

re-scaling $\epsilon$ finishes the proof. □

## C. Proofs of section 3.1

We first show that a finite dataset, under mild assumptions on the data, can be approximated using a random features model. The proof of the following lemma is exactly the same as the proof of Lemma 3.1 in (9).

**Lemma C.1.** *Let $\delta>0$, $x_1,\ldots,x_m\in\mathbb{R}^d$, and let $H$ be the $m\times m$ matrix with:*

$$H_{i,j}=\mathbb{E}_w\left[\sigma(\langle w,x_i\rangle)\sigma(\langle w,x_j\rangle)\right]$$

*Assume that $\lambda_{\min}(H)=\lambda>0$, then for $k>\frac{64m^2\log^2\left(\frac{m}{\delta}\right)}{\lambda^2}$, w.p $>1-\delta$ over sampling of $w_1,\ldots,w_k$ we have that $\lambda_{\min}(\tilde{H})\geq\frac{3}{4}\lambda$ where:*

$$\tilde{H}_{i,j}=\sum_{l=1}^{k}\sigma(\langle w_l,x_i\rangle)\sigma(\langle w_l,x_j\rangle)$$

Using the lemma above, and under the assumptions made on the data, w.h.p a two-layer network of size $\tilde{O}\left(\frac{m^2}{\lambda^2}\right)$ can overfit the data:

**Proposition C.2.** *Let $\delta>0$, $x_1,\ldots,x_m\in\mathbb{R}^d$ and $y_1,\ldots,y_m\in\{\pm1\}$. Assume that $\lambda_{\min}(H)=\lambda>0$, and $\sigma$ is L-Lipschitz then for $k>\frac{64m^2\log^2\left(\frac{m}{\delta}\right)}{\lambda^2}$ w.p $1-\delta$ over sampling of $w_1,\ldots,w_k$ there is $u\in\mathbb{R}^k$ with $\|u\|_\infty\leq\frac{4Lm}{3\lambda}$ such that for every $j=1,\ldots,m$ we have $\sum_{i=1}^{k}u_i\sigma(\langle w_i,x_j\rangle)=y_j$*

*Proof.* Set $X$ to be the $k \times m$ matrix defined by $X_{i,j} = \sigma(\langle w_i, x_j \rangle)$. By our assumption and the choice of $k$, w.p $> 1 - \delta$ we have that $\tilde{H} = X^\top X$ is invertible, and has a minimal eigenvalue of at least $\frac{3}{4}\lambda$. Define $u = y(X^\top X)^{-1}X^\top$, it is easy to see that $uX = y$, furthermore:

$$\begin{aligned}\|u\|_\infty = \|y(X^\top X)^{-1}X^\top\|_\infty &\leq \frac{4}{3\lambda}\|Xy\|_\infty \\ &\leq \frac{4}{3\lambda} m \max_{w,x} \sigma(\langle w, x \rangle) \leq \frac{4Lm}{3\lambda}\end{aligned}$$

□

For the second variation of Thm. 3.4 we consider functions from the class of functions $\mathcal{F}_C$. Here we use Theorem 3.3 from (36):

**Theorem C.3.** *Let $f(x) = c_d \int_{w \in \left[\frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]^d} g(w)\sigma(\langle w, x \rangle)dw$ where $\sigma : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz on $[-1, 1]$ with $\sigma(0) \leq L$, and $c_d = \left(\frac{\sqrt{d}}{2}\right)^d$ a normalization term. Assume that $max_{\|w\| \leq 1}|g(w)| \leq C$ for a constant $C$. Then for every $\delta > 0$ if $w_1, \ldots, w_k$ are drawn i.i.d from the uniform distribution on $\left[\frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]^d$, w.p $> 1 - \delta$ there is a function of the form*

$$\hat{f}(x) = \sum_{i=1}^{k} u_i \sigma(\langle w_i, x \rangle)$$

*where $|u_i| \leq \frac{C}{k}$ for every $1 \leq i \leq k$, such that:*

$$\sup_x \left|\hat{f}(x) - f(x)\right| \leq \frac{LC}{\sqrt{k}}\left(4 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right)$$

To prove the main theorem, we use the same argument as in the proof of Thm. 3.2, that pruning neurons can approximate random features models. Here the size of the target random features model depends on the complexity of the target (either a finite dataset or RKHS function).

*Proof of Thm. 3.4.* Although the proof for the two variations of the theorem are similar, for clarity and ease of notations we will prove them separately.

1. (Finite dataset) Let $\epsilon, \delta > 0$. Fix $\delta_1 = \frac{\delta}{2k_2}$, and fix some $j \in [k_2]$. Take $k_1 \geq \frac{64m^2 \log^2\left(\frac{m}{\delta_1}\right)}{\lambda^2}$, from Proposition C.2 w.p $> 1 - \delta_1$ we get the following: There exists some $v^{(j)} \in \mathbb{R}^{k_1}$ with $\|v^{(j)}\|_\infty \leq \frac{4Lm}{3\lambda}$ such that for the function $f^{(j)}(x) := \sum_{i=1}^{k_1} v_i^{(j)} \sigma\left(\langle w_i^{(j)}, x \rangle\right)$, and for every $l = 1, \ldots, m$, we have $f^{(j)}(x_l) = y_l$. Using union bound over all choices of $j$, we get that w.p $> 1 - \frac{\delta}{2}$ the above hold for every $j \in [k_2]$.

   Denote $M := \frac{4Lm}{3\lambda}$, $\epsilon' = \frac{\epsilon}{M} = \frac{3\lambda\epsilon}{4Lm}$ and let $k_2 > \frac{810L^8 m^4 k_1^4 \log\left(\frac{2k_1}{\delta}\right)}{\lambda^4 \epsilon^4}$. Using Lemma B.2 with $v^{(1)}, \ldots, v^{(k_2)}$ and $\epsilon'$ we have that there exist $b^{(1)}, \ldots, b^{(k_2)}$ such that for the functions $\tilde{g}^{(j)}(x) = \sum_{i=1}^{k_1} b_i^{(j)} \cdot u_i^{(j)} \sigma\left(\langle w_i^{(j)}, x \rangle\right)$ we get:

$$\sup_{x:\|x\| \leq 1} \left|\frac{c_1}{k_2}\sum_{j=1}^{k_2} \tilde{g}^{(j)}(x) - \frac{1}{k_2 M}\sum_{j=1}^{k_2} f^{(j)}(x)\right| \leq \epsilon' \tag{12}$$

   where $c_1 = \frac{8k_1 L}{\epsilon}$. Denote $\tilde{g}(x) = \sum_{j=1}^{k_2} g^{(j)}(x)$ and set $c = \frac{c_1 M}{k_2} = \frac{32k_1 Lm}{3\lambda\epsilon k_2}$. Using Eq. (12) we have that for every $l = 1, \ldots, m$:

$$|c\tilde{g}(x_l) - y_l| = \left|\frac{c_1 M}{k_2}\tilde{g}(x_l) - \frac{1}{k_2}\sum_{j=1}^{k_2} f^{(j)}(x_l)\right| \leq M\epsilon' \leq \epsilon$$

2. Let $\epsilon, \delta > 0$. Fix $\delta_1 = \frac{\delta}{2k_2}$, and fix some $j \in [k_2]$. Take $k_1 \geq \frac{128L^2C^2\log^2\left(\frac{m}{\delta_1}\right)}{\epsilon^2}$, from Thm. C.3 w.p $> 1 - \delta_1$ we get the following: There exists some $v^{(j)} \in \mathbb{R}^{k_1}$ with $\|v^{(j)}\|_\infty \leq \frac{C}{k_1} \leq 1$ such that for the function $f^{(j)}(x) := \sum_{i=1}^{k_1} v_i^{(j)} \sigma\left(\langle w_i^{(j)}, x\rangle\right)$ we have $\sup_{x:\|x\|\leq 1} \left|f^{(j)}(x) - f(x)\right| \leq \frac{\epsilon}{2}$. Using union bound over all choices of $j$, we get that w.p $> 1 - \frac{\delta}{2}$ the above hold for every $j \in [k_2]$.

   Let $k_2 > \frac{4010L^4k_1^4\log\left(\frac{2k_1}{\delta}\right)}{\epsilon^4}$, using Lemma B.2 with $v^{(1)}, \ldots, v^{(k_2)}$ and $\frac{\epsilon}{2}$ we have that there exist $b^{(1)}, \ldots, b^{(k_2)}$ such that for the functions $\tilde{g}^{(j)}(x) = \sum_{i=1}^{k_1} b_i^{(j)} \cdot u_i^{(j)} \sigma\left(\langle w_i^{(j)}, x\rangle\right)$ we get:

$$\sup_{x:\|x\|\leq 1} \left| \frac{c_1}{k_2} \sum_{j=1}^{k_2} \tilde{g}^{(j)}(x) - \frac{1}{k_2 M} \sum_{j=1}^{k_2} f^{(j)}(x) \right| \leq \frac{\epsilon}{2} \tag{13}$$

   where $c_1 = \frac{8k_1L}{\epsilon}$. Denote $\tilde{g}(x) = \sum_{j=1}^{k_2} g^{(j)}(x)$ and set $c = \frac{c_1}{k_2} = \frac{8k_1L}{\epsilon k_2}$. Using Eq. (13) we have that:

$$\begin{aligned}
&\sup_{x:\|x\|\leq 1} |c\tilde{g}(x) - f(x)| \\
&\leq \sup_{x:\|x\|\leq 1} \left| \frac{c_1}{k_2}\tilde{g}(x) - \frac{1}{k_2}\sum_{j=1}^{k_2} f^{(j)}(x) \right| + \sup_{x:\|x\|\leq 1} \left| \frac{1}{k_2}\sum_{j=1}^{k_2} f^{(j)}(x) - f(x) \right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon
\end{aligned}$$

□