A Proofs of tempered overfitting

A.1 Proof of Theorem 3.1

Throughout the proof, given a sample $S = (x_i, y_i)_{i=1}^m \sim \mathcal{D}^m$ we denote $S_x = (x_i)_{i=1}^m$, $S_y = (y_i)_{i=1}^m$, and assume without loss of generality that $x_1 \leq x_2 \leq \cdots \leq x_m$.

Lemma A.1. *Denote by E_x the event in which the samples $(x_i)_{i=1}^m \sim \mathcal{D}_x^m$ satisfy*

- *(maximal gap isn't too large)* $d_{\max} := \max_{i\in[m-1]}(x_{i+1} - x_i) \leq \frac{\log(8(m+1)/\delta)}{m+1}$.
- *(most gaps aren't too small)* $|\{i \in [m-1] \,:\, x_{i+1} - x_i < \frac{1}{10(m+1)}\}| < \frac{m+1}{8}$.
- *(no collisions)* $\forall i \neq j \in [m] :\ x_i \neq x_j$.

Then there exists absolute $m_0 \in \mathbb{N}$ such that $\forall m \geq m_0 :\ \Pr_{S_x\sim\mathcal{D}_x^m}[E_x] \geq 1 - \frac{\delta}{4}$.

Proof. Deferred to Appendix D. □

Following the lemma above, we continue by conditioning on the probable event E_x, after which we will conclude the proof by the union bound. We now state a lemma due to Safran et al. [2022] which is crucial for our analysis.

Lemma A.2 (Lemma E.6. Safran et al., 2022)**.** *Suppose that $i < j$ are such that $y_i, y_{i+1}, \ldots, y_j = -1$. Then in the interval $[x_i, x_j]$ there are at most two points at which N'_θ increases.*

We derive the following corollary:

Corollary A.3. *Suppose that $y_i, y_{i+1}, \ldots, y_{i+4} = -1$. Then there exists $i \leq \ell \leq i+3$ for which $N_{\boldsymbol{\theta}}|_{[x_\ell, x_{\ell+1}]} < 0$.*

Proof. Assume towards contradiction that $y_i, \ldots, y_{i+4} = -1$, yet for any $i \leq \ell \leq i+3$ there exists $z_\ell \in (x_\ell, x_{\ell+1})$ such that $N_{\boldsymbol{\theta}}(z_\ell) \geq 0$. Recall that $N_{\boldsymbol{\theta}}(x_i), \ldots, N_{\boldsymbol{\theta}}(x_{i+3}) \leq -1$ by Eq. (3). Thus for each $i \leq \ell \leq i+2$, looking at the segment $(z_\ell, z_{\ell+1}) \ni x_{\ell+1}$ we see that $N_{\boldsymbol{\theta}}(z_\ell) > 0, N_{\boldsymbol{\theta}}(x_{\ell+1}) \leq -1, N_{\boldsymbol{\theta}}(z_\ell) > 0$. In particular, by the mean value theorem, any such segment must contain a point at which N'_θ increases. Obtaining three such points which are distinct contradicts Lemma A.2. □

We assume without loss of generality that m is divisible by 5 and split the index set $[m]$ into groups consecutive five indices: we let $I_1 = \{1, \ldots, 5\}, I_2 = \{6, \ldots, 10\}$ and so on up to $I_{m/5}$. Denoting by μ the (one dimensional) Lebesgue measure we get that under the event E_x it holds that

$$
\begin{aligned}
\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}[N_{\boldsymbol{\theta}}(x) < 0]\right] &= \mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\mu(x : N_{\boldsymbol{\theta}}(x) < 0)\right] \\
&\geq \mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\sum_{i\in[m-1]} \mu(x_{i+1} - x_i) \cdot \mathbb{1}\left\{N|_{[x_i, x_{i+1}]} < 0\right\}\right] \\
&= \sum_{i\in[m-1]} \mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[(x_{i+1} - x_i) \cdot \mathbb{1}\left\{N|_{[x_i, x_{i+1}]} < 0\right\}\right] \\
&= \sum_{i\in[m/5]} \sum_{l\in[I_i]} \mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[(x_{l+1} - x_l) \cdot \mathbb{1}\left\{N|_{[x_l, x_{l+1}]} < 0\right\}\right] \\
[\text{Corollary A.3}] &\geq \sum_{i\in[m/5]} \sum_{l\in[I_i]} \mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\min_{\ell\in I_i}(x_{\ell+1} - x_\ell) \cdot \mathbb{1}\left\{\forall \ell \in I_i : y_\ell = -1\right\}\right] \\
&= \sum_{i\in[m/5]} \sum_{l\in[I_i]} \min_{\ell\in I_i}(x_{\ell+1} - x_\ell) \cdot \Pr[y_i, y_{i+1}, \ldots, y_{i+4} = -1] \\
&\geq \widetilde{c}p^5 \,,
\end{aligned}
$$

where the last inequality follows from our conditioning on E_x. To see why, note that the sum $\sum_{i\in[m/5]}\sum_{l\in[I_i]}\min_{\ell\in I_i}(x_{\ell+1}-x_\ell)$ is lower bounded by the sum of the $m/5$ smallest gaps, yet under E_x this sum contains at least $\Omega(m)$ summands larger than $\Omega(1/m)$ — hence it is at least some constant. We conclude that as long as E_x occurs we have $\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\right]\right]\geq cp^5$. Moreover, we see by the analysis above that if a *single* label y_l for some $l\in I_i\subset[m]$ is changed, this can affect $N_{\boldsymbol{\theta}}(x)$ only in the segment $[x_{\min_{I_i}\ell}, x_{\max_{I_i}\ell}]$ which is of length at most $7d_{\max}=O\left(\frac{\log(m/\delta)}{m}\right)$. Thus we can apply McDiarmid's inequality to obtain that under E_x, with probability at least $1-\delta/4$:

$$\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\right]\geq c\left(p^5-\sqrt{\frac{\log(m/\delta)}{m}}\right)\,.$$

Overall, by union bounding over E_x the inequality above holds with probability at least $1-\delta/2$, which proves the desired lower bound.

We now turn to prove the upper bound. Let $N^*(\cdot)$ be a 2-layer ReLU network of minimal width $n^*\in\mathbb{N}$ that classifies the data correctly, namely $y_iN^*(x_i)>0$ for all $i\in[m]$. Note that n^* is uniquely defined by the sample while N^* is not. Furthermore, n^* is upper bounded by the number of neighboring samples with different labels.[6] Hence,

$$\begin{aligned}\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[n^*\right]&\leq\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\left|\{i\in[m-1]:y_i\neq y_{i+1}\}\right|\right]\\&=(m-1)\cdot\mathbb{E}_{S_y\sim\mathcal{D}_y^m}[\mathbb{1}\left\{y_1\neq y_2\right\}]\\&=(m-1)\cdot\Pr_{S_y\sim\mathcal{D}_y^m}[y_1\neq y_2]\\&=2(m-1)p(1-p)=O(pm)\,.\end{aligned}\tag{6}$$

We conclude that the expected width of N^* (as a function of the sample) is at most $n^*=O(pm)$. By Safran et al. [2022, Theorem 4.2], this implies $N_{\boldsymbol{\theta}}$ belongs to a class of VC dimension $O(n^*)=O(pm)$. Thus by denoting the 0-1 loss $L^{0-1}(N_{\boldsymbol{\theta}})=\Pr_{(x,y)\sim\mathcal{D}}[\operatorname{sign}(N_{\boldsymbol{\theta}}(x))\neq y]$ and invoking a standard VC generalization bound we get

$$\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[L^{0-1}(N_{\boldsymbol{\theta}})\,|\,n^*\right]\leq\underbrace{\frac{1}{m}\sum_{i=1}^m\mathbb{1}\left\{\operatorname{sign}(N_{\boldsymbol{\theta}}(x_i))\neq y_i\right\}}_{=0}+O\left(\sqrt{\frac{n^*}{m}}\right)=O\left(\sqrt{\frac{n^*}{m}}\right)\,.$$

By Jensen's inequality $\mathbb{E}[\sqrt{n^*}]\leq\sqrt{\mathbb{E}[n^*]}\lesssim\sqrt{pm}\implies\mathbb{E}[\sqrt{n^*/m}]\lesssim\sqrt{p}$, so by the law of total expectation

$$\mathbb{E}_{S_y\sim\mathcal{D}_y^m}[L^{0-1}(N_{\boldsymbol{\theta}})]=\mathbb{E}_{n^*}\left[\mathbb{E}_{S_y\sim\mathcal{D}_y^m}[L^{0-1}(N_{\boldsymbol{\theta}})\,|\,n^*]\right]\lesssim\mathbb{E}_{n^*}\left[\sqrt{\frac{n^*}{m}}\right]\lesssim\sqrt{p}\,.\tag{7}$$

In order to relate the bound above to the clean test error, note that

$$\begin{aligned}L^{0-1}(N_{\boldsymbol{\theta}})&=\Pr_{(x,y)\sim\mathcal{D}}[\operatorname{sign}(N_{\boldsymbol{\theta}}(x))\neq y]\\&=(1-p)\cdot\Pr_{(x,y)\sim\mathcal{D}}[N_{\boldsymbol{\theta}}(x)\leq0|y=1]+p\cdot\Pr_{(x,y)\sim\mathcal{D}}[N_{\boldsymbol{\theta}}(x)>0|y=-1]\\&\geq\underbrace{(1-p)}_{\in[\frac{1}{2},1]}\cdot\Pr_{x\sim\mathcal{D}_x}[N_{\boldsymbol{\theta}}(x)\leq0]\\&\geq\frac{1}{2}\cdot\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)\leq0\right]\,,\end{aligned}$$

hence

$$\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)\leq0\right]\right]\leq\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[2L^{0-1}(N_{\boldsymbol{\theta}})\right]\overset{\text{Eq. (7)}}{\lesssim}\sqrt{p}\,.$$

[6] This can be seen by considering a network representing the linear spline of the data, for which it suffices to set a neuron for adjacent samples with alternating signs.

As in our argument for the lower bound, we now note by Eq. (6) that flipping a single label y_l for some $l \in [m]$ changes n^* by at most 1, hence changing the test error by at most $O(1/m)$. Therefore we can apply McDiarmid's inequality and see that under the event E_x, with probability at least $1 - \delta/4$:

$$\Pr_{x \sim \mathcal{D}_x} [N_{\boldsymbol{\theta}}(x) \leq 0] \leq C \left(\sqrt{p} + \sqrt{\frac{\log(1/\delta)}{m}} \right) .$$

Overall, by union bounding over E_x the inequality above holds with probability at least $1 - \delta/2$, which proves the upper bound and finishes the proof.

A.2 Proof of Theorem 3.2

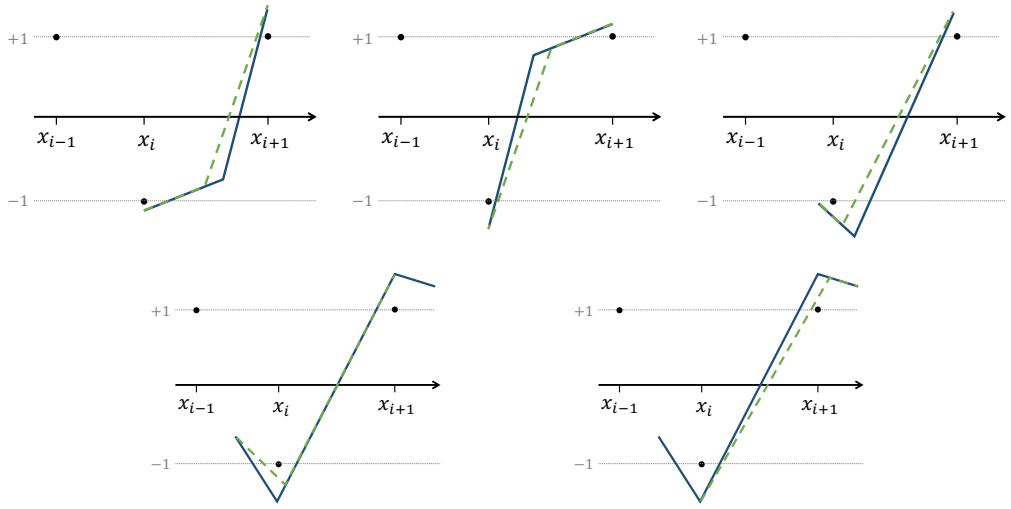

Figure 3: Illustration of the proof of Theorem 3.2 in case there is a single non-linearity along $[x_i, x_{i+1}]$. If the network is not linear along $[x_i, x_{i+1}]$, one of the cases illustrated in the top row (in blue) must occur. In each case, the dashed green perturbation classifies correctly by altering exactly two neurons while reducing the parameter norm. Moreover, if the network is linear along $[x_i, x_{i+1}]$, yet $N_{\boldsymbol{\theta}}(x_i) < -1$ or $N_{\boldsymbol{\theta}}(x_{i+1}) > 1$, one of the cases illustrated in the bottom row must occur. In either case, the dashed green perturbation classifies correctly by altering exactly two neurons while reducing the parameter norm.

Throughout the proof, given a sample $S = (x_i, y_i)_{i=1}^m \sim \mathcal{D}^m$ we denote $S_x = (x_i)_{i=1}^m$, $S_y = (y_i)_{i=1}^m$, and assume without loss of generality that $x_1 \leq x_2 \leq \cdots \leq x_m$. Denote by E_x the event in which the samples $(x_i)_{i=1}^m \sim \mathcal{D}_x^m$ satisfy

$$d_{\max} := \max_{i \in [m-1]} (x_{i+1} - x_i) \leq \frac{\log(8(m+1)/\delta)}{m+1} ,$$

and recall that by Lemma A.1 we have $\Pr_{S_x \sim \mathcal{D}_x^m}[E_x] \geq 1 - \frac{\delta}{4}$ for m larger than an absolute constant. Therefore from here on throughout the proof we condition on E_x, after which we can conclude using the union bound.

For any point $x \in (0, 1)$, denote by i_x the maximal index $i \in [m]$ such that $x_i \leq x$. Let A_x denote the event in which $y_{i_x - 1} = 1$, $y_{i_x} = -1$ and $y_{i_x + 1} = 1$. Note that for any x we have $\Pr_{S_y \sim \mathcal{D}_y^m}[A_x] = p(1-p)^2 \geq \frac{1}{4}p$, so we get

$$\begin{aligned}\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\right]\right] &= \mathbb{E}_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[\mathbb{1}\left\{N_{\boldsymbol{\theta}}(x)<0\right\}\right] \\ &\geq \mathbb{E}_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[\mathbb{1}\left\{N_{\boldsymbol{\theta}}(x)<0\right\}|A_x\right]\cdot\Pr_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[A_x\right] \\ &\geq \frac{p}{4}\cdot\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\,|\,A_x\right]\right]. \end{aligned} \tag{8}$$

We aim to show that $\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\,|\,A_x\right]=\Omega(1)$. In order to do so, note that conditioning the uniform measure $\mathcal{D}_x=\mathrm{Unif}([0,1])$ on the event A_x results in a uniform measure over the (union of) segments between -1 labeled samples to $+1$ labeled samples that also satisfy the additional property that the neighboring sample on their left is labeled $+1$. We will show that if $\boldsymbol{\theta}$ is a local minimum of Problem (1), then along any such segment $N_{\boldsymbol{\theta}}(x)$ is linear from -1 to $+1$:

Proposition A.4. *Let $i\in[m-1]$ be such that $x_{i-1}<x_i<x_{i+1}$ with $y_{i-1}=1$, $y_i=-1$ and $y_{i+1}=1$. If $\boldsymbol{\theta}$ is a local minimum of Problem (1), it holds that $N_{\boldsymbol{\theta}}(x_i)=-1$, $N_{\boldsymbol{\theta}}(x_{i+1})=1$ and $N_{\boldsymbol{\theta}}(\cdot)$ is linear over (x_i,x_{i+1}).*

In particular, the proposition above shows that

$$\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\,|\,A_x\right]=\frac{1}{2}\,,$$

which plugged into Eq. (8) gives

$$\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\right]\right]\geq\frac{p}{8}\,.$$

Moreover, noting that any local minimum is in particular a KKT point, we saw in the proof of Theorem 3.1 that under the event E_x, changing a single label $y_l,\ l\in[m]$ can change the test error by at most $O(d_{\max})=O(\log(m/\delta)/m)$. By McDiarmid's inequality this finishes the proof of the desired lower bound.

As for the upper bound, for $x\in(0,1)$ we denote by B_x the event in which $y_{i_x}=1=y_{i_x+1}=1$ — namely, x is between two positively labeled samples. Note that for any x we have $\Pr_{S_y\sim\mathcal{D}_y^m}[B_x]=(1-p)^2\geq 1-2p\implies\Pr_{S_y\sim\mathcal{D}_y^m}[B_x^c]\leq 2p$. We will show that if $\boldsymbol{\theta}$ is a local minimum of the margin maximization problem, then B_x implies that $N_{\boldsymbol{\theta}}(x)\geq 0$.

Proposition A.5. *Let $i\in[m-1]$ be such that $y_i=y_{i+1}=1$, and let $x\in[x_i,x_{i+1}]$. If $\boldsymbol{\theta}$ is a local minimum of Problem (1), it holds that $N_{\boldsymbol{\theta}}(x)\geq 0$.*

In particular, the proposition above shows that

$$\mathbb{E}_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[\mathbb{1}\left\{N_{\boldsymbol{\theta}}(x)<0\right\}|B_x\right]=0\,,$$

so we get

$$\begin{aligned}\mathbb{E}_{S_y\sim\mathcal{D}_y^m}\left[\Pr_{x\sim\mathcal{D}_x}\left[N_{\boldsymbol{\theta}}(x)<0\right]\right] &= \mathbb{E}_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[\mathbb{1}\left\{N_{\boldsymbol{\theta}}(x)<0\right\}\right] \\ &= \mathbb{E}_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[\mathbb{1}\left\{N_{\boldsymbol{\theta}}(x)<0\right\}|B_x\right]\cdot\Pr_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[B_x\right] \\ &+ \mathbb{E}_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[\mathbb{1}\left\{N_{\boldsymbol{\theta}}(x)<0\right\}|B_x^c\right]\cdot\Pr_{S_y\sim\mathcal{D}_y^m,x\sim\mathcal{D}_x}\left[B_x^c\right] \\ &\leq 0+1\cdot 2p=2p\,.\end{aligned}$$

As in the lower bound proof, recalling that any local minimum is in particular a KKT point, we saw in the proof of Theorem 3.1 that under the event E_x, changing a single label $y_l,\ l\in[m]$ can change the test error by at most $O(d_{\max})=O(\log(m/\delta)/m)$. Hence applying McDiarmid's inequality proves the upper bound thus finishing the proof.

Proof of Proposition A.4. Throughout the proof we fix $i\in[m]$ for which the conditions described in the proposition hold, and we assume without loss of generality that the neurons are ordered

with respect to their activation point: $-\frac{b_1}{w_1} \leq -\frac{b_2}{w_2} \leq \cdots \leq -\frac{b_n}{w_n}$. Moreover, we may assume without loss of generality that $w_1, \ldots, w_n \geq 0$. Indeed, note that Eq. (1) is invariant under the transformation $v_j \leftarrow -v_j, w_j \leftarrow -w_j, b_j \leftarrow -b_j$ which does not affect neither the parameter norm nor the parameterized network in function space. Hence, any local minimum of Eq. (1) corresponds to a local minimum with $w_1, \ldots, w_n \geq 0$. Lastly, we will make frequent use of the following simple observation. For differentiable x we have

$$N'_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{n} v_j \cdot \underbrace{w_j \mathbb{1}\left\{w_j x + b_j \geq 0\right\}}_{\geq 0}, \tag{9}$$

so if $N'_{\boldsymbol{\theta}}(z_1) < N'_{\boldsymbol{\theta}}(z_2)$ for some $z_1 < z_2$, then there must exist $j \in [n]$ with $v_j > 0$ and $z_1 < -\frac{b_j}{w_j} < z_2$. Similarly, $N'_{\boldsymbol{\theta}}(z_1) > N'_{\boldsymbol{\theta}}(z_2)$ implies that there exists $j \in [n]$ with $v_j < 0,\ z_1 < -\frac{b_j}{w_j} < z_2$.

We split the proof of Proposition A.4 into two lemmas.

Lemma A.6. *$N_{\boldsymbol{\theta}}(\cdot)$ is linear over (x_i, x_{i+1}).*

Proof. Recall that $N_{\boldsymbol{\theta}}(x_i) \leq -1$ and $N_{\boldsymbol{\theta}}(x_{i+1}) \geq 1$, so $N_{\boldsymbol{\theta}}$ increases along the segment (x_i, x_{i+1}). Thus, $N_{\boldsymbol{\theta}}(\cdot)$ is *not* linear over (x_i, x_{i+1}) only if (at least) one of the following occur: (1) There exist $z_1, z_2 \in (x_i, x_{i+1})$ such that $z_1 < z_2$ and $0 < N'_{\boldsymbol{\theta}}(z_1) < N'_{\boldsymbol{\theta}}(z_2)$; (2) there exist $z_1, z_2 \in (x_i, x_{i+1})$ such that $z_1 < z_2,\ N'_{\boldsymbol{\theta}}(z_1) > 0$ and $N'_{\boldsymbol{\theta}}(z_1) > N'_{\boldsymbol{\theta}}(z_2)$; or (3) there exists a single $z \in (x_i, x_{i+1})$ at which $N_{\boldsymbol{\theta}}$ is non-differentiable, such that $N'_{\boldsymbol{\theta}}|_{(x_i,z)} \leq 0$ and $N'_{\boldsymbol{\theta}}|_{(z,x_{i+1})} > 0$. We will show either of these contradict the assumption that $\boldsymbol{\theta}$ is a local optimum of the margin maximization problem.

Case (1). The assumption on z_1, z_2 implies that there exists $j_1 \in [n]$ such that $-\frac{b_{j_1}}{w_{j_1}} \in (z_1, z_2)$ and $v_{j_1} > 0$. Let $j_2 > j_1$ be the minimal index $j \in \{j_1 + 1, \ldots, n\}$ for which $v_j < 0$.[7] For some small $\delta > 0$, consider the perturbed network

$$\begin{aligned} N_{\boldsymbol{\theta}_\delta}(x) := & \sum_{j \in [n] \setminus \{j_1, j_2\}} v_j \sigma\left(w_j \cdot x + b_j\right) \\ & + (1-\delta) v_{j_1} \sigma\left(w_{j_1} \cdot x + \left(b_{j_1} - \frac{\delta}{1-\delta}\left(\frac{w_{j_1} b_{j_2}}{w_{j_2}} - b_{j_1}\right)\right)\right) \\ & + \left(1 + \delta \frac{v_{j_1} w_{j_1}}{v_{j_2} w_{j_2}}\right) v_{j_2} \sigma\left(w_{j_2} \cdot x + b_{j_2}\right) . \end{aligned} \tag{10}$$

It is clear that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\delta\| \xrightarrow{\delta \to 0} 0$, and we will show that for small enough δ the network above still satisfies the margin condition, yet has smaller parameter norm. To see why the margin condition is not violated for small enough δ, notice that $N_{\boldsymbol{\theta}_\delta}(x) = N_{\boldsymbol{\theta}}(x)$ for all $x \notin (x_i, -\frac{b_{j_2}}{w_{j_2}})$ so in particular $N_{\boldsymbol{\theta}_\delta}(x_l) = N_{\boldsymbol{\theta}}(x_l)$ for all $l \in [i]$, as well as for all $x_l \geq -\frac{b_{j_2}}{w_{j_2}}$. Furthermore, by minimality of j_2 we note that there cannot exist $y_k = -1$ for k such that $x_k \in (-\frac{b_{j_1}}{w_{j_1}}, -\frac{b_{j_2}}{w_{j_2}})$. A direct computation gives that $N_{\boldsymbol{\theta}_\delta} \geq N_{\boldsymbol{\theta}}$ along this segment, so overall the margin condition is indeed satisfied for the

[7] We can assume that such j_2 exists, by otherwise discarding j_2 in the rest of the proof which would work verbatim. Notably, the only case in which there does not exist such j_2 is when x_i is the last sample to be labeled -1.

entire sample. As to the parameter norm, we have

$$\begin{aligned}\|\boldsymbol{\theta}_\delta\|^2 &= \sum_{j\in[n]\setminus\{j_1,j_2\}} (v_j^2+w_j^2+b_j^2) + (1-\delta)^2 v_{j_1}^2 + w_{j_1}^2 + \left(b_{j_1} - \frac{\delta}{1-\delta}\left(\frac{w_{j_1}b_{j_2}}{w_{j_2}} - b_{j_1}\right)\right)^2 \\ &\quad + \left(1+\delta\frac{v_{j_1}w_{j_1}}{v_{j_2}w_{j_2}}\right)^2 v_{j_2}^2 + w_{j_2}^2 + b_{j_2}^2 \\ &= \sum_{j\in[n]\setminus\{j_1,j_2\}} (v_j^2+w_j^2+b_j^2) + (1-2\delta)v_{j_1}^2 + w_{j_1}^2 + b_{j_1}^2 - \frac{2\delta}{1-\delta}b_{j_1}\left(\frac{w_{j_1}b_{j_2}}{w_{j_2}} - b_{j_1}\right) \\ &\quad + \left(1+2\delta\frac{v_{j_1}w_{j_1}}{v_{j_2}w_{j_2}}\right) v_{j_2}^2 + w_{j_2}^2 + b_{j_2}^2 + O(\delta^2) \\ &= \sum_{j\in[n]} (v_j^2+w_j^2+b_j^2) - 2\delta\left(v_{j_1}^2 + \frac{b_{j_1}}{1-\delta}\left(\frac{w_{j_1}b_{j_2}}{w_{j_2}} - b_{j_1}\right) - \frac{v_{j_1}w_{j_1}}{v_{j_2}w_{j_2}}\right) + O(\delta^2)\,.\end{aligned}$$

Thus,

$$\begin{aligned}\|\boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_\delta\|^2 &= 2\delta\left(v_{j_1}^2 - \frac{v_{j_1}w_{j_1}v_{j_2}}{w_{j_2}} - \frac{b_{j_1}}{1-\delta}\left(b_{j_1} - \frac{w_{j_1}b_{j_2}}{w_{j_2}}\right)\right) + O(\delta^2) \\ \implies \|\boldsymbol{\theta}\|^2 &= \|\boldsymbol{\theta}_\delta\|^2 + 2\delta\left(v_{j_1}^2 - \frac{v_{j_1}w_{j_1}v_{j_2}}{w_{j_2}} - \frac{b_{j_1}}{1-\delta}\left(b_{j_1} - \frac{w_{j_1}b_{j_2}}{w_{j_2}}\right)\right) + O(\delta^2)\,. \qquad (11)\end{aligned}$$

By construction, $v_{j_1} > 0$ and $v_{j_2} < 0$ hence $-\frac{v_{j_1}w_{j_1}v_{j_2}}{w_{j_2}} > 0$. Moreover, $-\frac{b_{j_1}}{w_{j_1}} > 0$ and $w_{j_1} > 0$ so $-\frac{b_{j_1}}{1-\delta} > 0$ for $\delta < 1$. Lastly, recall that $j_2 > j_1 \implies -\frac{b_{j_2}}{w_{j_2}} \geq -\frac{b_{j_1}}{w_{j_1}} \implies b_{j_1} - \frac{w_{j_1}b_{j_2}}{w_{j_2}} \geq 0$. Overall, plugging these into Eq. (11) shows that $\|\boldsymbol{\theta}\|^2 > \|\boldsymbol{\theta}_\delta\|^2$ for small enough δ, contradicting the assumption the $\boldsymbol{\theta}$ is a local minimum.

Case (2). Since $N_{\boldsymbol{\theta}}(x_{i-1}) \geq 1$ and $N_{\boldsymbol{\theta}}(x_i) \leq -1$, $N'_{\boldsymbol{\theta}}$ must be negative somewhere along the segment (x_{i-1}, x_i). Yet, $N'_{\boldsymbol{\theta}}(z_1) > 0$ so there must exist $j_1 \in [n]$ such that $-\frac{b_{j_1}}{w_{j_1}} \in (x_{i-1}, z_1)$ and $v_{j_1} > 0$. Moreover, the assumption on z_1, z_2 implies that there exists $j_2 \in [n]$ such that $-\frac{b_{j_2}}{w_{j_2}} \in (z_1, z_2)$ and $v_{j_2} < 0$. For some small $\delta > 0$, consider the perturbed network

$$\begin{aligned}N_{\boldsymbol{\theta}_\delta}(x) &:= \sum_{j\in[n]\setminus\{j_1,j_2\}} v_j\sigma\left(w_j\cdot x + b_j\right) \qquad (12) \\ &\quad + (1-\delta)v_{j_1}\sigma\left(w_{j_1}\cdot x + b_{j_1}\right) + v_{j_2}\sigma\left(\left(w_{j_2} + \frac{\delta v_{j_1}w_{j_1}}{v_{j_2}}\right)\cdot x + \left(b_{j_2} + \frac{\delta v_{j_1}b_{j_1}}{v_{j_2}}\right)\right)\,.\end{aligned}$$

It is clear that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\delta\| \xrightarrow{\delta\to 0} 0$, and we will show that for small enough δ the network above still satisfies the margin condition, yet has smaller parameter norm. To see why the margin condition is not violated for small enough δ, notice that $N_{\boldsymbol{\theta}_\delta}(x) = N_{\boldsymbol{\theta}}(x)$ for all $x \notin (x_{i-1}, x_{i+1})$ so in particular $N_{\boldsymbol{\theta}_\delta}(x_l) = N_{\boldsymbol{\theta}}(x_l)$ for all $l \neq i$. Furthermore, a direct computation yields $N_{\boldsymbol{\theta}_\delta}(x_i) < N_{\boldsymbol{\theta}}(x_i) \leq 1$. As to the parameter norm, by a similar computation to that leading up to Eq. (11) we get that

$$\|\boldsymbol{\theta}\|^2 = \|\boldsymbol{\theta}_\delta\|^2 + 2\delta\left(v_{j_1}^2 - \frac{v_{j_1}w_{j_1}w_{j_2}}{v_{j_2}} - \frac{v_{j_1}b_{j_1}b_{j_2}}{v_{j_2}}\right) + O(\delta^2)\,. \qquad (13)$$

By construction, $v_{j_1}w_{j_1} > 0$ and $v_{j_2}w_{j_2} < 0$ hence $-\frac{v_{j_1}w_{j_1}w_{j_2}}{v_{j_2}} > 0$. Moreover, $-\frac{b_{j_1}}{w_{j_1}} > 0$ and $-\frac{b_{j_2}}{w_{j_2}} > 0$ so we also have $-\frac{v_{j_1}b_{j_1}b_{j_2}}{v_{j_2}} > 0$. Hence, Eq. (13) shows that $\|\boldsymbol{\theta}\|^2 > \|\boldsymbol{\theta}_\delta\|^2$ for small enough δ, contradicting the assumption the $\boldsymbol{\theta}$ is a local minimum.

Case (3). The assumption implies that there exists $j_1 \in [n]$ such that $-\frac{b_{j_1}}{w_{j_1}} = z \in (x_i, x_{i+1})$, $v_{j_1} > 0$ and $-\frac{b_{j_1+1}}{w_{j_1+1}} \geq x_{i+1}$. Let $j_2 > j_1$ be the minimal index $j \in \{j_1+1, \ldots, n\}$ such that $v_j < 0$ (see Footnote 7). Consider the perturbed netowrk as in Eq. (10) and continue the proof as in Case (1) verbatim.

□

Lemma A.7. *$N_{\boldsymbol{\theta}}(x_i) = -1$ and $N_{\boldsymbol{\theta}}(x_{i+1}) = 1$.*

Proof of Lemma A.7. Recall that $N_{\boldsymbol{\theta}}(x_i) \leq -1$, so assume towards contradiction that $N_{\boldsymbol{\theta}}(x_i) < -1$. Note that since $N_{\boldsymbol{\theta}}(x_{i-1}) \geq 1$, $N_{\boldsymbol{\theta}}(\cdot)$ must decrease along (x_{i-1}, x_i), so there must exist $j \in [n]$ with $v_j < 0$ and $-\frac{b_j}{w_j} < x_i$. Denote by $j_1 \in [n]$ the maximal such index. Similarly, since $N_{\boldsymbol{\theta}}(x_{i+1}) \geq 1$ there must exist $j_2 \in [n]$ with $v_{j_2} > 0$ and $x_{i-1} < -\frac{b_{j_2}}{w_{j_2}} < x_{i+1}$. Consider the perturbed network

$$N_{\boldsymbol{\theta}_\delta}(x) := \sum_{j \in [n] \setminus \{j_1, j_2\}} v_j \sigma\left(w_j \cdot x + b_j\right) + (1-\delta) v_{j_1} \sigma\left(w_{j_1} \cdot x + b_{j_1}\right) + v_{j_2} \sigma\left(\left(w_{j_2} + \frac{\delta v_{j_1} w_{j_1}}{v_{j_2}}\right) \cdot x + \left(b_{j_2} + \frac{\delta v_{j_1} b_{j_1}}{v_{j_2}}\right)\right) \tag{14}$$

for some small $\delta > 0$. It is clear that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\delta\| \xrightarrow{\delta \to 0} 0$, and we will show that for small enough δ the network above still satisfies the margin condition, yet has smaller parameter norm. To see why the margin condition is not violated for small enough δ, first notice that $N_{\boldsymbol{\theta}_\delta}(x) = N_{\boldsymbol{\theta}}(x)$ for all $x \notin (x_{i-1}, x_{i+1})$. Furthermore, $N_{\boldsymbol{\theta}_\delta}(x_i)$ is continuous with respect to δ so $N_{\boldsymbol{\theta}}(x_i) < -1$ implies that $N_{\boldsymbol{\theta}_\delta}(x_i) < -1$ for small enough δ. As to the parameter norm, by a similar computation to that leading up to Eq. (11) we get that

$$\|\boldsymbol{\theta}\|^2 = \|\boldsymbol{\theta}_\delta\|^2 + 2\delta \left(v_{j_1}^2 - \frac{v_{j_1} w_{j_1} w_{j_2}}{v_{j_2}} - \frac{v_{j_1} b_{j_1} b_{j_2}}{v_{j_2}}\right) + O(\delta^2)\,. \tag{15}$$

By construction, $v_{j_1} < 0$ and $v_{j_2} > 0$ hence $-\frac{v_{j_1} w_{j_1} w_{j_2}}{v_{j_2}} > 0$. Moreover, $-\frac{b_{j_1}}{w_{j_1}}$ and $-\frac{b_{j_2}}{w_{j_2}} > 0$ so we also have $-\frac{v_{j_1} b_{j_1} b_{j_2}}{v_{j_2}} > 0$. Hence, Eq. (15) shows that $\|\boldsymbol{\theta}\|^2 > \|\boldsymbol{\theta}_\delta\|^2$ for small enough δ, contradicting the assumption the $\boldsymbol{\theta}$ is a local minimum.

Having proved that $N_{\boldsymbol{\theta}}(x_i) = -1$, we turn to prove that $N_{\boldsymbol{\theta}}(x_{i+1}) = 1$. Knowing that $N_{\boldsymbol{\theta}}(x_{i+1}) \geq 1$, we assume towards contradiction that $N_{\boldsymbol{\theta}}(x_{i+1}) > 1$. Recalling that $N_{\boldsymbol{\theta}}(x_{i-1}) \geq 1$, $N_{\boldsymbol{\theta}}(x_i) \leq 1$ and that $N_{\boldsymbol{\theta}}(\cdot)$ is linear along (x_i, x_{i+1}) due to Lemma A.6, we conclude that there must exist $j_1 \in [n]$ such that $v_{j_1} > 0$ and $x_{i-1} < -\frac{b_{j_1}}{w_{j_1}} \leq x_i$. Denote by $j_2 > j_1$ the minimal index such that $v_{j_2} < 0$. Consider once again the perturbed network in Eq. (14) (only now j_1, j_2 are different, as we just described). The same argument as in the previous part of the proof shows that for small enough δ the network above still satisfies the margin condition, while Eq. (15) once again implies that $\|\boldsymbol{\theta}\|^2 > \|\boldsymbol{\theta}_\delta\|^2$ for small enough δ – contradicting the assumption the $\boldsymbol{\theta}$ is a local minimum.

□

Overall, combining Lemma A.6 and Lemma A.7 finishes the proof of Proposition A.4. □

Proof of Proposition A.5. The proof is essentially the same as Case (1) in the proof of Proposition A.4.

Throughout the proof we fix $i \in [m-1]$ for which the conditions described in the proposition hold, and we assume without loss of generality that the neurons are ordered with respect to their activation point: $-\frac{b_1}{w_1} \leq -\frac{b_2}{w_2} \leq \cdots \leq -\frac{b_n}{w_n}$. As in the proof of Proposition A.4, we may assume without loss of generality that $w_1, \ldots, w_n \geq 0$. Moreover, as explained there as a consequence of Eq. (9), we observe that $N'_{\boldsymbol{\theta}}(z_1) < N'_{\boldsymbol{\theta}}(z_2)$ for some $z_1 < z_2$ implies the existence of $j_1 \in [n]$ with $v_{j_1} > 0$ and $z_1 < -\frac{b_{j_1}}{w_{j_1}} < z_2$. Similarly, $N'_{\boldsymbol{\theta}}(z_1) > N'_{\boldsymbol{\theta}}(z_2)$ implies that there exists $j_2 \in [n]$ with $v_{j_2} < 0,\ z_1 < -\frac{b_{j_2}}{w_{j_2}} < z_2$. With this choice of j_1, j_2, the proof continues as in Case (1) in the proof of Proposition A.4 verbatim.

□

B Proofs of benign overfitting

B.1 Proof of Theorem 4.1

We sample a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and labels $y_1, \ldots, y_m \sim \mathcal{D}_y$, for $p \leq c_1$ for some universal constant c_1. We first prove that the following properties holds with probability $> 1 - \delta$:

1. For every $i, j \in [m]$, $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq \sqrt{\frac{2\log\left(\frac{3m^2}{\delta}\right)}{d}}$
2. $\|XX^\top\| \leq C$ where C is some universal constant, and X is a matrix whose rows are equal to $\mathbf{x}_i$.
3. $|I_-| \leq \frac{3pm}{2}$.

Lemma B.1. *Let $\delta > 0$, assume that we sample $\mathbf{x}_1, \ldots \mathbf{x}_m \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and $y_1, \ldots, y_m \sim \mathcal{D}_y$ for $p \leq c_1$, and that $m > c_2 \frac{\log\left(\frac{1}{\delta}\right)}{p}$, for some universal constant $c_1, c_2 > 0$. Then, with probability $> 1 - \delta$ properties 1, 2, 3 holds.*

Proof. By Lemma D.1 we have that $\Pr\left[|\mathbf{x}_i^\top \mathbf{x}_j| \geq \sqrt{\frac{2\log\left(\frac{1}{\delta'}\right)}{d}}\right] \leq \delta'$. Take $\delta' = \frac{\delta}{3m^2}$, and use union bound over all $i, j \in [m]$ with $i \neq j$. This shows that for every $i \neq j$ we have:

$$\Pr\left[|\mathbf{x}_i^\top \mathbf{x}_j| \geq \sqrt{\frac{2\log\left(\frac{3m^2}{\delta}\right)}{d}}\right] \leq \frac{\delta}{3}\,.$$

This proves Property 1. Next, set X to be the matrix whose rows are equal to $\mathbf{x}_i$. By Lemma D.2 there is a constant $c' > 0$ such that:

$$\Pr\left[\|XX^\top - I\| \geq \frac{c'}{d}\left(\sqrt{\frac{d + \log\left(\frac{3}{\delta}\right)}{m}} + \frac{d + \log\left(\frac{3}{\delta}\right)}{m}\right)\right] \leq \frac{\delta}{3}\,. \tag{16}$$

We can also bound $\|XX^\top\| \leq \|I\| + \|XX^\top - I\| \leq 1 + \|XX^\top - I\|$. Combining both bounds and using the assumption that $m \geq \log\left(\frac{3}{\delta}\right)$ we get that there is a universal constant $C > 0$ such that $\Pr\left[\|XX^\top\| > C\right] \leq \frac{\delta}{3}$. This proves Property 2.

Finally, using Bernstein's inequality over the choice of the labels y_i have that:

$$\Pr\left[|I_-| \geq \frac{3pm}{2}\right] \leq \exp\left(-\frac{p^2m^2/4}{mp(1-p) + mp/6}\right) \leq \exp\left(-\frac{pm}{5}\right)\,,$$

where we used that $p \leq 1$. Hence, if $m \geq c_2 \frac{\log\left(\frac{1}{\delta}\right)}{p}$ for some universal constant $c_2 \geq 0$, then $|I_-| \leq \frac{3pm}{2}$. Applying union over those three arguments proves the lemma. □

From now on we condition on the event that properties 1, 2, 3 hold, and our bounds will depend on the probability of this event.

In the following lemma we show that if $\boldsymbol{\theta}$ converges to a solution of the max margin problem and the bias terms are relatively small, then the norm of $\boldsymbol{\theta}$ is relatively large:

Lemma B.2. *Assume that $m \geq c_2 \log\left(\frac{3}{\delta}\right)$ and that $\boldsymbol{\theta} = (\mathbf{w}_j, v_j, b_j)_{j=1}^n$ is a solution to the max margin problem of Eq. (1) with $\sum_{j=1}^n v_j \sigma(b_j) \leq \frac{1}{2}$. Then, $\sum_{j=1}^n \|\mathbf{w}_j\|^2 + v_j^2 + b_j^2 \geq C\sqrt{|I_+|}$ where C is some universal constant.*

Proof. Take $i \in [m]$, then we have that $N_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq 1$. By our assumption, this implies that:

$$\begin{aligned}
\frac{1}{2} &\leq N_{\boldsymbol{\theta}}(\mathbf{x}_i) - \sum_{j=1}^{n} v_j \sigma(b_j) \\
&= \sum_{j=1}^{n} v_j \sigma(\mathbf{w}_j^\top \mathbf{x}_i + b_j) - \sum_{j=1}^{n} v_j \sigma(b_j) \\
&\leq \sum_{j=1}^{n} |v_j| \cdot \left|\sigma(\mathbf{w}_j^\top \mathbf{x}_i + b_j) - \sigma(b_j)\right| \\
&\leq \sum_{j=1}^{n} |v_j| \cdot |\mathbf{w}_j^\top \mathbf{x}_i| \\
&\leq \sqrt{\sum_{j=1}^{n} v_j^2} \sqrt{\sum_{j=1}^{n} (\mathbf{w}_j^\top \mathbf{x}_i)^2} \,,
\end{aligned}$$

where in the last inequality we used Cauchy-Schwartz. Denote by $S := \sum_{j=1}^{n} \|\mathbf{w}_j\|^2 + v_j^2 + b_j^2$. Combining the above and that $\sqrt{\sum_{j=1}^{n} v_j^2} \leq \sqrt{S}$ we get:

$$\sum_{j=1}^{n} (\mathbf{w}_j^\top \mathbf{x}_i)^2 \geq \frac{1}{4S} \,.$$

We sum the above inequality for every $i \in I_+$ to get:

$$\begin{aligned}
\frac{|I_+|}{4S} &\leq \sum_{j=1}^{n} \sum_{i=1}^{m} (\mathbf{w}_j^\top \mathbf{x}_i)^2 \\
&= \sum_{j=1}^{n} \sum_{i=1}^{m} \mathbf{w}_j^\top (\mathbf{x}_i \mathbf{x}_i^\top) \mathbf{w}_j \\
&\leq \sum_{j=1}^{n} \|\mathbf{w}_j\|^2 \cdot \left\| \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \right\| \\
&\leq \sum_{j=1}^{n} \|\mathbf{w}_j\|^2 \cdot C \leq S \cdot C
\end{aligned}$$

where in the second to last to last inequality we used the Property 2 for some constant $C > 0$, and in the last inequality we used that $\sum_{j=1}^{n} \|\mathbf{w}_j\|^2 \leq S$. Rearranging the above terms yields: $S \geq \sqrt{\frac{|I_+|}{12C}}$. □

We now prove a lemma which constructs a specific solution that achieves a norm bound that depends on $|I_+|$.

Lemma B.3. *Assume $d \geq 50m^2 \log\left(\frac{3m^2}{\delta}\right)$ and $p \leq \frac{1}{4}$. There exists weights $\boldsymbol{\theta} = (\mathbf{w}_j, v_j, b_j)_{j=1}^{n}$ that attain a margin of at least 1 on every sample and have $\sum_{j=1}^{n} \|\mathbf{w}_j\|^2 + v_j^2 + b_j^2 \leq 9\sqrt{|I_-|}$.*

Proof. Assume without loss of generality that n is even (otherwise fix the last neuron to be 0). We consider the following weight assignment: For every $j \leq \frac{n}{2}$, consider $\mathbf{w}_j = -\sqrt{\frac{4}{n\sqrt{|I_-|}}} \sum_{i \in I_-} \mathbf{x}_i$, $v_j = 2\sqrt{\frac{\sqrt{|I_-|}}{n}}$ and $b_j = \sqrt{\frac{4}{n\sqrt{|I_-|}}}$. For every $j > \frac{n}{2}$, consider $\mathbf{w}_j = \sqrt{\frac{4}{n\sqrt{|I_-|}}} \sum_{i \in I_-} \mathbf{x}_i$, $v_j = -2\sqrt{\frac{\sqrt{|I_-|}}{n}}$ and $b_j = 0$.

We first show that this solution attains a margin of at least 1. For every $i \in I_+$ we have:

$$\begin{aligned}
N_{\boldsymbol{\theta}}(\mathbf{x}_i) =& \sum_{j=1}^{n} v_j \sigma(\mathbf{w}_j^\top \mathbf{x}_i + b_j) \\
=& \sum_{j \le n/2} 2\sqrt{\frac{\sqrt{|I_-|}}{n}} \sigma\left(\sqrt{\frac{4}{n\sqrt{|I_-|}}} - \sqrt{\frac{4}{n\sqrt{|I_-|}}} \sum_{r \in I_-} \mathbf{x}_r^\top \mathbf{x}_i\right) \\
& - \sum_{j > n/2} 2\sqrt{\frac{\sqrt{|I_-|}}{n}} \sigma\left(\sqrt{\frac{4}{n\sqrt{|I_-|}}} \sum_{r \in I_-} \mathbf{x}_r^\top \mathbf{x}_i\right) \\
\ge& \sum_{j \le n/2} \frac{4}{n} \sigma\left(1 - \frac{|I_-|\sqrt{2\log\left(\frac{3m^2}{\delta}\right)}}{\sqrt{d}}\right) - \sum_{j > n/2} \frac{4}{n} \sigma\left(\frac{|I_-|\sqrt{2\log\left(\frac{3m^2}{\delta}\right)}}{\sqrt{d}}\right) \\
\ge& \frac{n}{2} \cdot \frac{4}{n} \cdot \left(1 - \frac{1}{10}\right) - \frac{n}{2} \cdot \frac{4}{n} \cdot \frac{1}{10} \ge 1 \,,
\end{aligned}$$

where we Property 1, that $p \le \frac{1}{4}$ hence by Property 3 $|I_-| \le \frac{m}{2}$ and our assumption on m and d. For $i \in I_-$ we have:

$$\begin{aligned}
N_{\boldsymbol{\theta}}(\mathbf{x}_i) =& \sum_{j=1}^{n} v_j \sigma(\mathbf{w}_j^\top \mathbf{x}_i + b_j) \\
=& \sum_{j \le n/2} 2\sqrt{\frac{\sqrt{|I_-|}}{n}} \sigma\left(\sqrt{\frac{4}{n\sqrt{|I_-|}}} - \sqrt{\frac{4}{n\sqrt{|I_-|}}} \sum_{r \in I_-} \mathbf{x}_r^\top \mathbf{x}_i\right) \\
& - \sum_{j > n/2} 2\sqrt{\frac{\sqrt{|I_-|}}{n}} \sigma\left(\sqrt{\frac{4}{n\sqrt{|I_-|}}} \sum_{r \in I_-} \mathbf{x}_r^\top \mathbf{x}_i\right) \\
=& \sum_{j \le n/2} \frac{4}{n} \sigma\left(-\sum_{r \in I_- \setminus \{i\}} \mathbf{x}_r^\top \mathbf{x}_i\right) - \sum_{j > n/2} \frac{4}{n} \sigma\left(1 + \sum_{r \in I_- \setminus \{i\}} \mathbf{x}_r^\top \mathbf{x}_i\right) \\
\le& \sum_{j \le n/2} \frac{4}{n} \sigma\left(-\frac{(|I_-|-1)\sqrt{2\log\left(\frac{3m^2}{\delta}\right)}}{\sqrt{d}}\right) - \sum_{j > n/2} \frac{4}{n} \sigma\left(1 - \frac{(|I_-|-1)\sqrt{2\log\left(\frac{3m^2}{\delta}\right)}}{\sqrt{d}}\right) \\
\le& \frac{n}{2} \cdot \frac{4}{n} \cdot \frac{1}{10} - \frac{n}{2} \cdot \frac{4}{n} \cdot \left(1 - \frac{1}{10}\right) \le -1 \,,
\end{aligned}$$

where again we used properties 1 and 3, that $p \le \frac{1}{4}$ hence $|I_-| \le \frac{m}{2}$ and our assumption on m and d. This shows that this is indeed a feasible solution. We turn to calculate the norm of this solution. First we bound the following:

$$\begin{aligned}
\left\| \sum_{i \in I_-} \mathbf{x}_i \right\|^2 =& \sum_{i \in I_-} \|\mathbf{x}_i\|^2 + \sum_{i \ne j,\, i,j \in I_-} \mathbf{x}_i^\top \mathbf{x}_j \\
& \le |I_-| + |I_-|^2 \cdot \frac{\sqrt{2\log\left(\frac{3m^2}{\delta}\right)}}{\sqrt{d}} \le |I_-| \cdot \frac{11}{10} \,,
\end{aligned}$$

where we used Property 1 and the assumption on m and d. We now use the above calculation to bound the norm of our solution:

$$\sum_{j=1}^{n} \|\mathbf{w}_j\|^2 + v_j^2 + b_j^2 = n \cdot \frac{4}{n\sqrt{|I_-|}} \left\| \sum_{i \in I_-} \mathbf{x}_i \right\|^2 + \frac{2n\sqrt{|I_-|}}{n} + \frac{n}{2} \cdot \frac{4}{n\sqrt{|I_-|}}$$
$$\leq \frac{44\sqrt{|I_-|}}{10} + 2\sqrt{|I_-|} + \frac{2}{\sqrt{|I_-|}} \leq 9\sqrt{|I_-|}$$

□

We are now ready to prove the main theorem of this subsection:

Proof of Theorem 4.1. Denote by $K := \sum_{j=1}^{n} \|\mathbf{w}_j\|^2 + v_j^2 + b_j^2$, and assume that $K \leq \frac{a}{\sqrt{p}} \|\boldsymbol{\theta}^*\|^2$, where a will be chosen later and $\boldsymbol{\theta}^*$ is a solution to the max margin solution from Eq. (1). Assume on the way of contradiction that $\sum_{j=1}^{n} v_j \sigma(b_j) \leq \frac{1}{2}$, then by Lemma B.2 we know that: $K \geq C\sqrt{|I_+|} \geq C\sqrt{\left(1 - \frac{3p}{2}\right) m}$. On the other hand, by Lemma B.3 we know that $\|\boldsymbol{\theta}^*\| \leq 9\sqrt{|I_-|} \leq 9\sqrt{\frac{3p}{2} m}$. Combining this with the assumption we have on K we get that:

$$C\sqrt{\left(1 - \frac{3p}{2}\right) m} \leq \frac{9a}{\sqrt{p}} \sqrt{\frac{3pm}{2}} \,.$$

Picking a to be some constant with $a < \frac{C\sqrt{2}}{18\sqrt{3}}$ contradicts the above inequality. Thus, there exists a constant $c_4 := \frac{a}{2}$ such that if $K \leq \frac{c_4}{\sqrt{p}} S$ then $\sum_{j=1}^{n} v_j \sigma(b_j) > \frac{1}{2}$.

Suppose we sample $\mathbf{x} \sim \mathcal{N}\left(0, \frac{1}{d} I\right)$, then we have:

$$\begin{aligned}
N_{\boldsymbol{\theta}}(\mathbf{x}) &= \sum_{j=1}^{n} v_j \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) \\
&= \sum_{j=1}^{n} v_j \sigma(b_j) - \left(\sum_{j=1}^{n} v_j \sigma(b_j) - \sum_{j=1}^{n} v_j \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) \right) \\
&\geq \frac{1}{2} - \left(\sum_{j=1}^{n} v_j \sigma(b_j) - \sum_{j=1}^{n} v_j \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) \right) \\
&\geq \frac{1}{2} - \sum_{j=1}^{n} |v_j| \cdot |\mathbf{w}_j^\top \mathbf{x}| \\
&\geq \frac{1}{2} - \sqrt{\sum_{j=1}^{n} v_j^2} \sqrt{\sum_{j=1}^{n} (\mathbf{w}_j^\top \mathbf{x})^2} \,. \qquad (17)
\end{aligned}$$

We will now bound the terms of the above equation. Note that $\sum_{j=1}^{n} v_j^2 \leq K \leq 9\sqrt{m}$. For the second term, we denote by $\bar{\mathbf{w}}_j := \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$, and write:

$$\begin{aligned}
\sum_{j=1}^{n} (\mathbf{w}_j^\top \mathbf{x})^2 &= \sum_{j=1}^{n} \|\mathbf{w}_j\|^2 (\bar{\mathbf{w}}_j^\top \mathbf{x})^2 \\
&\leq \max_{j \in [n]} (\bar{\mathbf{w}}_j^\top \mathbf{x})^2 \cdot \sum_{j=1}^{n} \|\mathbf{w}_j\|^2 \,.
\end{aligned}$$

Again, we have that $\sum_{j=1}^{n} \|\mathbf{w}_j\|^2 \leq K \leq 9\sqrt{m}$. By using the stationarity KKT condition from Eq. (2) , we get that $\bar{\mathbf{w}}_j \in \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, thus we can write $\bar{\mathbf{w}}_j = \sum_{i=1}^{m} \alpha_{i,j} \mathbf{x}_i = X \boldsymbol{\alpha}_j$ where

X is a matrix with rows equal to $\mathbf{x}_i$, and $(\boldsymbol{\alpha}_j)_i = \alpha_{i,j}$. We will bound $a := \arg\max_i |\alpha_{i,j}|$:

$$\begin{aligned}
1 = \|\bar{\mathbf{w}}_j\|^2 &= \left\|\sum_{i=1}^m \alpha_{i,j}\mathbf{x}_i\right\|^2 \\
&= \|X\boldsymbol{\alpha}_j\|^2 = \boldsymbol{\alpha}_j^\top X^\top X \boldsymbol{\alpha}_j \\
&= \boldsymbol{\alpha}_j^\top I \boldsymbol{\alpha}_j + \boldsymbol{\alpha}_j^\top \left(X^\top X - I\right)\boldsymbol{\alpha}_j \\
&\geq \|\boldsymbol{\alpha}_j\|^2 - \|\boldsymbol{\alpha}_j\|^2 \|X^\top X - I\| \\
&= \|\boldsymbol{\alpha}_j\|^2 - \|\boldsymbol{\alpha}_j\|^2 \|XX^\top - I\| \,,
\end{aligned}$$

where the last equality is true by using the SVD decomposition of X. Namely, write $X = USV^\top$, then $\|X^\top X - I\| = \|VS^2V^\top - I\| = \|S^2 - I\| = \|U^\top S^2 U - I\| = \|XX^\top - I\|$. Note that in Lemma B.1, Eq. (16) we have shown that $\|XX^\top - I\| \leq c'$ for some constant c'. Note that we can choose m large enough such that $c' \leq \frac{1}{2}$. In total, this shows that $\|\boldsymbol{\alpha}_j\|^2 \leq c''$ for some constant c''.

We now use Lemma D.1 and the union bound to get that with probability $> 1 - \epsilon$ we have for every $i \in I$ that $|\mathbf{x}^\top \mathbf{x}_i| \leq \sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}}$. We condition on this event from now on. Applying both bounds we get:

$$\begin{aligned}
\max_{j\in[n]}(\bar{\mathbf{w}}_j^\top \mathbf{x}) &\leq \max_{j\in[n]}\left(\sum_{i=1}^m \alpha_{i,j}|\mathbf{x}_i^\top \mathbf{x}|\right) \\
&\leq \sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \max_{j\in[n]} \|\boldsymbol{\alpha}_j\| \\
&\leq c''\sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \,,
\end{aligned}$$

Plugging in the bounds above to Eq. (17) we get:

$$\begin{aligned}
N_{\boldsymbol{\theta}}(\mathbf{x}) &\geq \frac{1}{2} - \sqrt{\sum_{j=1}^n v_j^2}\sqrt{\sum_{j=1}^n (\mathbf{w}_j^\top \mathbf{x})^2} \\
&\geq \frac{1}{2} - 9\sqrt{m}\cdot 9\sqrt{m}\cdot c''\sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \\
&\geq \frac{1}{2} - \frac{81mc''\sqrt{2\log\left(\frac{m}{\epsilon}\right)}}{\sqrt{d}} \,.
\end{aligned}$$

By choosing a constant $c_3 > 0$ large enough such that if $d \geq c_3 m^2\sqrt{\log\left(\frac{m}{\epsilon}\right)}$ then , then $\frac{81mc''\sqrt{2\log\left(\frac{m}{\epsilon}\right)}}{\sqrt{d}} \leq \frac{1}{4}$ we get that $N_{\boldsymbol{\theta}}(\mathbf{x}) > 0$. □

B.2 Proof of Theorem 4.3

We sample a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and labels $y_1, \ldots, y_m \sim \mathcal{D}_y$, for $p \leq c_1$ for some universal constant c_1. We first prove that the following properties holds with probability $> 1 - \delta$:

1. For every $i, j \in [m]$, $|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \leq \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}$
2. $|I_-| \leq \frac{c_1 m}{n^2}$.

Lemma B.4. *Let $\delta > 0$, assume that we sample $\mathbf{x}_1, \ldots \mathbf{x}_m \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and $y_1, \ldots, y_m \sim \mathcal{D}_y$ for $p \leq \frac{c}{n^2}$, and that $m > c'\log\left(\frac{2}{\delta}\right)$, for some universal constant $c, c' > 0$. Then, with probability $> 1 - \delta$ properties 1 and 2 holds.*

The proof is the same as in Lemma B.1 so we will not repeat it for conciseness. The only different is that here we only need $|I_-|$ to be smaller than $\frac{c_1 m}{n^2}$ which is independent of p. Hence, for $p \leq \frac{c}{n^2}$ we get this concentration where m depends only on the probability. From now on we condition on the event that properties 1 and 2 hold, and our bounds will depend on the probability that this event happens.

In this section we consider a networks of the form:

$$N_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j)$$

That is, the output weights are all fixed to be ± 1, equally divided between the neurons. We will use the stationarity condition (Eq. (2)) freely throughout the proof. For our network it means that we can write for every $j \in [n]$:

$$\mathbf{w}_j = v_j \sum_{i \in [m]} \lambda_i \sigma'_{i,j} y_i \mathbf{x}_i$$
$$b_j = v_j \sum_{i \in [m]} \lambda_i \sigma'_{i,j} y_i \ ,$$

where $\sigma'_{i,j} = \mathbb{1}(\mathbf{w}_j^\top \mathbf{x}_i + b_j >)$, $v_j = 1$ for $j \in \{1, \ldots, n/2\}$ and $v_j = -1$ for $j \in \{n/2+1, \ldots, n\}$.

The next lemma shows that all the biases are positive. Note that this lemma relies only on that the dimension is large enough, and that Property 1 of the data holds:

Lemma B.5. *Assume that $d \geq 8m^4 \log\left(\frac{2m^2}{\delta}\right)$. Then for every $j \in [n]$ we have $b_j \geq 0$.*

Proof. Assume on the way of contradiction that $b_j < 0$ for some $j \in [n]$. We assume without loss of generality that $j \in \{1, \ldots, n/2\}$, the proof for $j \in \{n/2+1, \ldots, n\}$ is done similarly. Denote by $I'_+ = \{i \in I_+ := \sigma_{i,j} = 1\}$ and $I'_- = \{i \in I_- := \sigma_{i,j} = 1\}$. We can write:

$$b_j = \sum_{i \in [m]} \lambda_i \sigma'_{i,j} y_i = \sum_{i \in I'_+} \lambda_i - \sum_{i \in I'_-} \lambda_i \ .$$

Note that $\lambda_i \geq 0$ for every i, hence I'_- is non empty, otherwise $b_j \geq 0$. Take $\lambda_r = \arg\max_{i \in I'_-} \lambda_i$, we have:

$$\begin{aligned} 0 < w_j^\top \mathbf{x}_r + b_j &= b_j + \sum_{i \in I'_+ \cup I'_-} \lambda_i y_i \mathbf{x}_r^\top \mathbf{x}_i \\ &< -\lambda_r + \sum_{i \in I'_+ \cup I'_- \setminus \{r\}} \lambda_i y_i \mathbf{x}_i^\top \mathbf{x}_r \ , \end{aligned}$$

where we used that $b_j < 0$ and $\|\mathbf{x}_r\| = 1$. Rearranging the terms and using Property 1:

$$\begin{aligned} \lambda_r &< \sum_{i \in I'_+ \cup I'_- \setminus \{r\}} \lambda_i |y_i \mathbf{x}_i^\top \mathbf{x}_r| \\ &\leq \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \sum_{i \in I'_+ \cup I'_- \setminus \{r\}} \lambda_i \\ &\leq \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \sum_{i \in I'_+ \cup I'_-} \lambda_i \ . \end{aligned}$$

Denote $\lambda_s := \arg\max_{i \in I'_+ \cup I'_-} \lambda_i$, then from the above we have shown that $\lambda_r \leq \lambda_s m \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}$, since $|I'_+ \cup I'_-| \leq m$. In particular, $s \in I'_+$, otherwise, $\lambda_r = \lambda_s$ which means that $\lambda_r < 0$ since

$m\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} < 1$ which is a contradiction. We can now write:

$$\begin{aligned} b_j &= \sum_{i\in I'_+} \lambda_i - \sum_{i\in I'_-} \lambda_i \\ &\geq \sum_{i\in I'_+} \lambda_i - \sum_{i\in I'_-} \lambda_r \\ &\geq \lambda_s - \lambda_s m^2 \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \\ &= \lambda_s \left(1 - m^2\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\right) . \end{aligned}$$

By our assumption on d, we have that $m^2\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$, hence $b_j \geq \frac{\lambda_s}{2}$. But, by our assumption, $b_j < 0$ which is a contradiction since $\lambda_s \geq 0$. □

The following lemma is a general property of the KKT conditions:

Lemma B.6. *Let $i \in I_+$ (resp. $i \in I_-$) with $\lambda_i > 0$. Then, there is a neuron k with positive output weight (resp. negative output weight) s.t $\sigma'_{i,k} = 1$.*

Proof. We prove it for $i \in I_+$, the other case is similar. Assume otherwise, that is for every neurons with positive output j we have $\sigma'_{i,j} = 0$, that is $\mathbf{w}_j^\top \mathbf{x}_i + b_j \leq 0$ by the definition of $\sigma'_{i,j}$. This means that $N_{\boldsymbol{\theta}}(\mathbf{x}_i) \leq 0$, since all the positive neurons are inactive on $\mathbf{x}_i$, which is a contradiction to $N_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq 1$. □

The next lemma shows that if the bias terms are smaller than $\frac{1}{4}$, then the λ_i's for $i \in I_+$ cannot be too small.

Lemma B.7. *Assume that $d \geq 32m^4n^2\log\left(\frac{2m^2}{\delta}\right)$, and that $\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \leq \frac{1}{4}$. Then, for every $i \in I_+$ we have $\lambda_i \geq \frac{1}{4n}$.*

Proof. Take $r \in I_+$, we have:

$$\begin{aligned} 1 \leq & N_{\boldsymbol{\theta}}(\mathbf{x}_r) = \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) \\ =& \sum_{j=1}^{n/2} \sigma\left(\lambda_r \sigma'_{r,j} + \sum_{i\in I\setminus\{r\}} \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_r + b_j\right) \\ & - \sum_{j=n/2+1}^{n} \sigma\left(-\lambda_r \sigma'_{r,j} - \sum_{i\in I\setminus\{r\}} \lambda_i y_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_r + b_j\right) \\ \leq & \sum_{j=1}^{n/2} \sigma\left(\lambda_r + \sum_{i\in I\setminus\{r\}} \lambda_i \sigma'_{i,j} |\mathbf{x}_i^\top \mathbf{x}_r| + b_j\right) - \sum_{j=n/2+1}^{n} \sigma\left(-\lambda_r - \sum_{i\in I\setminus\{r\}} \lambda_i \sigma'_{i,j} |\mathbf{x}_i^\top \mathbf{x}_r| + b_j\right) \end{aligned} \tag{18}$$

We will bound the two terms above. For the first term, note that all the terms inside the ReLU are positive, since by Lemma B.5 we have $b_j \geq 0$, and $\lambda_i \geq 0$ by Eq. (4) hence we can remove the ReLU

function. Using Property 1 we can bound:

$$\sum_{j=1}^{n/2}\sigma\left(\lambda_r+\sum_{i\in I\setminus\{r\}}\lambda_i\sigma'_{i,j}|\mathbf{x}_i^\top\mathbf{x}_r|+b_j\right)\leq\frac{n}{2}\lambda_r+\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\sum_{j=1}^{n/2}\sum_{i\in I\setminus\{r\}}\lambda_i\sigma'_{i,j}+\sum_{j=1}^{n/2}b_j$$
$$\leq\frac{n}{2}\lambda_r+\frac{n}{2}\cdot\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\sum_{i\in I\setminus\{r\}}\lambda_i+\sum_{j=1}^{n/2}b_j\ , \quad (19)$$

where we used that $\sigma'_{i,j}\leq 1$. For the second term in Eq. (18) we use the fact that the ReLU function is 1-Lipschitz to get that:

$$-\sum_{j=n/2+1}^{n}\sigma\left(-\lambda_r-\sum_{i\in I\setminus\{r\}}\lambda_i\sigma'_{i,j}|\mathbf{x}_i^\top\mathbf{x}_r|+b_j\right)$$
$$=-\sum_{j=n/2+1}^{n}\sigma\left(-\lambda_r-\sum_{i\in I\setminus\{r\}}\lambda_i\sigma'_{i,j}|\mathbf{x}_i^\top\mathbf{x}_r|+b_j\right)+\sum_{j=n/2+1}^{n}\sigma(b_j)-\sum_{j=n/2+1}^{n}\sigma(b_j)$$
$$\leq-\sum_{j=n/2+1}^{n}b_j+\sum_{j=n/2+1}^{n}\left|\lambda_r+\sum_{i\in I\setminus\{r\}}\lambda_i\sigma'_{i,j}|\mathbf{x}_i^\top\mathbf{x}_r|\right|$$
$$\leq-\sum_{j=n/2+1}^{n}b_j+\frac{n}{2}\lambda_r+\frac{n}{2}\cdot\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\sum_{i\in I\setminus\{r\}}\lambda_i\ , \quad (20)$$

where we again used Lemma B.5 to get that $b_j\geq 0$, and Property 1. Combining Eq. (19) and Eq. (20) with Eq. (18) we get that:

$$1\leq\sum_{j=1}^{n/2}b_j-\sum_{j=n/2+1}^{n}b_j+n\lambda_r+n\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\sum_{i\in I\setminus\{r\}}\lambda_i\ .$$

Rearranging the terms above, and using our assumption on the biases we get that:

$$\frac{3}{4}\leq n\lambda_r+n\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\sum_{i\in I\setminus\{r\}}\lambda_i\ .$$

If $\lambda_r\geq\frac{1}{4n}$ then we are done. Assume otherwise, then $n\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\sum_{i\in I\setminus\{r\}}\lambda_i\geq\frac{1}{2}$ and $\lambda_r<\frac{1}{4n}$. Denote $\lambda_s:=\arg\max_{i\in[m]}\lambda_i$, then we have that $\sum_{i\in[m]\setminus\{r\}}\lambda_i\leq m\lambda_s$, which by the above inequality means that $\lambda_s\geq\frac{\sqrt{d}}{2mn\sqrt{2\log\left(\frac{2m^2}{\delta}\right)}}$. We split into cases:

Case I. Assume $s\in I_+$. By Lemma B.6 there is $k\in\{1,\ldots,n/2\}$ with $\sigma'_{s,k}=1$. For the k-th neuron we have:

$$\sigma(\mathbf{w}_k^\top\mathbf{x}_s+b_k)=\sigma\left(\sum_{i\in[m]}y_i\lambda_i\sigma'_{i,k}\mathbf{x}_i^\top\mathbf{x}_s+b_k\right)$$
$$\geq\sigma\left(\lambda_s-\sum_{i\in I\setminus\{s\}}\lambda_i|\mathbf{x}_i^\top\mathbf{x}_s|+b_k\right)$$
$$\geq\sigma\left(\lambda_s-m\lambda_s\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}+b_k\right)\ , \quad (21)$$

and note that since $\lambda_s \geq 0$, $m\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} < 1$ and $b_j \geq 0$ this neuron is active. For every other neuron j with a positive output weight we have:

$$\begin{aligned}
\sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) - \sigma(b_j) &= \sigma\left(\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j\right) - \sigma(b_j) \\
&\geq \sigma\left(\sum_{i\in[m]} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j\right) - \sigma(b_j) \\
&\geq \sigma\left(-m\lambda_s\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} + b_j\right) - \sigma(b_j) \\
&\geq -m\lambda_s\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\,, \qquad (22)
\end{aligned}$$

where we used that σ is 1-Lipschitz and that λ_s is the largest among the λ_i's. For a neuron with a negative output weight, we can write:

$$\begin{aligned}
\sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) &= \sigma\left(-\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j\right) \\
&\leq \sigma\left(-\sum_{i\in I\setminus\{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j\right) \leq b_j + m\lambda_s\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}
\end{aligned}$$

In total, combining the above bound with Eq. (21) and Eq. (22) we have that:

$$\begin{aligned}
1 = N_{\boldsymbol{\theta}}(\mathbf{x}_s) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) \\
&= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) + \sum_{j=1}^{n/2} \sigma(b_j) - \sum_{j=1}^{n/2} \sigma(b_j) \\
&\geq -\lambda_s mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} + \lambda_s + \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j
\end{aligned}$$

By our assumption on d we have that $mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$. Hence, rearranging the terms above, we get that:

$$\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \leq 1 - \frac{\lambda_s}{2}\,.$$

But then, looking at the output of the network on λ_r (recall that by our assumption $\lambda_r \leq \frac{1}{4n}$) we get:

$$
\begin{aligned}
1 \leq N_{\boldsymbol{\theta}}(\mathbf{x}_r) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) \\
&= \sum_{j=1}^{n/2} \sigma\left(\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_r + b_j \right) - \sum_{j=n/2+1}^{n} \sigma\left(- \sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_r + b_j \right) + \\
&+ \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\leq \sum_{j=1}^{n/2} \sigma\left(\lambda_r + \sum_{i\in I\setminus\{r\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_r| + b_j \right) - \sum_{j=n/2+1}^{n} \sigma\left(- \sum_{i\in I\setminus\{r\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_r| + b_j \right) + \\
&+ \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\leq \frac{n\lambda_r}{2} + \lambda_s mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} + \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \\
&\leq \frac{1}{8} + \lambda_s mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} + 1 - \frac{\lambda_s}{2} \\
&\leq \frac{9}{8} + \lambda_s \left(mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} - \frac{1}{2} \right) . \qquad (23)
\end{aligned}
$$

By our assumption $mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{4}$, hence rearranging the terms above we get that $\lambda_s \leq \frac{1}{2}$ which is a contradiction to $\lambda_s \geq \frac{\sqrt{d}}{2mn\sqrt{2\log\left(\frac{2m^2}{\delta}\right)}} > 1$.

Case II. Assume $s \in I_-$. By Lemma B.6 there is $k \in \{n/2+1, \ldots, n\}$ with $\sigma'_{s,k} = 1$. Note that since $\lambda_s \neq 0$ we have that $N_{\boldsymbol{\theta}}(\mathbf{x}_s) = -1$ (i.e this sample is on the margin). We have that:

$$\begin{aligned}
-1 = N_{\boldsymbol{\theta}}(\mathbf{x}_s) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) \\
&= \sum_{j=1}^{n/2} \sigma\left(\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j \right) - \sum_{j=n/2+1}^{n} \sigma\left(- \sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j \right) + \\
&+ \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\leq \sum_{j=1}^{n/2} \sigma\left(\sum_{i\in I\setminus\{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j \right) - \sum_{j=n/2+1}^{n} \sigma\left(\lambda_s - \sum_{i\in I\setminus\{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j \right) + \\
&+ \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\leq \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j + \lambda_s mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} - \lambda_s \\
&\leq \frac{1}{4} + \lambda_s \left(mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} - 1 \right) ,
\end{aligned} \tag{24}$$

where we used that $mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$ and $b_j \geq 0$ by Lemma B.5, hence in the second to last inequality the term inside the ReLU is positive. In addition, we used that the ReLU function is 1-Lipschitz, and that λ_s is the largest amont the λ_i's. Since $mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$, rearranging the terms above give us that $\lambda_s \leq \frac{10}{4}$, which is a contradiction to that $\lambda_s \geq \frac{\sqrt{d}}{2mn\sqrt{2\log\left(\frac{2m^2}{\delta}\right)}} \geq 3$.

To conclude, both cases are not possible, hence $\lambda_r \geq \frac{1}{4n}$, which is true for every $r \in I_+$.

□

Lemma B.8. *Assume that* $d \geq 16m^4n^4 \log\left(\frac{2m^2}{\delta}\right)$, *and that* $\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \leq \frac{1}{4}$. *Then, for every* $i \in I_-$ *we have* $\lambda_i \leq \frac{10}{4}$.

Proof. Take $\lambda_s := \arg\max_{i\in[m]} \lambda_i$. We split into two cases:

Case I. Assume $s \in I_-$. Note that $\lambda_s > 0$, otherwise $\lambda_i = 0$ for every i, which means that $N_{\boldsymbol{\theta}}(\mathbf{x})$ is the zero function. Hence, $\mathbf{x}_s$ lies on the margin, and also by Lemma B.6 there is a neuron j with a negative output weight such that $\sigma'_{s,j} = 1$. By the same calculation as in Eq. (24) we have::

$$\begin{aligned}
-1 = N_{\boldsymbol{\theta}}(\mathbf{x}_s) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) \\
&\leq \frac{1}{4} + \lambda_s \left(mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} - 1 \right)
\end{aligned}$$

Using our assumption that $mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$ and rearranging the above terms we get: $\lambda_s \leq \frac{10}{4}$ which finishes the proof.

Case II. Assume $s \in I_+$. Denote by $\lambda_r := S \arg\max_{i \in I_-} \lambda_i$. By Lemma B.6 there is at least one neuron k with a positive output weight such that $\sigma'_{s,k} = 1$, for this neuron we can bound:

$$b_k = \sum_{i \in [m]} y_i \lambda_i \sigma'_{i,k} \geq \lambda_s - m\lambda_r \ .$$

Every neuron j with a negative output weight can be bounded similarly by:

$$b_j = -\sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j} \leq m\lambda_r \ .$$

Combining the above bounds, and using that $b_j \geq 0$ by Lemma B.5 we can bound:

$$\begin{aligned}
\frac{1}{4} &\geq \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \\
&\geq \lambda_s - m\lambda_r - \sum_{j=n/2+1}^{n} m\lambda_r \\
&\geq \lambda_s - mn\lambda_r \ .
\end{aligned}$$

Rearranging the terms we get that: $\lambda_s \leq mn\lambda_r + \frac{1}{4}$. If $\lambda_r \leq 1$ we are finished, otherwise $\lambda_s \leq 2mn\lambda_r$ and also $\mathbf{x}_r$ lies on the margin since $\lambda_r > 0$, hence. By doing a similar calculation to Case I we can write:

$$\begin{aligned}
-1 = N_{\boldsymbol{\theta}}(\mathbf{x}_r) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) \\
&= \sum_{j=1}^{n/2} \sigma\left(\sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_r + b_j \right) - \sum_{j=n/2+1}^{n} \sigma\left(-\sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_r + b_j \right) + \\
&+ \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\leq \sum_{j=1}^{n/2} \sigma\left(\sum_{i \in I \setminus \{r\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_r| + b_j \right) - \sum_{j=n/2+1}^{n} \sigma\left(\lambda_r - \sum_{i \in I \setminus \{r\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_r| + b_j \right) + \\
&+ \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\leq -\lambda_r + \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j + \lambda_s mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \\
&\leq -\lambda_r + \frac{1}{4} + 2\lambda_r m^2 n^2 \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \ .
\end{aligned}$$

By the assumption on d we have $2m^2n^2\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$. Thus, rearranging the terms above and using these bounds we get that $\lambda_r \leq \frac{10}{4}$.

□

Lemma B.9. *Assume that* $p \leq \frac{c_1}{n^2}$, $m \geq c_2 n \log\left(\frac{1}{\delta}\right)$ *and* $d \geq 32m^4n^4 \log\left(\frac{2m^2}{\delta}\right)$. *Then,* $\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j > \frac{1}{4}$.

Proof. Assume on the way of contradiction that $\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \leq \frac{1}{4}$. By Lemma B.7 we have that $\lambda_i \geq \frac{1}{4n}$ for every $i \in I_+$, and by Lemma B.8 we have that $\lambda_i \leq \frac{10}{4}$ for every $i \in I_-$. By

Lemma B.6, for every $i \in I_+$ there is some $j \in \{1, \ldots, n/2\}$ with $\sigma'_{i,j} = 1$. This means that:

$$\sum_{j=1}^{n/2} b_j = \sum_{j=1}^{n/2} \sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j}$$
$$\geq \frac{|I_+|}{4n} - \sum_{j=1}^{n/2} \sum_{i \in I_-} \lambda_i \geq \frac{|I_+|}{4n} - \frac{10n|I_-|}{8} .$$

In a similar manner, we can bound:

$$\sum_{j=n/2+1}^{n} b_j = \sum_{j=n/2+1}^{n} \sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j}$$
$$\geq \sum_{j=n/2+1}^{n} \sum_{i \in I_-} \lambda_i \geq -\frac{10n|I_-|}{8} .$$

Combining the two bounds above we get that:

$$\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \geq \frac{|I_+|}{4n} - \frac{10n|I_-|}{4} .$$

But, by our assumptions we have that $m \geq 2n$ (for a large enough constant), and by Lemma B.4 we have $|I_-| \leq \frac{cm}{n^2}$ and $|I_+| \geq \left(1 - \frac{c}{n^2}\right) m$ for some universal constant $c > 0$ which we will later choose. Plugging these bounds to the displayed equation above, we get that:

$$\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \geq \frac{1}{2} \cdot \left(1 - \frac{c}{n^2}\right) - \frac{10c}{2} \geq \frac{1}{2} - \frac{11c}{2} ,$$

where in the last inequality we used that $n \geq 1$. Taking c to be a small enough constant (i.e. $c < \frac{1}{22}$), we get that $\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j > \frac{1}{4}$, which is a contradiction. □

We are now ready to prove the main theorem of this section:

Proof of Theorem 4.3. By Lemma B.9 we have that $\sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j > \frac{1}{4}$. Denote by $\lambda_s := \arg\max_{i \in [m]} \lambda_i$, we would first like to show that $\lambda_s \leq 3mn$. We split into two cases:

Case I. Assume $s \in I_+$. Note that $\mathbf{x}_s$ lies on the margin, otherwise $\lambda_i = 0$ for every $i \in I$, which means that $N_{\boldsymbol{\theta}}$ is the zero predictor, this contradicts the assumption that $N_{\boldsymbol{\theta}}$ classifies the data correctly. Using a similar analysis to Case I in Lemma B.7 we have:

$$1 = N_{\boldsymbol{\theta}}(\mathbf{x}_s) = \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_s + b_j)$$
$$= \sum_{j=1}^{n/2} \sigma\Big(\sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j\Big) - \sum_{j=n/2+1}^{n} \sigma\Big(-\sum_{i \in [m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j\Big) + \sum_{j=1}^{n/2} \sigma(b_j) - \sum_{j=1}^{n/2} \sigma(b_j)$$
$$\geq \lambda_s + \sum_{j=1}^{n/2} \sigma\Big(-\sum_{i \in I \setminus \{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j\Big) - \sum_{j=n/2+1}^{n} \sigma\Big(\sum_{i \in I \setminus \{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j\Big) + \sum_{j=1}^{n/2} \sigma(b_j) - \sum_{j=1}^{n/2} \sigma(b_j)$$
$$\geq \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j + \lambda_s - \lambda_s mn \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}$$
$$\geq \frac{1}{4} + \lambda_s \left(1 - mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}}\right) ,$$

where we used Lemma B.6 to show that there is at least one $k \in \{1, \ldots, n/2\}$ with $\sigma'_{s,k} = 1$. Using that $mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \leq \frac{1}{2}$ for a large enough constant c_3 and rearranging the terms above we have that $\lambda_s \leq \frac{6}{4} \leq 3mn$ since $m, n \geq 1$.

Case II. Assume $s \in I_-$. Take $\lambda_r := \arg\max_{i \in I_+} \lambda_i$. By Lemma B.5 there is at least one neuron $k \in \{n/2+1, \ldots, n\}$ with $\sigma'_{s,k} = 1$. We have that:

$$\begin{aligned}\frac{1}{4} &\leq \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j \\ &\leq \sum_{j=1}^{n/2} \sum_{i\in[m]} \lambda_i y_i \sigma'_{i,j} - \sum_{j=n/2+1}^{n} \sum_{i\in[m]} \lambda_i y_i \sigma'_{i,j} \leq mn\lambda_r - \lambda_s \,.\end{aligned}$$

Rearranging the terms we get $\lambda_s \leq mn\lambda_r - \frac{1}{4} \leq mn\lambda_r$. Note that $\mathbf{x}_r$ lies on the margin, otherwise $\lambda_s = 0$, which similarly to Case I is a contradiction. Again, using a similar analysis to Case I we get:

$$\begin{aligned}1 = N_{\boldsymbol{\theta}}(\mathbf{x}_r) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) \\ &= \sum_{j=1}^{n/2} \sigma\Big(\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j\Big) - \sum_{j=n/2+1}^{n} \sigma\Big(-\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + b_j\Big) + \sum_{j=1}^{n/2} \sigma(b_j) - \sum_{j=1}^{n/2} \sigma(b_j) \\ &\geq \lambda_r + \sum_{j=1}^{n/2} \sigma\Big(-\sum_{i\in I\setminus\{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j\Big) - \sum_{j=n/2+1}^{n} \sigma\Big(\sum_{i\in I\setminus\{s\}} \lambda_i |\mathbf{x}_i^\top \mathbf{x}_s| + b_j\Big) + \sum_{j=1}^{n/2} \sigma(b_j) - \sum_{j=1}^{n/2} \sigma(b_j) \\ &\geq \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j + \lambda_r - \lambda_s mn\sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \\ &\geq \frac{1}{4} + \lambda_r - \lambda_r m^2 n^2 \sqrt{\frac{2\log\left(\frac{2m^2}{\delta}\right)}{d}} \,,\end{aligned}$$

By our assumption, $\frac{m^2 n^2 \log(d) 2\sqrt{2}}{\sqrt{d}} \leq \frac{1}{2}$ for a large enough constant c_3. Hence, rearranging the terms we get $\lambda_r \leq \frac{6}{4}$. This means that $\lambda_s \leq 2mn\lambda_r \leq 3mn$.

We now turn to calculate the output of $N_{\boldsymbol{\theta}}$ on the sample $\mathbf{x}$. Suppose we sample $\mathbf{x} \sim \mathrm{Unif}(\mathbb{S}_{d-1})$, by Lemma D.1 we have with probability $> 1 - \epsilon$ that $|\mathbf{x}_i^\top \mathbf{x}| \leq \sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}}$ for every $i \in [m]$. We condition on this event for the rest of the proof. Note that for every positive neuron j we have that:

$$\begin{aligned}\sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) - \sigma(b_j) &= \sigma\left(\sum_{i\in[m]} y_i \lambda_i \sigma'_{i,j} \mathbf{x}_i^\top \mathbf{x} + b_j\right) - \sigma(b_j) \\ &\geq \sigma\left(-\lambda_s m\sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} + b_j\right) - \sigma(b_j) \\ &\geq -\lambda_s m\sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \,.\end{aligned}$$

Similarly, for every negative neuron j we have:

$$\sigma(b_j) - \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) \geq -\lambda_s m\sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \,.$$

In both bounds above we used that λ_s is the largest among the λ_i's. Combining both bounds, we have that:

$$\begin{aligned}
N_{\boldsymbol{\theta}}(\mathbf{x}) &= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) \\
&= \sum_{j=1}^{n/2} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) - \sum_{j=n/2+1}^{n} \sigma(\mathbf{w}_j^\top \mathbf{x}_r + b_j) + \\
&+ \sum_{j=1}^{n/2} \sigma(b_j) - \sum_{j=1}^{n/2} \sigma(b_j) + \sum_{j=n/2+1}^{n} \sigma(b_j) - \sum_{j=n/2+1}^{n} \sigma(b_j) \\
&\geq \sum_{j=1}^{n/2} b_j - \sum_{j=n/2+1}^{n} b_j - \lambda_s mn \sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \\
&\geq \frac{1}{4} - 3m^2n^2\sqrt{\frac{2\log\left(\frac{m}{\epsilon}\right)}{d}} \,.
\end{aligned}$$

We use that $\frac{3m^2n^2\log(d)\sqrt{2}}{\sqrt{d}} \leq \frac{1}{8}$ for a large enough constant c_3, hence $N_{\boldsymbol{\theta}}(\mathbf{x}) \geq \frac{1}{4} - \frac{1}{8} > 0$, this finishes the proof. □

C Proofs from Section 4.2

C.1 Proof of Proposition 4.4.

We first note that there exists $j \in \{1, 2\}$ with $v_j < 0$, since otherwise the network wouldn't be able to classify samples with a negative label (which exist by assumption). Assuming without loss of generality that $v_2 < 0$, for any $\mathbf{w}_1$ we also have that $\Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\mathbf{w}_1^\top \mathbf{x} \leq 0] = \frac{1}{2}$. If this even occurs, then $N_{\boldsymbol{\theta}}(\mathbf{x}) \leq v_2\sigma(\mathbf{w}_2 \top \mathbf{x}) \leq 0$.

C.2 Proof of Proposition 4.5

Suppose $[n] = J_+ \cup J_-$ are such that

$$N_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j\in J_+} v_j\sigma(\mathbf{w}_j^\top \mathbf{x}) + \sum_{j\in J_-} v_j\sigma(\mathbf{w}_j^\top \mathbf{x}) \,,$$

where $v_j \geq 0$ for $j \in J_+$ and $v_j < 0$ for $j \in J_-$. Then, for any choice of $(\mathbf{w}_j)_{j\in J_+}$:

$$\Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[N_{\boldsymbol{\theta}}(\mathbf{x}) \leq 0] \geq \Pr_{\mathbf{x}\sim\mathbb{S}^{d-1}}[\forall j \in J_+,\ \mathbf{w}_j^\top \mathbf{x} < 0] \geq \frac{1}{2^{|J_+|}} \geq \frac{1}{2^n} \,.$$

C.3 Proof of Proposition 4.6

Assume without loss of generality that $I_- = [k]$. Consider the weights $\mathbf{w}_i = \mathbf{x}_i,\ v_i = -1$ for every $i \in I_-$, and $\mathbf{w}_{k+1} = \sum_{i\in I_+} \mathbf{w}_i,\ v_{k+1} = 1$. For this network we have $y_i N_{\boldsymbol{\theta}}(\mathbf{x}_i) = 1$, thus all points lie on the margin. In addition, it is not difficult to see that this network satisfies the other KKT conditions with $\lambda_i = 1$ for every $i \in [m]$. Note that $\Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\mathbf{w}_{k+1}^\top \mathbf{x} < 0] = \frac{1}{2}$ and $\Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\forall i \in I_-,\ \mathbf{w}_i^\top \mathbf{x} < 0] = \frac{1}{2^k}$, since all the $\mathbf{x}_i$ are orthogonal. Thus, we get that

$$\Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[N_{\boldsymbol{\theta}}(\mathbf{x}) \leq 0] \geq \Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\mathbf{w}_{k+1}^\top \mathbf{x} < 0] - \Pr_{\mathbf{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\forall i \in I_-,\ \mathbf{w}_i^\top \mathbf{x} < 0] \geq \frac{1}{2} - \frac{1}{2^k} \,.$$

D Additional probabilistic lemmas

D.1 Proof of Lemma A.1

Clearly, the third (no collisions) condition holds almost surely, so it suffices to analyze the gaps between samples. The distribution of distances between uniformly random points on a segment is well studied [Pyke, 1965, Holst, 1980] and the lemma can be derived by known results. Nonetheless we provide a proof for completeness.

Denote by $\Delta_1 \leq \Delta_2 \leq \cdots \leq \Delta_{m+1}$ the *ordered* spacings $x_1, (x_2 - x_1), \ldots, (x_m - x_{m-1}), (1 - x_m)$. With this notation, note that E_x occurs if $\Delta_{m+1} \leq \frac{\log(8(m+1)/\delta)}{m+1}$ and $\Delta_{m/8} \geq \frac{1}{10m}$. We will show that each of these conditions holds with probability at least $1 - \frac{\delta}{4}$, under which we would conclude by the union bound. Let $Z_1, \ldots, Z_{m+1} \overset{iid}{\sim} \mathrm{Exp}(1)$ be unit mean exponential random variables, and denote their ordering by $Z_{(1)} \leq \cdots \leq Z_{(m+1)}$. The main well known observation which we use is that for any $j \in [m+1]$: $\Delta_j \overset{d}{=} \frac{Z_{(j)}}{\sum_{i=1}^{m+1} Z_i}$ (see Holst, 1980). Hence for any $t > 0$:

$$\begin{aligned}
&\Pr[(m+1)\Delta_j - \log(m+1) \leq t] \\
&= \Pr\left[Z_{(j)} - \log(m+1) \leq t + (t + \log(m+1)) \left(\frac{\sum_{i=1}^{m+1} Z_i}{m+1} - 1 \right) \right] .
\end{aligned}$$

Note that $\frac{\sum_{i=1}^{m+1} Z_i}{m+1} - 1 \xrightarrow{p} 0$ as $\mathbb{E}\left[\frac{\sum_{i=1}^{m+1} Z_i}{m+1}\right] = 1$ and $\mathrm{Var}\left[\frac{\sum_{i=1}^{m+1} Z_i}{m+1}\right] = \frac{m+1}{(m+1)^2} \overset{m\to\infty}{\longrightarrow} 0$. Furthermore,

$$\begin{aligned}
\Pr\left[Z_{(j)} - \log(m+1) \leq t\right] &= \Pr\left[Z_{(j)} \leq t + \log(m+1)\right] = \Pr[Z_{(1)}, \ldots, Z_{(j)} \leq t + \log(m+1)] \\
&= \left(1 - e^{-t - \log(m+1)}\right)^j = \left(1 - \frac{e^{-t}}{m+1}\right)^j .
\end{aligned}$$

By introducing the change of variables $r = \frac{t + \log(m+1)}{m+1}$ we conclude that

$$\lim_{m\to\infty} \Pr[\Delta_j \leq r] = \left(1 - \frac{e^{-(m+1)r + \log(m+1)}}{m+1}\right)^j .$$

It remains to plug in our parameters of interest. For $j = m+1$ we get

$$\lim_{m\to\infty} \Pr[\Delta_{m+1} \leq r] = \lim_{m\to\infty} \left(1 - \frac{e^{-(m+1)r + \log(m+1)}}{m+1}\right)^{m+1} = \lim_{m\to\infty} \exp\left(-e^{-(m+1)r + \log(m+1)}\right) .$$

Noting that for $r = \frac{8\log(8(m+1)/\delta)}{m+1}$ it holds that $e^{-(m+1)r + \log(m+1)} \to 0$ so we can use the Taylor approximation $\exp(z) \approx 1 + z$ and conclude that

$$\lim_{m\to\infty} \Pr[\Delta_{m+1} \leq r] = 1 - e^{-(m+1)r + \log(m+1)} = 1 - \frac{\delta}{8} ,$$

where the last equality holds for $r = \frac{\log(8(m+1)/\delta)}{m+1}$. Overall for m larger than some numerical constant, the left hand side is larger than $1 - \delta/4$ as required.

The second condition follows a similar computation for $j = \frac{m+1}{8}$, $r = \frac{\log(m/8\log(8/\delta))}{8(m+1)}$, while using the fact $\Pr\left[\Delta_{m/8} > \frac{1}{10(m+1)}\right] = 1 - \Pr\left[\Delta_{m/8} \leq \frac{1}{10(m+1)}\right]$.

D.2 Concentration bounds for vectors on the unit sphere

Here we present some standard concentration bounds for uniformly sampled points on the unit sphere in high dimension.

Lemma D.1 (Lemma 2.2 from Ball, 1997)**.** $\Pr_{\mathbf{u},\mathbf{v}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[|\mathbf{u}^\top \mathbf{v}| \geq t] \leq 2\exp\left(\frac{-dt^2}{2}\right)$.

Lemma D.2 (Exercise 4.7.3 from Vershynin, 2018)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and denote by X the matrix whose i'th row is $\mathbf{x}_i$. Then, there is a universal constant $c > 0$ such that:*

$$\Pr\left[\left\|XX^\top - I\right\| \geq \frac{c}{d} \cdot \left(\sqrt{\frac{d+t}{m}} + \frac{d+t}{m}\right)\right] \leq 2e^{-t} .$$