Towards Understanding Neural Networks with Linear Teachers
Supplementary Material

Here we provide proofs of the theorems stated in the paper, and additional empirical results.

1. Gradient Flow Definitions

We next formally define gradient flow. A function $f : \mathbb{X} \to \mathbb{R}$ is locally Lipschitz if for every $\boldsymbol{x} \in \mathbb{X}$ there exists a neighborhood $\mathbb{U}$ of $\boldsymbol{x}$ such that the restriction of f on $\mathbb{U}$ is Lipschitz continuous. For a locally Lipschitz function $f : \mathbb{X} \to \mathbb{R}$, the Clarke subdifferential at $\boldsymbol{x} \in \mathbb{X}$ is the convex set:

$$\partial^\circ f(\boldsymbol{x}) \coloneqq \text{conv}\left\{\lim_{k\to\infty} \nabla f(\boldsymbol{x}_k) : \boldsymbol{x}_k \to \boldsymbol{x}, f \text{ is differentiable at } \boldsymbol{x}_k\right\} \tag{1}$$

As in (Lyu & Li, 2020) and (Ji & Telgarsky, 2020), a curve z from an interval I to a real space $\mathbb{R}^m$ is called an arc if it is absolutely continuous on any compact subinterval of I. For an arc z we use $z'(t)$ (or $\frac{dz}{dt}(t)$) to denote the derivative at t if it exists. We say that a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ admits a chain rule if for any arc $z : [0; +\infty) \to \mathbb{R}^d, \forall h \in \partial^\circ f(z(t)) : (f \circ z)'(t) = \langle h, z'(t)\rangle$ holds for a.e. $t \geq 0$. It holds that an arc is a.e. differentiable, and the composition of an arc and a locally Lipschitz function is still an arc.

Given the definitions above, we define gradient flow $\boldsymbol{W} : [0, \infty) \to \mathbb{R}^k$ to be an arc that satisfies the following differential inclusion for a.e. $t \geq 0$:

$$\frac{d\boldsymbol{W}_t}{dt} \in -\partial^\circ L_{\mathbb{S}}(\boldsymbol{W}_t) \tag{2}$$

2. Proof of Theorem 4.1

Throughout this proof we will sometimes use the notation $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ as the dot product between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ for readability purposes.

Let $\overrightarrow{\boldsymbol{W}}^* = (\overbrace{\boldsymbol{w}^* \dots \boldsymbol{w}^*}^{k}, \overbrace{-\boldsymbol{w}^* \cdots - \boldsymbol{w}^*}^{k}) \in \mathbb{R}^{2kd}$.

Define the following two functions:

$$F(\boldsymbol{W}_t) = \langle \overrightarrow{\boldsymbol{W}}_t, \overrightarrow{\boldsymbol{W}}^* \rangle = \sum_{i=1}^{k} \langle \boldsymbol{w}_t^{(i)}, \boldsymbol{w}^* \rangle - \sum_{i=1}^{k} \langle \boldsymbol{u}_t^{(i)}, \boldsymbol{w}^* \rangle$$

and

$$G(\boldsymbol{W}_t) = ||\overrightarrow{\boldsymbol{W}}_t|| = \sqrt{\sum_{i=1}^{k} ||\boldsymbol{w}_t^{(i)}||^2 + \sum_{i=1}^{k} ||\boldsymbol{u}_t^{(i)}||^2}$$

Then, from Cauchy-Schwartz inequality we have:

$$\frac{|F(\boldsymbol{W}_t)|}{G(\boldsymbol{W}_t)||\overrightarrow{\boldsymbol{W}}^*||} = \frac{|\langle \overrightarrow{\boldsymbol{W}}_t, \overrightarrow{\boldsymbol{W}}^* \rangle|}{||\overrightarrow{\boldsymbol{W}}_t||||\overrightarrow{\boldsymbol{W}}^*||} \leq 1 \tag{3}$$

Recall we define: $N_{\boldsymbol{W}}(\boldsymbol{x}) = v \sum\limits_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - v \sum\limits_{j=1}^{k} \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})$.

We consider minimizing the objective function:

$$L_{\mathbb{S}}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)}\right)$$

using SGD on $\mathbb{S}$ where each point is sampled without replacement at each epoch. WLOG, we set $\sigma'(0) = \alpha$.

We first outline the proof structure. Let's assume we run SGD for N_e epochs and denote $T = nN_e$. Furthermore, we assume that for all epochs up to this point there is at least one point in the epoch s.t. $\ell(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t)) > \varepsilon_0$ for some $\varepsilon_0 > 0$ (recall that n is the number of training points, and $(y_t, \boldsymbol{x}_t)$ is some training point selected during some epoch).

First, we will show that after at most $T \leq M(n, \epsilon_0)$ iterations, there exists an epoch i_e such that for each point $(\boldsymbol{x}_t, y_t) \in \mathbb{S}$ sampled in the epoch, it holds that:

$$\ell(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t)) \leq \varepsilon_0 \tag{4}$$

Next, using the Lipschitzness of $\ell(x)$ we will show that the loss on points cannot change too much during an epoch. Specifically, we will use this to show that at the end of epoch i_e, which we denote by time T^*, it holds for all $(\boldsymbol{x}_i, y_i) \in \mathbb{S}$:

$$\ell(y_i N_{\boldsymbol{W}_{T^*}}(\boldsymbol{x}_i)) \leq (1 + 2v^2 R_x^2 \eta k n)\varepsilon_0 \tag{5}$$

now by choosing $\varepsilon_0 = \frac{\varepsilon}{1+2v^2R_x^2\eta kn}$ we will get that $\forall 1 \leq i \leq n$ $\ell(y_i N_{\boldsymbol{W}_{T^*}}(\boldsymbol{x}_i)) \leq \varepsilon$ which shows that $L_{\mathbb{S}}(\boldsymbol{W}_{T^*}) \leq \varepsilon$ as required.

We start by showing Eq. (4).

For the gradient of each neuron we have:

$$\begin{aligned}
\frac{\partial L_{\{(\boldsymbol{x}_i, y_i)\}}(\boldsymbol{W})}{\partial \boldsymbol{w}^{(j)}} &= \frac{e^{-y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)}}{1 + e^{-y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)}} \cdot \frac{-y_i \partial N_{\boldsymbol{W}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}^{(j)}} \\
&= \frac{-y_i e^{-y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)}}{1 + e^{-y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)}} \cdot v \boldsymbol{x}_i \sigma'(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_i) \\
&= -v y_i \boldsymbol{x}_i \left|\ell'(y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i))\right| \sigma'(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_i)
\end{aligned}$$

and similarly:

$$\frac{\partial L_{\{(\boldsymbol{x}_i, y_i)\}}(\boldsymbol{W})}{\partial \boldsymbol{u}^{(j)}} = v y_i \boldsymbol{x}_i \left|\ell'(y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i))\right| \sigma'(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_i)$$

where $\ell'(x) = -\frac{e^{-x}}{1+e^{-x}} = -\frac{1}{1+e^{x}}$ and $\ell(x) = log(1 + e^{-x})$.

Optimizing by SGD yields the following update rule:

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta \frac{\partial}{\partial \boldsymbol{W}} L_{\{(\boldsymbol{x}_t, y_t)\}}(\boldsymbol{W}_{t-1})$$

where $\boldsymbol{W}_t = (\boldsymbol{w}_t^{(1)}, ..., \boldsymbol{w}_t^{(k)}, \boldsymbol{u}_t^{(1)}, ..., \boldsymbol{u}_t^{(k)})$.

For every neuron we get the following updates:

1. $\boldsymbol{w}_t^{(j)} = \boldsymbol{w}_{t-1}^{(j)} + \eta v y_t \boldsymbol{x}_t \left|\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))\right| p_{t-1}^{(j)}$
2. $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_{t-1}^{(j)} - \eta v y_t \boldsymbol{x}_t \left|\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))\right| q_{t-1}^{(j)}$

where $p_t^{(j)} := \sigma'(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{x}_{t+1})$; $q_t^{(j)} := \sigma'(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_{t+1})$.

Next we will show recursive upper bounds for $G(\boldsymbol{W}_t)$ and $F(\boldsymbol{W}_t)$.

$$
\begin{aligned}
G(\boldsymbol{W}_t)^2 &= \sum_{j=1}^{k} ||\boldsymbol{w}_t^{(j)}||^2 + \sum_{j=1}^{k} ||\boldsymbol{u}_t^{(j)}||^2 \\
&\leq \sum_{j=1}^{k} ||\boldsymbol{w}_{t-1}^{(j)}||^2 + \sum_{j=1}^{k} ||\boldsymbol{u}_{t-1}^{(j)}||^2 \\
&+ 2\eta y_t |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| \left(\sum_{j=1}^{k} \langle \boldsymbol{w}_{t-1}^{(j)}, \boldsymbol{x}_t \rangle p_{t-1}^{(j)} v - \sum_{j=1}^{k} \langle \boldsymbol{u}_{t-1}^{(j)}, \boldsymbol{x}_t \rangle q_{t-1}^{(j)} v \right) \\
&+ 2k\eta^2 v^2 ||\boldsymbol{x}_t||^2 |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))|^2 = \sum_{j=1}^{k} ||\boldsymbol{w}_{t-1}^{(j)}||^2 + \sum_{j=1}^{k} ||\boldsymbol{u}_{t-1}^{(j)}||^2 \\
&+ 2\eta |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t) + 2k\eta^2 v^2 ||\boldsymbol{x}_t||^2 |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))|^2 \\
&= G(\boldsymbol{W}_{t-1})^2 + 2\eta |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t) + 2k\eta^2 v^2 ||\boldsymbol{x}_t||^2 |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))|^2
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
F(\boldsymbol{W}_t) &= \sum_{j=1}^{k} \langle \boldsymbol{w}_t^{(j)}, \boldsymbol{w}^* \rangle - \sum_{j=1}^{k} \langle \boldsymbol{u}_t^{(j)}, \boldsymbol{w}^* \rangle = \sum_{j=1}^{k} \langle \boldsymbol{w}_{t-1}^{(j)}, \boldsymbol{w}^* \rangle - \sum_{j=1}^{k} \langle \boldsymbol{u}_{t-1}^{(j)}, \boldsymbol{w}^* \rangle \\
&+ \eta |\ell'(y_t N_{W_{t-1}}(\boldsymbol{x}_t))| \sum_{j=1}^{k} \langle y_t \boldsymbol{x}_t, \boldsymbol{w}^* \rangle p_{t-1}^{(j)} v + \eta |\ell'(y_t N_{W_{t-1}}(\boldsymbol{x}_t))| \sum_{j=1}^{k} \langle y_t \boldsymbol{x}_t, \boldsymbol{w}^* \rangle q_{t-1}^{(j)} v \\
&\geq \sum_{j=1}^{k} \langle \boldsymbol{w}_{t-1}^{(j)}, \boldsymbol{w}^* \rangle - \sum_{j=1}^{k} \langle \boldsymbol{u}_{t-1}^{(j)}, \boldsymbol{w}^* \rangle + 2k\eta v \alpha |\ell'(y_t N_{W_{t-1}}(\boldsymbol{x}_t))|
\end{aligned}
$$

Where we used the inequalities $\langle y_t \boldsymbol{x}_t, \boldsymbol{w}^* \rangle \geq 1$ and $q_t^{(j)}, p_t^{(j)} \geq \alpha$.

To summarize we have:

$$
G(\boldsymbol{W}_t)^2 \leq G(\boldsymbol{W}_{t-1})^2 + 2\eta |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t) + 2k\eta^2 v^2 R_x^2 |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))|^2 \tag{6}
$$

$$
F(\boldsymbol{W}_t) \geq F(\boldsymbol{W}_{t-1}) + 2k\eta v \alpha |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| \tag{7}
$$

For an upper bound on $G(\boldsymbol{W}_t)$ we use the following inequalities (which hold for the cross entropy loss):

$\forall x \in \mathbb{R} \quad \frac{x}{1+e^x} \leq 1 \Rightarrow |\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t) = \frac{y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t)}{1+e^{y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t)}} \leq 1$ and $|\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| \leq 1$. Together we have for any t:

$$
G(\boldsymbol{W}_t)^2 \leq G(\boldsymbol{W}_{t-1})^2 + 2\eta + 2k\eta^2 v^2 R_x^2
$$

Using this recursively up until $T = nN_e$ we get:

$$
G(\boldsymbol{W}_T)^2 \leq G(\boldsymbol{W}_0)^2 + T(2k\eta^2 v^2 R_x^2 + 2\eta) \tag{8}
$$

Now, for $F(\boldsymbol{W}_t)$, let $\varepsilon_0 > 0$, under our assumption, in any epoch i_e until N_e $(1 \leq i_e \leq N_e)$ there exists at least one point in the epoch $(y_{t_{i_e}}, \boldsymbol{x}_{t_{i_e}}) \in \mathbb{S}$ s.t. $\ell(y_{t_{i_e}} N_{\boldsymbol{W}_{t_{i_e}}}(\boldsymbol{x}_{t_{i_e}})) > \varepsilon_0$.

Now, since in our case $\ell(x) = log(1 + e^{-x})$ and $\ell'(x) = -\frac{1}{1+e^x}$, we see that the condition $\ell(x) > \varepsilon_0$ implies that:

$$
|\ell'(x)| > 1 - e^{-\varepsilon_0} \tag{9}
$$

In any other case $|\ell'(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t))| \geq 0$, so if we assume at least one point violation per epoch (i.e. $\ell(y_{t_{i_e}} N_{\boldsymbol{W}_{t_{i_e}}}(\boldsymbol{x}_{t_{i_e}})) \geq \varepsilon_0$ for some point $\left(y_{t_{i_e}}, \boldsymbol{x}_{t_{i_e}}\right)$ in the epoch) we would get that at the end of epoch N_e:

$$F(\boldsymbol{W}_T) \geq F(\boldsymbol{W}_{T-n}) + 2k\eta v\alpha(1 - e^{-\varepsilon_0}) \tag{10}$$

This implies that (recursively using Eq. (10)):

$$F(\boldsymbol{W}_T) \geq F(\boldsymbol{W}_0) + 2k\eta v\alpha N_e(1 - e^{-\varepsilon_0}) \tag{11}$$

where N_e is the number of epochs and n the number of training points, $T = nN_e$.

Now, using the Cauchy-Schwartz, Eq. (8) and Eq. (11) we have:

$$\begin{aligned}
&-G(\boldsymbol{W}_0)||\overrightarrow{\boldsymbol{W}}^*|| + 2k\eta v\alpha N_e(1 - e^{-\varepsilon_0}) \leq F(\boldsymbol{W}_0) + 2k\eta v\alpha N_e(1 - e^{-\varepsilon_0}) \\
&\leq F(\boldsymbol{W}_T) \leq ||\overrightarrow{\boldsymbol{W}}^*||G(\boldsymbol{W}_T) \leq ||\overrightarrow{\boldsymbol{W}}^*||\sqrt{G(\boldsymbol{W}_0)^2 + T(2k\eta^2 v^2 R_x^2 + 2\eta)}
\end{aligned}$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ the above implies:

$$-G(\boldsymbol{W}_0)||\overrightarrow{\boldsymbol{W}}^*|| + 2k\eta v\alpha N_e(1 - e^{-\varepsilon_0}) \leq ||\overrightarrow{\boldsymbol{W}}^*||G(\boldsymbol{W}_0) + ||\overrightarrow{\boldsymbol{W}}^*||\sqrt{T}\sqrt{2k\eta^2 v^2 R_x^2 + 2\eta}$$

Now using $\left|\left|\boldsymbol{w}_0^{(i)}\right|\right|, \left|\left|\boldsymbol{u}_0^{(i)}\right|\right| \leq R_0$ we get $G(\boldsymbol{W}_0) \leq \sqrt{2k}R_0$.

Noting that $\left\|\overrightarrow{\boldsymbol{W}}^*\right\| = \sqrt{2k}||\boldsymbol{w}^*||$ and that $N_e = \frac{T}{n}$, we get :

$$\left(\frac{2k\eta v\alpha(1 - e^{-\varepsilon_0})}{n}\right)T \leq \sqrt{4k^2\eta^2 v^2 R_x^2 + 4k\eta}||\boldsymbol{w}^*||\sqrt{T} + 4kR_0||\boldsymbol{w}^*||$$

Therefore, we have an inequality of the form:

$$aT \leq b\sqrt{T} + c$$

where $a = \dfrac{2k\eta v\alpha(1 - e^{-\varepsilon_0})}{n}, b = \sqrt{4k^2\eta^2 v^2 R_x^2 + 4k\eta}||\boldsymbol{w}^*||$ and $c = 4kR_0||\boldsymbol{w}^*||$.

By inspecting the roots of the parabola $P(X) = x^2 - \frac{b}{a}x - \frac{c}{a}$ we conclude that:

$$\begin{aligned}
T \leq \left(\frac{b}{a}\right)^2 + \sqrt{\frac{c}{a}}\frac{b}{a} + \frac{c}{a} &= \frac{(4k^2\eta^2 v^2 R_x^2 + 4k\eta)||\boldsymbol{w}^*||^2 n^2}{4k^2\eta^2 v^2\alpha^2(1 - e^{-\varepsilon_0})^2} + \frac{\sqrt{4k^2\eta^2 v^2 R_x^2 + 4k\eta}||\boldsymbol{w}^*||n}{2k\eta v\alpha(1 - e^{-\varepsilon_0})}\sqrt{\frac{4kR_0||\boldsymbol{w}^*||n}{2k\eta v\alpha(1 - e^{-\varepsilon_0})}} \\
&+ \frac{4kR_0||\boldsymbol{w}^*||n}{2k\eta v\alpha(1 - e^{-\varepsilon_0})} = \left(\frac{R_x^2}{\alpha^2} + \frac{1}{k\eta v^2\alpha^2}\right)\frac{||\boldsymbol{w}^*||^2 n^2}{(1 - e^{-\varepsilon_0})^2} + \frac{\sqrt{R_0(8k^2\eta^2 v^2 R_x^2 + 8k\eta)}||\boldsymbol{w}^*||^{1.5} n^{1.5}}{2k(\eta v\alpha)^{1.5}(1 - e^{-\varepsilon_0})^{1.5}} \\
&+ \frac{2R_0||\boldsymbol{w}^*||n}{\eta v\alpha(1 - e^{-\varepsilon_0})}
\end{aligned}$$

By the inequality $1 - e^{-x} > \frac{x}{1+x}$ for $x > 0$ (which is equivalent to $\frac{1}{1-e^{-x}} < \frac{x+1}{x}$), with $x = \varepsilon_0 > 0$ we get $\frac{1}{1-e^{-\varepsilon_0}} < \frac{\varepsilon_0+1}{\varepsilon_0} = 1 + \frac{1}{\varepsilon_0}$. Therefore for $\beta > 0$ (all arguments are positive):

$$\frac{1}{(1 - e^{-\varepsilon_0})^\beta} < \left(1 + \frac{1}{\varepsilon_0}\right)^\beta$$

By using the above inequality we can reach a polynomial bound on T:

$$\begin{aligned}
T &\leq \left(\frac{R_x^2}{\alpha^2} + \frac{1}{k\eta v^2\alpha^2}\right)||\boldsymbol{w}^*||^2 n^2\left(1 + \frac{1}{\varepsilon_0}\right)^2 \\
&+ \frac{\sqrt{R_0(8k^2\eta^2 v^2 R_x^2 + 8k\eta)}||\boldsymbol{w}^*||^{1.5} n^{1.5}(1 + \frac{1}{\varepsilon_0})^{1.5}}{2k(\eta v\alpha)^{1.5}} + \frac{2R_0||\boldsymbol{w}^*||n(1 + \frac{1}{\varepsilon_0})}{\eta v\alpha}
\end{aligned} \tag{12}$$

We have shown that there is at most a finite amount of epochs $N_e = \frac{T}{n}$ such that there exists at least one point in each of them with a loss greater than ε_0. Therefore, there exists an epoch $1 \leq i_e \leq N_e + 1$ such that each point sampled in the epoch has a loss smaller than ε_0. Formally, for any $(i_e - 1)n + 1 \leq t \leq i_e n, \;\; \ell\left(y_t N_{\boldsymbol{W}_{t-1}}(\boldsymbol{x}_t)\right) \leq \varepsilon_0$. Recall that SGD samples without replacement and therefore, each point is sampled at some t in the epoch i_e.

Next, we will show that there exists a time t such that $L_{\mathbb{S}}(\boldsymbol{W}_t) < \varepsilon$ by bounding the change in the loss values during the epoch. We'll start by noticing that our loss function $\ell(x)$ is locally Lipschitz with coefficient 1, that is because $\forall \boldsymbol{x} \; |\ell'(\boldsymbol{x})| = \frac{1}{1+e^x} \leq 1$. With this in mind for any point $(y_i, \boldsymbol{x}_i) \in \mathbb{S}$ if we can bound $|y_i N_{\boldsymbol{W}_{t+s}}(\boldsymbol{x}_i) - y_i N_{\boldsymbol{W}_t}(\boldsymbol{x}_i)|$ we would also bound $|\ell\left(y_i N_{\boldsymbol{W}_{t+s}}(\boldsymbol{x}_i)\right) - \ell\left(y_i N_{\boldsymbol{W}_t}(\boldsymbol{x}_i)\right)|$.

For any iteration $(i_e - 1)n + 1 \leq t \leq i_e n$ and $1 \leq s \leq n$ we have:

$$
\begin{aligned}
&|y_i N_{\boldsymbol{W}_{t+s}}(\boldsymbol{x}_i) - y_i N_{\boldsymbol{W}_t}(\boldsymbol{x}_i)| = |N_{\boldsymbol{W}_{t+s}}(\boldsymbol{x}_i) - N_{\boldsymbol{W}_t}(\boldsymbol{x}_i)| \\
&= \left| v \sum_{j=1}^{k} \left(\sigma(\boldsymbol{w}_{t+s}^{(j)} \cdot \boldsymbol{x}_i) - \sigma(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{x}_i)\right) - v \sum_{j=1}^{k} \left(\sigma(\boldsymbol{u}_{t+s}^{(j)} \cdot \boldsymbol{x}_i) - \sigma(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_i)\right) \right| \\
&\leq v \sum_{j=1}^{k} \left|\sigma(\boldsymbol{w}_{t+s}^{(j)} \cdot \boldsymbol{x}_i) - \sigma(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{x}_i)\right| + v \sum_{j=1}^{k} \left|\sigma(\boldsymbol{u}_{t+s}^{(j)} \cdot \boldsymbol{x}_i) - \sigma(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_i)\right| \\
&\leq v \sum_{j=1}^{k} \left|\left(\boldsymbol{w}_{t+s}^{(j)} - \boldsymbol{w}_t^{(j)}\right) \cdot \boldsymbol{x}_i\right| + v \sum_{j=1}^{k} \left|\left(\boldsymbol{u}_{t+s}^{(j)} - \boldsymbol{u}_t^{(j)}\right) \cdot \boldsymbol{x}_i\right| && (13) \\
&\leq v \sum_{j=1}^{k} ||\boldsymbol{w}_{t+s}^{(j)} - \boldsymbol{w}_t^{(j)}|| \cdot ||\boldsymbol{x}_i|| + v \sum_{j=1}^{k} ||\boldsymbol{u}_{t+s}^{(j)} - \boldsymbol{u}_t^{(j)}|| \cdot ||\boldsymbol{x}_i|| && (14) \\
&\leq v R_x \sum_{j=1}^{k} \left|\left| \sum_{h=1}^{s} \eta v y_{t+h} \boldsymbol{x}_{t+h} \left|\ell'(y_{t+h} N_{\boldsymbol{W}_{t+h-1}}(\boldsymbol{x}_{t+h}))\right| p_{t+h-1}^{(j)} \right|\right| \\
&+ v R_x \sum_{j=1}^{k} \left|\left| \sum_{h=1}^{s} \eta v y_{t+h} \boldsymbol{x}_{t+h} \left|\ell'(y_{t+h} N_{\boldsymbol{W}_{t+h-1}}(\boldsymbol{x}_{t+h}))\right| q_{t+h-1}^{(j)} \right|\right| && (15) \\
&\leq v R_x \sum_{j=1}^{k} \sum_{h=1}^{s} \eta v \left|\ell'(y_{t+h} N_{\boldsymbol{W}_{t+h-1}}(\boldsymbol{x}_{t+h}))\right| ||\boldsymbol{x}_{t+h}|| + v R_x \sum_{j=1}^{k} \sum_{h=1}^{s} \eta v \left|\ell'(y_{t+h} N_{\boldsymbol{W}_{t+h-1}}(\boldsymbol{x}_{t+h}))\right| ||\boldsymbol{x}_{t+h}|| \\
&\leq 2v^2 R_x^2 \eta k \sum_{h=1}^{s} \left|\ell'(y_{t+h} N_{\boldsymbol{W}_{t+h-1}}(\boldsymbol{x}_{t+h}))\right| \leq 2v^2 R_x^2 \eta k s (1 - e^{-\varepsilon_0}) \leq 2v^2 R_x^2 \eta k n (1 - e^{-\varepsilon_0}) \leq 2v^2 R_x^2 \eta k n \varepsilon_0 && (16)
\end{aligned}
$$

Where in Eq. (13) we used the Lipschitzness of $\sigma(\cdot): \; \forall x_1, x_2 \in \mathbb{R} |\sigma(x_1) - \sigma(x_2)| \leq |x_1 - x_2|$, in Eq. (14) we used the Cauchy-Shwartz inequality, in Eq. (15) we used the update rule Eq. (2) recursively and finally in Eq. (16) we used that if $\ell(x) \leq \varepsilon_0$ then $|\ell'(x)| \leq 1 - e^{-\varepsilon_0}$ (follows from a similar derivation to Eq. (9)) and that $1 - e^{-\varepsilon_0} \leq \varepsilon_0$.

Now we can use the bound we just derived and the Lipschitzness of ℓ and reach

$$
|\ell\left(y_i N_{\boldsymbol{W}_{t+s}}(\boldsymbol{x}_i)\right) - \ell\left(y_i N_{\boldsymbol{W}_t}(\boldsymbol{x}_i)\right)| \leq 2v^2 R_x^2 \eta k n \varepsilon_0 \tag{17}
$$

for any time $(i_e - 1)n + 1 \leq t \leq i_e n$ and $1 \leq s \leq n$. We know that for all $1 \leq i \leq n$, there exists $(i_e - 1)n + 1 \leq t_i^* \leq i_e n$ such that $\ell(y_i N_{\boldsymbol{W}_{t_i^*-1}}(\boldsymbol{x}_i)) \leq \varepsilon_0$. Therefore, by Eq. (17), for time $T^* = i_e n + 1$ and any $(y_i, \boldsymbol{x}_i) \in \mathbb{S}$ we have:

$$
\ell\left(y_i N_{\boldsymbol{W}_{T^*}}(\boldsymbol{x}_i)\right) \leq \ell\left(y_i N_{\boldsymbol{W}_{t_i^*-1}}(\boldsymbol{x}_i)\right) + 2v^2 R_x^2 \eta k n \varepsilon_0 \leq \varepsilon_0 + 2v^2 R_x^2 \eta k n \varepsilon_0 \tag{18}
$$

If $\forall 1 \leq i \leq n \; \ell(y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)) \leq \varepsilon$ we would get our bound $L_{\mathbb{S}}(\boldsymbol{W}) \leq \varepsilon$.

Therefore, if we set $\varepsilon_0 = \frac{\varepsilon}{1+2v^2 R_x^2 \eta k n}$ in Eq. (18) we'll reach our result.

Setting this ε_0 at Eq. (12) leads to:

$$T \leq \left(\frac{R_x^2}{\alpha^2} + \frac{1}{k\eta v^2 \alpha^2}\right) ||\boldsymbol{w}^*||^2 n^2 \left(1 + \frac{1 + 2v^2 R_x^2 \eta k n}{\varepsilon}\right)^2$$
$$+ \frac{\sqrt{R_0(8k^2\eta^2 v^2 R_x^2 + 8k\eta)}||\boldsymbol{w}^*||^{1.5} n^{1.5} \left(1 + \frac{1+2v^2 R_x^2 \eta k n}{\varepsilon}\right)^{1.5}}{2k(\eta v \alpha)^{1.5}} + \frac{2R_0||\boldsymbol{w}^*||n\left(1 + \frac{1+2v^2R_x^2\eta k n}{\varepsilon}\right)}{\eta v \alpha} \tag{19}$$

We denote the right hand side of Eq. (19) plus n by $M(n, \epsilon)$. [1] Note that $M(n, \epsilon) = O(\frac{n^4}{\varepsilon^2})$ and therefor for simplicity we can alternatively denote $M(n, \epsilon)$ to be a less tight bound of the form $\frac{Cn^4}{\varepsilon^2}$ where C is a constant that depends polynomially on $R_x, R_0, k, \frac{1}{\alpha}, \max\left\{\eta, \frac{1}{\eta}\right\}, \max\left\{v, \frac{1}{v}\right\}$ and $||\boldsymbol{w}^*||$. Overall, we proved that after $O(\frac{n^4}{\varepsilon^2})$ steps, SGD will converge to a solution with $L_{\mathbb{S}}(\boldsymbol{W}_t) < \varepsilon$ empirical loss for some $t \leq M(n, \varepsilon)$.

[1] We need to add n to Eq. (18) because we may consider the epoch immediately after T.

3. Proof of Theorem 5.1

Before we start proving the main theorem we will prove some useful lemmas and corollaries.

We first show the following.

Corollary 3.1. *if* $|(\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}|\geq 2r||\boldsymbol{x}||$ *then* $|\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\vee|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||$.

Proof. Assume in contradiction that $|\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|<r||\boldsymbol{x}||\wedge|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|<r||\boldsymbol{x}||$. then by the triangle inequality and the Cauchy-Shwartz inequality we'll get:

$|(\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}|\leq|\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|+|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|<r||\boldsymbol{x}||+r||\boldsymbol{x}||=2r||\boldsymbol{x}||$ in contradiction to the assumption $|(\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}|\geq 2r||\boldsymbol{x}||$. □

Next, we prove the following lemma, which will be used throughout the proof of the main theorem. The lemma ties the dot products with the center of the cluster to the dot products with the individual neurons:

Lemma 3.1. *If* $\forall 1\leq j\leq k:\ \boldsymbol{w}^{(j)}\in Ball(\overline{\boldsymbol{w}},r)\wedge\boldsymbol{u}^{(j)}\in Ball(\overline{\boldsymbol{u}},r)$ *then:* $\forall\boldsymbol{x}\in\mathbb{R}^d$ *s.t* $|\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||:[\forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}>0]\vee[\forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}<0]$ *and similarly for u type neurons* $\forall\boldsymbol{x}\in\mathbb{R}^d$ *s.t* $|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||:[\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}>0]\vee[\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}<0]$.

Proof. Let's assume that $\overline{\boldsymbol{w}}\cdot\boldsymbol{x}\geq r||\boldsymbol{x}||$, therefore $\forall 1\leq j\leq k:\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}=(\boldsymbol{w}^{(j)}-\overline{\boldsymbol{w}})\cdot\boldsymbol{x}+\overline{\boldsymbol{w}}\cdot\boldsymbol{x}\geq-||\boldsymbol{w}^{(j)}-\overline{\boldsymbol{w}}||\cdot||\boldsymbol{x}||+r||\boldsymbol{x}||>-r||\boldsymbol{x}||+r||\boldsymbol{x}||=0$ where we had used Cauchy-Shwartz inequality and that $||\boldsymbol{w}^{(j)}-\overline{\boldsymbol{w}}||<r$.

If $\overline{\boldsymbol{w}}\cdot\boldsymbol{x}\leq-r||\boldsymbol{x}||$, $\forall 1\leq j\leq k:\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}=(\boldsymbol{w}^{(j)}-\overline{\boldsymbol{w}})\cdot\boldsymbol{x}+\overline{\boldsymbol{w}}\cdot\boldsymbol{x}<||\boldsymbol{w}^{(j)}-\overline{\boldsymbol{w}}||\cdot||\boldsymbol{x}||-r||\boldsymbol{x}||<r||\boldsymbol{x}||-r||\boldsymbol{x}||=0$ the same derivation would work for $\boldsymbol{u}$. □

We are now ready to move forward with proving the main lemma.

By Corollary 3.1 we see that $\{\boldsymbol{x}\in\mathbb{R}^d|\ |(\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}|\geq 2r||\boldsymbol{x}||\}\subseteq\{\boldsymbol{x}\in\mathbb{R}^d|\ |\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\vee|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\}$ so if we prove that:

$\forall\boldsymbol{x}\in\mathbb{R}^d\in\{\boldsymbol{x}\in\mathbb{R}^d|\ |(\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}|\geq 2r||\boldsymbol{x}||\}\cap\{\boldsymbol{x}\in\mathbb{R}^d|\ |\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\vee|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\}:\ \text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right)=\text{sign}\left((\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}\right)$ we will be done.

We'll start by showing first our lemma holds $\forall\boldsymbol{x}\in\mathbb{R}^d$ s.t $|\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\wedge|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||$ and then deal with the points in which only one of the above conditions holds.

Proposition 3.1. $\forall\boldsymbol{x}\in\mathbb{R}^d$ *s.t* $|\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\wedge|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||:$ *sign* $\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right)=$ *sign* $\left((\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}\right)$

Proof. Under our clusterization assumption $\forall 1\leq j\leq k:\ \boldsymbol{w}^{(j)}\in Ball(\overline{\boldsymbol{w}},r)\wedge\boldsymbol{u}^{(j)}\in Ball(\overline{\boldsymbol{u}},r)$ so we can use Lemma 3.1 and we are left with proving that $\forall\boldsymbol{x}\in\mathbb{R}^d$ such that for the $\boldsymbol{w}$ neurons $\{[\forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}>0]\vee[\forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}<0]\}$ and for the $\boldsymbol{u}$ neurons $\{[\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}>0]\vee[\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}<0]\}$ we get $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right)=\text{sign}\left((\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}\right)$.

We can represent $\{\boldsymbol{x}\in\mathbb{R}^d|\ |\overline{\boldsymbol{w}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\wedge|\overline{\boldsymbol{u}}\cdot\boldsymbol{x}|\geq r||\boldsymbol{x}||\}$ as a union of $\{C_+^+,C_-^-,C_+^-,C_-^+\}$ where:

$$C_+^+=\{\boldsymbol{x}\in\mathbb{R}^d|\ \ \forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}>0\text{ and }\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}>0\}$$

$$C_-^-=\{\boldsymbol{x}\in\mathbb{R}^d|\ \ \forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}<0\text{ and }\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}<0\}$$

$$C_+^-=\{\boldsymbol{x}\in\mathbb{R}^d|\ \ \forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}>0\text{ and }\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}<0\}$$

$$C_-^+=\{\boldsymbol{x}\in\mathbb{R}^d|\ \ \forall 1\leq j\leq k\ \ \boldsymbol{w}^{(j)}\cdot\boldsymbol{x}<0\text{ and }\forall 1\leq j\leq k\ \ \boldsymbol{u}^{(j)}\cdot\boldsymbol{x}>0\}$$

Now we will show that $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right)=\text{sign}\left((\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}\right)$ in each region, from which the claim follows.

1. If $\boldsymbol{x}\in C_+^+$ then $N_{\boldsymbol{W}}(\boldsymbol{x})=v\left(\sum_{j=1}^{k}\sigma(\boldsymbol{w}^{(j)}\cdot\boldsymbol{x})-\sigma(\boldsymbol{u}^{(j)}\cdot\boldsymbol{x})\right)=v\left(\sum_{j=1}^{k}\boldsymbol{w}^{(j)}-\boldsymbol{u}^{(j)}\right)\cdot\boldsymbol{x}$ and therefore $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right)=\text{sign}\left((\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}})\cdot\boldsymbol{x}\right)$.

2. If $\boldsymbol{x} \in C_-^-$ then $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = \alpha v \left(\sum_{j=1}^{k} \boldsymbol{w}^{(j)} - \boldsymbol{u}^{(j)}\right) \cdot \boldsymbol{x}$ and therefore $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$

3. If $\boldsymbol{x} \in C_+^-$ then both $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = v\left(\sum_{j=1}^{k} \boldsymbol{w}^{(j)} \cdot \boldsymbol{x} - \alpha \boldsymbol{u}^{(j)} \cdot \boldsymbol{x}\right) > 0$ and $\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - \overline{\boldsymbol{u}} \cdot \boldsymbol{x} > 0$. Therefore, $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$.

4. If $\boldsymbol{x} \in C_-^+$ then both $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = v\left(\sum_{j=1}^{k} \alpha \boldsymbol{w}^{(j)} \cdot \boldsymbol{x} - \boldsymbol{u}^{(j)} \cdot \boldsymbol{x}\right) < 0$ and $\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - \overline{\boldsymbol{u}} \cdot \boldsymbol{x} < 0$. Therefore, $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$.

□

We are left with proving $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$ holds when exactly one condition holds ,i.e., either $|\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}||$ or $|\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}||$.

Proposition 3.2.

$$\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\} :\ \mathit{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \mathit{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$$

and similarly our decision boundary is linear for points in which our condition only holds for $\overline{\boldsymbol{w}}$*:*

$$\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\} :\ \mathit{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \mathit{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$$

Proof. We start with the domain $\{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$

i.e. our condition only holds for $\overline{\boldsymbol{u}}$.

There are two cases, and we'll prove the result for each of them:

If $\overline{\boldsymbol{u}} \cdot \boldsymbol{x} \geq r||\boldsymbol{x}||$:

In this case $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - k\overline{\boldsymbol{u}} \cdot \boldsymbol{x}\right)$.

Next, for any $\boldsymbol{x}$ in the domain, we'll denote $J_+^w(\boldsymbol{x}) \coloneqq \{j | \boldsymbol{w}^{(j)} \cdot \boldsymbol{x} > 0\}$ and $k_+^w(\boldsymbol{x}) \coloneqq |J_+^w(\boldsymbol{x})|$ similarly $J_-^w(\boldsymbol{x}) = \{j | \boldsymbol{w}^{(j)} \cdot \boldsymbol{x} < 0\}$ and $k_-^w(\boldsymbol{x}) \coloneqq |J_-^w(\boldsymbol{x})|$. Using these definitions, our network has the following form:

$$N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j_+ \in J_+^w(\boldsymbol{x})} \boldsymbol{w}^{(j_+)} \cdot \boldsymbol{x} + \alpha \sum_{j_- \in J_-^w(\boldsymbol{x})} \boldsymbol{w}^{(j_-)} \cdot \boldsymbol{x} - k\overline{\boldsymbol{u}} \cdot \boldsymbol{x}\right) = v\left(k\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - k\overline{\boldsymbol{u}} \cdot \boldsymbol{x} + (\alpha - 1) \sum_{j_- \in J_-^w(\boldsymbol{x})} \boldsymbol{w}^{(j_-)} \cdot \boldsymbol{x}\right)$$

Next, we bound $\forall j\ |\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}| = |(\boldsymbol{w}^{(j)} - \overline{\boldsymbol{w}} + \overline{\boldsymbol{w}}) \cdot \boldsymbol{x}| \leq ||\boldsymbol{w}^{(j)} - \overline{\boldsymbol{w}}|| \cdot ||\boldsymbol{x}|| + |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < 2r||\boldsymbol{x}||$ where we used $||\boldsymbol{w}^{(j)} - \overline{\boldsymbol{w}}|| < r$ and $|\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}||$.

Now, if $(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} \geq 2r||\boldsymbol{x}|| > 0$ we get that $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(k(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} - (1 - \alpha) \sum_{j_- \in J_-^w(\boldsymbol{x})} \boldsymbol{w}^{(j_-)} \cdot \boldsymbol{x}\right) > v\left(2r||\boldsymbol{x}||k - 2r||\boldsymbol{x}||k_-^w(\boldsymbol{x})(1 - \alpha)\right) > 0$ since $(1 - \alpha) < 1$ and $k_-^w(\boldsymbol{x}) \leq k$ and therefore $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right) = 1$ for this case.

If $(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} \leq -2r||\boldsymbol{x}|| < 0$ we get that $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(k(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} - (1 - \alpha) \sum_{j_- \in J_-^w(\boldsymbol{x})} \boldsymbol{w}^{(j_-)} \cdot \boldsymbol{x}\right) < v\left(-2r||\boldsymbol{x}||k + 2r||\boldsymbol{x}||k_-^w(\boldsymbol{x})(1 - \alpha)\right) < 0$ since $(1 - \alpha) < 1$ and $k_-^w(\boldsymbol{x}) \leq k$. Therefore, we get that $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right) = -1$ in this case.

At any rate, we have shown that $\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d | \ \ |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge \overline{\boldsymbol{u}} \cdot \boldsymbol{x} \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$: $\text{sign}\,(N_{\boldsymbol{W}}(\boldsymbol{x})) = \text{sign}\,((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x})$.

If $\overline{\boldsymbol{u}} \cdot \boldsymbol{x} \leq -r||\boldsymbol{x}||$:

First, we notice that $(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} > -r||\boldsymbol{x}|| + r||\boldsymbol{x}|| = 0$ so $\text{sign}\,((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}) = 1$ again we use Lemma 3.1 and from our assumption $\overline{\boldsymbol{u}} \cdot \boldsymbol{x} \leq -r||\boldsymbol{x}||$ we have $\forall 1 \leq j \leq k \ \boldsymbol{u}^{(j)} \cdot \boldsymbol{x} < 0$ and we can see that our network takes the form: $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \alpha \cdot k\overline{\boldsymbol{u}} \cdot \boldsymbol{x}\right) \geq v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) + \alpha k r||\boldsymbol{x}||\right)$.
Next, we prove the following lemma:

Lemma 3.2. *If* $|\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}||$ *then* $\alpha \cdot k \cdot r||\boldsymbol{x}|| > -\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x})$.

Proof. Let's assume by contradiction that $-\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) \geq \alpha \cdot k \cdot r||\boldsymbol{x}||$. We notice that regardless of the sign of the dot product $\forall j : -\sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) \leq -\alpha \boldsymbol{w}^{(j)} \cdot \boldsymbol{x}$ so we have $-\alpha \sum_{j=1}^{k} \boldsymbol{w}^{(j)} \cdot \boldsymbol{x} \geq -\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) \geq \alpha \cdot k \cdot r||\boldsymbol{x}||$, which leads to $-\alpha k \overline{\boldsymbol{w}} \cdot \boldsymbol{x} \geq \alpha \cdot k \cdot r||\boldsymbol{x}||$ (where we used the definition of $\overline{\boldsymbol{w}}$) finally we reach $\overline{\boldsymbol{w}} \cdot \boldsymbol{x} \leq -r||\boldsymbol{x}||$. This contradicts $|\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}||$. □

Therefore, we have $-\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) < \alpha \cdot k \cdot r||\boldsymbol{x}||$ and $\text{sign}\,(N_{\boldsymbol{W}}(\boldsymbol{x})) = \text{sign}\,((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x})) = 1$ as desired.

To conclude we proved that $\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d | \ \ |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$, $\text{sign}\,(N_{\boldsymbol{W}}(\boldsymbol{x})) = \text{sign}\,((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x})$.

Next we look at $\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d | \ \ |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$ and through a similar derivation of two cases we will prove that $\text{sign}\,(N_{\boldsymbol{W}}(\boldsymbol{x})) = \text{sign}\,((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x})$.

If $\overline{\boldsymbol{w}} \cdot \boldsymbol{x} \geq r||\boldsymbol{x}||$:

Through a similar derivation for the case of $\overline{\boldsymbol{u}} \cdot \boldsymbol{x} \geq r||\boldsymbol{x}||$, our network has the following form:

$$
\begin{aligned}
N_{\boldsymbol{W}}(\boldsymbol{x}) &= v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = v\left(k\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - \sum_{j=1}^{k} \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) \\
&= vk\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - v\left(\sum_{j_+ \in J_+^u(\boldsymbol{x})} \boldsymbol{u}^{(j_+)} \cdot \boldsymbol{x} + \sum_{j_- \in J_-^u(\boldsymbol{x})} \alpha \boldsymbol{u}^{(j_-)} \cdot \boldsymbol{x}\right) \\
&= vk\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - v\left(\sum_{j_+ \in J_+^u(\boldsymbol{x})} \boldsymbol{u}^{(j_+)} \cdot \boldsymbol{x} + \sum_{j_- \in J_-^u(\boldsymbol{x})} \boldsymbol{u}^{(j_-)} \cdot \boldsymbol{x} + (\alpha - 1) \sum_{j_- \in J_-^u(\boldsymbol{x})} \boldsymbol{u}^{(j_-)} \cdot \boldsymbol{x}\right) \\
&= v\left(k\overline{\boldsymbol{w}} \cdot \boldsymbol{x} - k\overline{\boldsymbol{u}} \cdot \boldsymbol{x} + (1 - \alpha) \sum_{j_- \in J_-^u(\boldsymbol{x})} \boldsymbol{u}^{(j_-)} \cdot \boldsymbol{x}\right)
\end{aligned}
$$

where $J_-^u(\boldsymbol{x}) := \{j | \boldsymbol{u}^{(j)} \cdot \boldsymbol{x} < 0\}$, $J_+^u(\boldsymbol{x}) := \{j | \boldsymbol{u}^{(j)} \cdot \boldsymbol{x} > 0\}$ and $k_-^u(\boldsymbol{x}) = |J_-^u(\boldsymbol{x})|, k_+^u(\boldsymbol{x}) = |J_+^u(\boldsymbol{x})|$.

If $(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} \geq 2r||\boldsymbol{x}|| > 0$ then $N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(k(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} + (1-\alpha) \sum\limits_{j_- \in J_-^u(\boldsymbol{x})} \boldsymbol{u}^{(j_-)} \cdot \boldsymbol{x}\right) \geq v\left(2kr||\boldsymbol{x}|| - 2r||\boldsymbol{x}||(1-\alpha)k_-^u(\boldsymbol{x})\right) > 0$ (because $(1-\alpha) < 1$ and $k_-^u(\boldsymbol{x}) \leq k$) and $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right) = 1$ (where we used the fact that $\forall j : |\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}| < 2r||\boldsymbol{x}||$ which follows from $|\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}||$ and $||\boldsymbol{u}^{(j)} - \overline{\boldsymbol{u}}|| < r$).

If $(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x} \leq -2r||\boldsymbol{x}|| < 0$ we get that $N_{\boldsymbol{W}}(\boldsymbol{x}) \leq v\left(-2r||\boldsymbol{x}||k + 2r||\boldsymbol{x}||(1-\alpha)k_-^u(\boldsymbol{x})\right) < 0$ and $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right) = 1$.

To summarize, we showed that $\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge \overline{\boldsymbol{w}} \cdot \boldsymbol{x} \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$, $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$.

<u>If $\overline{\boldsymbol{w}} \cdot \boldsymbol{x} \leq -r||\boldsymbol{x}||$:</u>

We again use Lemma 3.1 which yields from $\overline{\boldsymbol{w}} \cdot \boldsymbol{x} \leq -r||\boldsymbol{x}||$ that $\forall 1 \leq j \leq k\ \boldsymbol{w}^{(j)} \cdot \boldsymbol{x} < 0$ and we can see that our network takes the form:

$$N_{\boldsymbol{W}}(\boldsymbol{x}) = v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) = \alpha k v \overline{\boldsymbol{w}} \cdot \boldsymbol{x} - v\left(\sum_{j=1}^{k} \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right) \leq v\left(-\alpha k r||\boldsymbol{x}|| - \sum_{j=1}^{k} \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x})\right)$$

If $-\sum\limits_{j=1}^{k} \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}) < \alpha k r||\boldsymbol{x}||$ we have $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right) = -1$ as desired.

The same contradiction proof from $\overline{\boldsymbol{u}} \cdot \boldsymbol{x} \leq -r||\boldsymbol{x}||$ segment above (Lemma 3.2) would show

$-\sum\limits_{j=1}^{k} \sigma(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}) < \alpha \cdot k \cdot r||\boldsymbol{x}||$ (just exchange w and u) and we'll get $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right) = -1$.

Finally, we proved that

$$\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}\ \text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$$

and that

$$\forall \boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}\ \text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$$

as required. □

We can now combine Corollary 3.1, Proposition 3.1 and Proposition 3.2 and prove Theorem 5.1:

We have $\forall \boldsymbol{x} \in \mathbb{R}^d$ s.t $|(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||$ then $|\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \vee |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}||$. If $\boldsymbol{x}$ is such that $|\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}||$ we can use Proposition (3.1) and get $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$.

If only one condition holds i.e. $\boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$ or $\boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d |\ |\overline{\boldsymbol{u}} \cdot \boldsymbol{x}| < r||\boldsymbol{x}|| \wedge |\overline{\boldsymbol{w}} \cdot \boldsymbol{x}| \geq r||\boldsymbol{x}|| \wedge |(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||\}$ then we can use Proposition (3.2) and get $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$.

Therefore, overall for $|(\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}| \geq 2r||\boldsymbol{x}||$ we get $\text{sign}\left(N_{\boldsymbol{W}}(\boldsymbol{x})\right) = \text{sign}\left((\overline{\boldsymbol{w}} - \overline{\boldsymbol{u}}) \cdot \boldsymbol{x}\right)$ as required.

3.1. Proof of Corollary 6.1

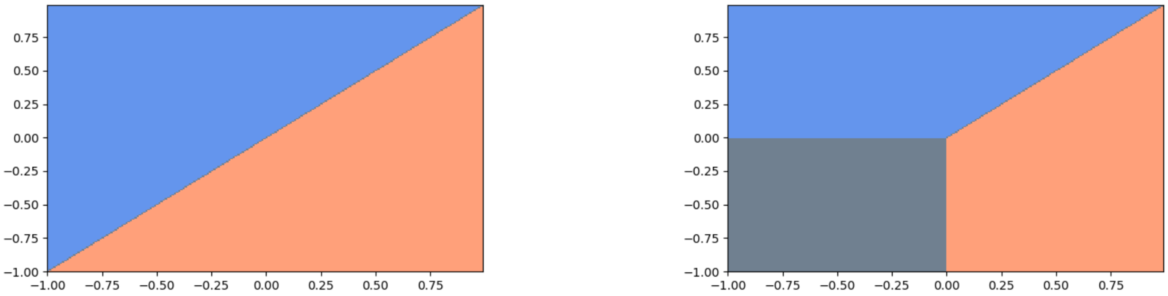
Since the network is perfectly clustered, the corollary follows by Proposition (3.1) with $r = 0$.

4. Additional Experiments - Linear Decision Boundary

In this section we provide additional empirical evaluations of the decision boundary that SGD converges to in our setting.

4.1. Leaky ReLU vs ReLU decision boundary

Theorem 5.1 addresses the case of Leaky ReLU activation. Here we show that the result is indeed not true for ReLU networks. We compare two perfectly clustered networks (i.e., each with two neurons) one with a Leaky ReLU activation and the other with a ReLU activation. Figure 1 shows a decision boundary for a two neuron network, in the case of Leaky ReLU (Figure 1a) and ReLU (Figure 1b). It can be seen that the leaky ReLU indeed provides a linear decision boundary, as predicted by Theorem 5.1, whereas the ReLU case is non-linear (we explicitly show the regime where the network output is zero. This can be orange or blue, depending on whether zero is given label positive or negative. In any case the resulting boundary is non-linear).



(a) Leaky ReLU network - Linear Decision Boundary

(b) ReLU network - Non Linear Decision Boundary

Figure 1: The prediction landscape for two neuron networks with Leaky ReLU and ReLU activations. Orange for positive prediction, blue for a negative prediction and grey for zero prediction. The $\boldsymbol{w}$ neuron is $(1, 0) \in \mathbb{R}^2$ and the $\boldsymbol{u}$ neuron is $(0, 1) \in \mathbb{R}^2$.

4.2. MNIST - Linear Regime

In Figure 2 in the main text we saw how for MNIST digit pairs (0,1) and (3,5) the network enters the linear regime at some point in the training process. In Figure 2 we see the robustness of this behavior across the MNIST data-set by showing the above holds for more pairs of digits.

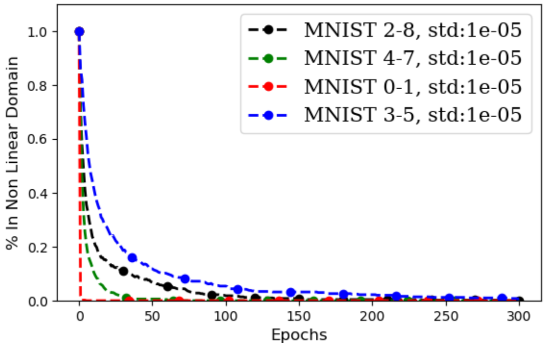
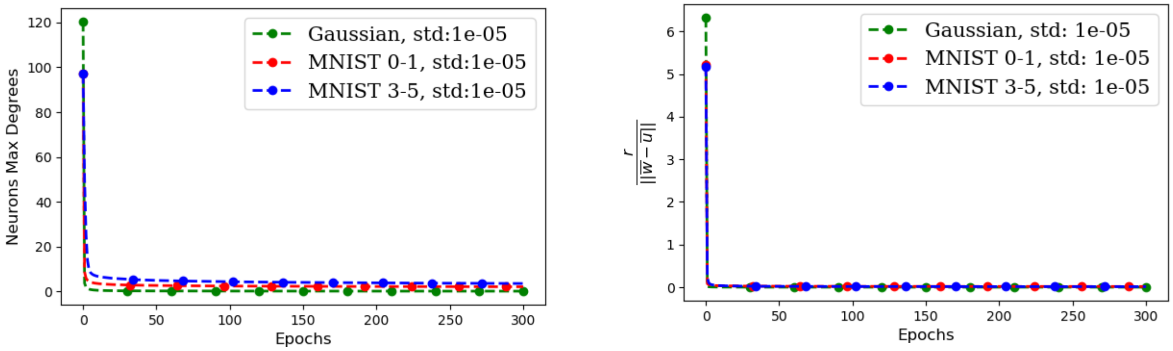


Figure 2: Convergence to a classifier that is linear on the data, for MNIST pairs. Each line corresponds to an average over 5 initializations.

4.3. Clustering of Neurons - Empirical Evidence

In Section 5 in the main text and Figure 2 above, we saw that learning converges to a linear decision boundary on the train and test points. Theorem 5.1 suggests that this will happen if neurons are well clustered (in the $\boldsymbol{w}$ and $\boldsymbol{u}$ groups). Here we show that indeed clustering occurs.

We consider two different measures of clustering. The first is the ratio $\frac{r}{||\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}}||}$, and the second is the maximum angle between the neurons of the same type (i.e., the maximal angle between vectors in the same cluster). Figure 3 shows these two measures as a function of the training epochs. They can indeed be seen to converge to zero, which by Theorem 5.1 implies convergence to a linear decision boundary.



(a) Max Angle In Same Cluster Neurons

(b) $\frac{r}{||\overline{\boldsymbol{w}}-\overline{\boldsymbol{u}}||}$

Figure 3: Evaluation of clustering measures during training. We consider two different clustering measures in (a) and (b) (see text). It can be seen that both measures converge to zero.

5. Assumptions for Gradient Flow Analysis

In the paper we use results from (Lyu & Li, 2020) and (Ji & Telgarsky, 2020). Here we show that the assumptions required by these theorems are satisfied in our setup.

The assumptions in (Lyu & Li, 2020) and (Ji & Telgarsky, 2020) are:

(**A1**) . (Regularity). For any fixed $\boldsymbol{x}, \Phi(\cdot; \boldsymbol{x})$ is locally Lipschitz and admits a chain rule;

(**A2**) . (Homogeneity). There exists $L > 0$ such that $\forall \alpha > 0 : \Phi(\alpha \boldsymbol{W}; \boldsymbol{x}) = \alpha^L \Phi(\boldsymbol{W}; \boldsymbol{x})$;

(**B3**) . The loss function $\ell(q)$ can be expressed as $\ell(q) = e^{-f(q)}$ such that

(B3.1). $f : \mathbb{R} \to \mathbb{R}$ is $\mathcal{C}^1$-smooth.

(B3.2). $f'(q) > 0$ for all $q \in \mathbb{R}$.

(B3.3). There exists $b_f \geq 0$ such that $f'(q)q$ is non-decreasing for $q \in (b_f, +\infty)$, and $f'(q)q \to +\infty$ as $q \to +\infty$.

(B3.4). Let $g : [f(b_f), +\infty) \to [b_f, +\infty)$ be the inverse function of f on the domain $[b_f, +\infty)$. There exists $b_g \geq max\{2f(b_f), f(2b_f)\}, K \geq 1$ such that $g'(x) \leq Kg'(\theta x)$ and $f'(y) \leq Kf'(\theta y)$ for all $x \in (b_g, +\infty), y \in (g(b_g), +\infty)$ and $\theta \in [1/2, 1)$

(**B4**). (Separability). There exists a time t_0 such that $\mathcal{L}(\boldsymbol{W}) < e^{-f(b_f)} = \ell(b_f)$

We next show that these are satisfied in our setup.

Proof. (**A1**). (Regularity) first we show that $\Phi(\cdot;\boldsymbol{x})$ is locally Lipschitz, with slight abuse of notations, let $\boldsymbol{W}_1 = \overrightarrow{\boldsymbol{W}}_1, \boldsymbol{W}_2 = \overrightarrow{\boldsymbol{W}}_2 \in \mathbb{R}^{2kd}$ so in our case:

$$\begin{aligned}\Phi(\boldsymbol{W}_1;\boldsymbol{x}) - \Phi(\boldsymbol{W}_2;\boldsymbol{x}) &= \boldsymbol{v}\cdot\sigma(\boldsymbol{W}_1\cdot\boldsymbol{x}) - \boldsymbol{v}\cdot\sigma(\boldsymbol{W}_2\cdot\boldsymbol{x})\\ &= v\left[\sum_{j=1}^{k}\sigma\left(\boldsymbol{w}_1^{(j)}\cdot\boldsymbol{x}\right) - \sigma\left(\boldsymbol{w}_2^{(j)}\cdot\boldsymbol{x}\right) - \left(\sigma\left(\boldsymbol{u}_1^{(j)}\cdot\boldsymbol{x}\right) - \sigma\left(\boldsymbol{u}_2^{(j)}\cdot\boldsymbol{x}\right)\right)\right]\end{aligned}$$

and therefore

$$\begin{aligned}&||\Phi(\boldsymbol{W}_1;\boldsymbol{x}) - \Phi(\boldsymbol{W}_2;\boldsymbol{x})||\\ &= \left|\left|v\left[\sum_{j=1}^{k}\sigma\left(\boldsymbol{w}_1^{(j)}\cdot\boldsymbol{x}\right) - \sigma\left(\boldsymbol{w}_2^{(j)}\cdot\boldsymbol{x}\right) - \left(\sigma\left(\boldsymbol{u}_1^{(j)}\cdot\boldsymbol{x}\right) - \sigma\left(\boldsymbol{u}_2^{(j)}\cdot\boldsymbol{x}\right)\right)\right]\right|\right|\\ &\leq v\left[\sum_{j=1}^{k}\left|\left|\sigma\left(\boldsymbol{w}_1^{(j)}\cdot\boldsymbol{x}\right) - \sigma\left(\boldsymbol{w}_2^{(j)}\cdot\boldsymbol{x}\right)\right|\right| + \left|\left|\sigma\left(\boldsymbol{u}_1^{(j)}\cdot\boldsymbol{x}\right) - \sigma\left(\boldsymbol{u}_2^{(j)}\cdot\boldsymbol{x}\right)\right|\right|\right]\\ &\leq 2v||\boldsymbol{x}||\left[\sum_{j=1}^{k}||\boldsymbol{w}_1^{(j)} - \boldsymbol{w}_2^{(j)}|| + ||\boldsymbol{u}_1^{(j)} - \boldsymbol{u}_2^{(j)}||\right] = 2v\cdot||\boldsymbol{x}||\cdot||\overrightarrow{\boldsymbol{W}}_1 - \overrightarrow{\boldsymbol{W}}_2||\end{aligned}$$

And we showed $\Phi(\cdot;\boldsymbol{x})$ is globally Lipschitz (and therfor locally Lispchitz). Next for the chain rule, as shown in (Davis et al., 2018) (corollary for deep learning therein), any function definable in an o-minimal structure admits a chain rule. Our network is definable because algebraic, composition, inverse, maximum and minimum operations over definable functions are also definable. Leaky ReLUs are definable as maximum operations over two linear functions (linear functions are definable).and because Leaky ReLUs are definable our network is also definable.

(**A2**). (Homogeneity). It is easy to see from the definition that in our case, the trainable parameters are only the first layer weights and the network $\Phi(\cdot;\boldsymbol{x})$ is $L = 1$ homogeneous.

(**B3**). As seen in Lyu & Li (2020) (Remark A.2. therein) the logistic loss $\ell(q) = \log(1 + e^{-q})$ satisfies (**B3**) with $f(q) = -\log\left(\log(1 + e^{-q})\right), g(q) = -\log\left(e^{e^{-q}} - 1\right), b_f = 0$.

(**B4**). (Separability). This is Assumption 6.1 in the main text. As we mentioned in the main text, this assumption is satisfied with SGD by Theorem 4.1.

□

6. Proof of Theorem 6.1

In this proof we will show that the normalized parameters $\hat{\boldsymbol{W}}_t \coloneqq \frac{\boldsymbol{W}_t}{||\boldsymbol{W}_t||}$ under gradient flow optimization, converges to a solution in $\mathcal{N}$ and that the network $N_{\hat{\boldsymbol{W}}}$ at convergence is perfectly clustered. Under our assumption $\forall t \geq T_{NAR}$ $\hat{\boldsymbol{W}}_t \in \mathcal{N}$. From the definition of the NAR it's easy to see that the NAR is a closed domain. Therefore any limit point of $\hat{\boldsymbol{W}}_t$ is also in the NAR. From Ji & Telgarsky (2020) (Theorem 3.1. therein) we have that the normalized parameters flow converges when using gradient flow. To conclude so far, we had shown that $\hat{\boldsymbol{W}}_t$ converges to a point inside the NAR $\mathcal{N}$.

We are left with showing that the limit point of $\lim_{t\to\infty}\hat{\boldsymbol{W}}_t \coloneqq \hat{\boldsymbol{W}}_*$ has a perfectly clustered form.

Lyu & Li (2020) (Theorem A.8. therein) shows that every limit point of $\hat{\boldsymbol{W}}_t$ is along the direction of a KKT point of the following optimization problem (P):

$$\begin{aligned}&\min \frac{1}{2}||\boldsymbol{W}||_2^2\\ \text{s.t.}\quad &q_i(\boldsymbol{W}) \geq 1 \qquad \forall i \in [n]\end{aligned}$$

where $q_i(\boldsymbol{W}) = y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i)$ is the network margin on the sample point $(y_i, \boldsymbol{x}_i)$.[2]

We are left with showing that at convergence the neurons align in two directions. We will use a characterization of the KKT points of (P) and show that they are perfectly clustered. Since every limit point of the normalized parameters flow is along the direction of a KKT point of (P) that would mean $\hat{\boldsymbol{W}}_*$ has a perfectly clustered form.

A feasible point $\boldsymbol{W}$ of (P) is a KKT point if there exist $\lambda_1, \dots, \lambda_n \geq 0$ such that:

1. $\boldsymbol{W} - \sum\limits_{i=1}^{n} \lambda_i \boldsymbol{h}_i = 0$ for some $\boldsymbol{h}_1, \dots, \boldsymbol{h}_n$ satisfying $\boldsymbol{h}_i \in \partial^\circ q_i(\boldsymbol{W})$
2. $\forall i \in [n] : \lambda_i(q_i(\boldsymbol{W}) - 1) = 0$

From Lyu & Li (2020) (Theorem A.8. therein) we know $\exists \beta$ s.t. $\beta \hat{\boldsymbol{W}}_*$ is a KKT point of (P). Since our limit point is in an NAR we don't need to worry about the non differential points of the network because $\forall 1 \leq j \leq k, i \in [n] : \boldsymbol{w}_*^{(j)} \cdot \boldsymbol{x}_i \neq 0 \wedge \boldsymbol{u}_*^{(j)} \cdot \boldsymbol{x}_i \neq 0$. (where $\boldsymbol{w}_*^{(j)}$ and $\boldsymbol{u}_*^{(j)}$ stands for the $\boldsymbol{w}$ and $\boldsymbol{u}$ type neurons of $\boldsymbol{W}_*$, respectively). Therefore the Clarke subdifferential coincides with the gradient in our domain, and we can derive it using calculus rules.

By looking at the gradient of the margin for any point $(y_i, \boldsymbol{x}_i)$:

- $\dfrac{\partial q_i(\boldsymbol{W})}{\partial \boldsymbol{w}^{(j)}} = \dfrac{y_i \partial N_{\boldsymbol{W}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}^{(j)}} = y_i v \boldsymbol{x}_i \sigma'(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_i) = y_i v \boldsymbol{x}_i \sigma'(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_i)$
- $\dfrac{\partial q_i(\boldsymbol{W})}{\partial \boldsymbol{u}^{(j)}} = \dfrac{y_i \partial N_{\boldsymbol{W}}(\boldsymbol{x}_i)}{\partial \boldsymbol{u}^{(j)}} = -y_i v \boldsymbol{x}_i \sigma'(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_i) = -y_i v \boldsymbol{x}_i \sigma'(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_i)$

Now using the above gradients implies that: $\partial q_i(\boldsymbol{W}) = y_i v \boldsymbol{x}_i(\overbrace{\sigma'(\boldsymbol{w}^{(1)} \cdot \boldsymbol{x}_i), \dots, \sigma'(\boldsymbol{w}^{(k)} \cdot \boldsymbol{x}_i)}^{k}, \overbrace{-\sigma'(\boldsymbol{u}^{(1)} \cdot \boldsymbol{x}_i), \dots, -\sigma'(\boldsymbol{u}^{(k)} \cdot \boldsymbol{x}_i)}^{k})$

By the definition of the NAR $\mathcal{N}$ with parameters $(\beta, c_i^{\boldsymbol{w}}, c_i^{\boldsymbol{u}})$ the dot product of a point $\boldsymbol{x}_i$ with all neurons of the same type is of the same sign, i.e.:

$$\forall i \in [n], \forall 1 \leq l, p \leq k : \sigma'(\boldsymbol{w}^{(l)} \cdot \boldsymbol{x}_i) = \sigma'(\boldsymbol{w}^{(p)} \cdot \boldsymbol{x}_i) = c_i^{\boldsymbol{w}}$$

and

$$\forall i \in [n], \forall 1 \leq l, p \leq k : \sigma'(\boldsymbol{u}^{(l)} \cdot \boldsymbol{x}_i) = \sigma'(\boldsymbol{u}^{(p)} \cdot \boldsymbol{x}_i) = c_i^{\boldsymbol{u}}$$

It follows that for $\boldsymbol{W} \in \mathcal{N}$, $\partial q_i(\boldsymbol{W}) = y_i \cdot v \cdot \boldsymbol{x}_i(\overbrace{c_i^{\boldsymbol{w}}, \dots, c_i^{\boldsymbol{w}}}^{k}, \overbrace{-c_i^{\boldsymbol{u}}, \dots, -c_i^{\boldsymbol{u}}}^{k})$.

Therefore, by the definition of a KKT point we have:

$$\hat{\boldsymbol{W}}_* = \frac{1}{\beta}\left(\overbrace{\sum_{i=1}^{n} \lambda_i y_i v \boldsymbol{x}_i c_i^{\boldsymbol{w}}, \dots, \sum_{i=1}^{n} \lambda_i y_i\ v \boldsymbol{x}_i c_i^{\boldsymbol{w}}}^{k}, \overbrace{-\sum_{i=1}^{n} \lambda_i y_i v \boldsymbol{x}_i c_i^{\boldsymbol{u}}, \dots, -\sum_{i=1}^{n} \lambda_i y_i v \boldsymbol{x}_i c_i^{\boldsymbol{u}}}^{k}\right) \in \mathbb{R}^{2kd}$$

We can see that the first k entries are equal, as well as the next k entries (equal to each other and not to the first k entries).

Therefore the normalized parameters flow $\hat{\boldsymbol{W}}_t$ converges to a perfectly clustered solution.

6.1. Proof Of Corollary 6.2.

By Theorem 6.1, we know the normalized parameters $\hat{\boldsymbol{W}}_t$ are perfectly clustered at convergence so by Corollary 6.1 we get that the decision boundary of $N_{\hat{\boldsymbol{W}}}$ is linear at convergence. From the homogeneity of the network we have $N_{\boldsymbol{W}}(\boldsymbol{x}) = ||\boldsymbol{W}|| N_{\hat{\boldsymbol{W}}}(\boldsymbol{x})$ for any $\boldsymbol{W} \in \mathbb{R}^{2kd}$ and because the norm is a non negative scalar we get $\operatorname{sign}(N_{\boldsymbol{W}}(\boldsymbol{x})) = \operatorname{sign}(N_{\hat{\boldsymbol{W}}}(\boldsymbol{x}))$, i.e., $N_{\boldsymbol{W}}$ and $N_{\hat{\boldsymbol{W}}}$ are the same classifiers. Therefore, this implies that the decision boundary of $N_{\boldsymbol{W}}$ is linear at convergence.[3]

[2] It is not hard to see that given that the solution is in an NAR, then this optimization problem is convex.

[3] We use $\operatorname{sign}(\infty) = 1$ and $\operatorname{sign}(-\infty) = -1$, since the norm $||\boldsymbol{W}||$ diverges.

7. Proof of Theorem 6.2

We divide the proof of Theorem 6.2 into two parts. First, we show that the NAR is a PAR, and then we show that if a network enters and remains in the PAR the network weights at convergence are proportional to the solutions of the SVM problem we defined in the main text.

7.1. The NAR is a PAR

In this subsection we will prove the NAR is in fact a PAR under the conditions of the theorem. In the first step we show that for all $\boldsymbol{w}^{(i)}$'s, $\left(\frac{\boldsymbol{w}^{(i)}}{||\boldsymbol{w}^{(i)}||}\right)\cdot\boldsymbol{x}_+ \geq \beta$ for all positive $\boldsymbol{x}_+ \in \mathbb{S}_+$ and times $t \geq T_{Margin}$. Assume by contradiction that the latter does not hold. Thus, by assumption 2 the network is in a NAR(β) and there exists a positive $\boldsymbol{x}_+ \in \mathbb{S}_+$ such that $\left(\frac{\boldsymbol{w}^{(i)}}{||\boldsymbol{w}^{(i)}||}\right)\cdot\boldsymbol{x}_+ \leq -\beta$ for all $\boldsymbol{w}^{(i)}$. Denote by $\overline{\gamma}_{t,\{\boldsymbol{x}\}}$ the margin of the network at time $t \geq T_{Margin}$ on the point $\boldsymbol{x}$. we notice that $\overline{\gamma}_t \leq \overline{\gamma}_{t,\{\boldsymbol{x}\}}$ by definition. Then:

$$\tilde{\gamma}_t \leq \overline{\gamma}_t \leq \overline{\gamma}_{t,\{\boldsymbol{x}_+\}} = \frac{+1\cdot N_{\boldsymbol{W}}(\boldsymbol{x}_+)}{\left\|\overrightarrow{\boldsymbol{W}}\right\|} = \frac{v\left(\sum\limits_{i=1}^{k}\sigma\left(\boldsymbol{w}_t^{(i)}\cdot\boldsymbol{x}_+\right) - \sum\limits_{i=1}^{k}\sigma\left(\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+\right)\right)}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{w}_t^{(i)}\right\|^2 + \left\|\boldsymbol{u}_t^{(i)}\right\|^2}} \tag{20}$$

$$\leq \frac{v\left(\sum\limits_{i=1}^{k}\sigma\left(\boldsymbol{w}_t^{(i)}\cdot\boldsymbol{x}_+\right) - \sum\limits_{i=1}^{k}\sigma\left(\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+\right)\right)}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}} \leq \frac{v\left(-\sum\limits_{i=1}^{k}\sigma\left(\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+\right)\right)}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}} \leq \frac{v\alpha\left(\sum\limits_{i=1}^{k}\left|\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+\right|\right)}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}} \tag{21}$$

$$\leq \frac{v\alpha\left(\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|\cdot||\boldsymbol{x}_+||\right)}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}} \leq \frac{v\alpha\left(\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|\right)\cdot \underset{i\in[n]}{max}||\boldsymbol{x}_i||}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}}$$

$$= \frac{v\cdot\alpha\cdot\left\|\left(\overbrace{\left\|\boldsymbol{u}_t^{(1)}\right\|,\ldots,\left\|\boldsymbol{u}_t^{(k)}\right\|}^{k}\right)\right\|_1\cdot \underset{i\in[n]}{max}||\boldsymbol{x}_i||}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}}$$

where the first inequality follows by Lyu & Li (2020) (Theorem A.7. therein). In Eq. (21) we noticed that $-\sum\limits_{i=1}^{k}\sigma\left(\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+\right)$ is largest when $\forall 1 \leq i \leq k \quad \boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+ < 0$ and therefore $\sigma\left(\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+\right) = \alpha\boldsymbol{u}_t^{(i)}\cdot\boldsymbol{x}_+$. Therefore, by the inequality $\forall\boldsymbol{v}\in\mathbb{R}^k \quad ||\boldsymbol{v}||_1 \leq \sqrt{k}\cdot||\boldsymbol{v}||_2$, we have:

$$\tilde{\gamma}_t \leq \frac{v\cdot\alpha\cdot\sqrt{k}\left\|\left(\overbrace{\left\|\boldsymbol{u}_t^{(1)}\right\|,\ldots,\left\|\boldsymbol{u}_t^{(k)}\right\|}^{k}\right)\right\|_2\cdot \underset{i\in[n]}{max}||\boldsymbol{x}_i||}{\sqrt{\sum\limits_{i=1}^{k}\left\|\boldsymbol{u}_t^{(i)}\right\|^2}} = \sqrt{k}\cdot\alpha\cdot v\cdot \underset{i\in[n]}{max}||\boldsymbol{x}_i|| \tag{22}$$

Now under assumption 3 there exists a time $T_{Margin} \geq T_{NAR}$ such that $\tilde{\gamma}_{T_{Margin}} > \sqrt{k}\alpha v\cdot\max\limits_{i\in[n]}||\boldsymbol{x}_i||$. By Lyu & Li (2020) (Theorem A.7. therein) the smoothed margin $\tilde{\gamma}_t$ is a non-decreasing function and we will get that $\forall t \geq T_{Margin}$:

$\tilde{\gamma}_t > \sqrt{k}\alpha v \cdot \max\limits_{i\in[n]} ||\boldsymbol{x}_i||$ which is a contradiction to Eq. (22). Hence, $\forall 1 \le i \le k \wedge \boldsymbol{x} \in \mathbb{S}_+ : \left(\frac{\boldsymbol{w}_t^{(i)}}{||\boldsymbol{w}_t^{(i)}||}\right) \cdot \boldsymbol{x} \ge \beta$.

In a similar fashion, assume there is some $\boldsymbol{x}_- \in \mathbb{S}_-$ such that for $\left(\frac{\boldsymbol{u}_t^{(l)}}{||\boldsymbol{u}_t^{(l)}||}\right) \cdot \boldsymbol{x}_- \ge \beta$ doesn't hold. Then by assumption 1 the network is in a NAR, $\left(\frac{\boldsymbol{u}_t^{(l)}}{||\boldsymbol{u}_t^{(l)}||}\right) \cdot \boldsymbol{x}_- \le -\beta$ and by symmetry again we get:

$$\begin{aligned}\tilde{\gamma}_t \le \overline{\gamma}_t \le \overline{\gamma}_{t,\{\boldsymbol{x}_-\}} = \frac{-1 \cdot N_{\boldsymbol{W}}(\boldsymbol{x}_-)}{||\overrightarrow{\boldsymbol{W}}_t||} = \frac{-v\left(\sum\limits_{i=1}^{k} \sigma\left(\boldsymbol{w}_t^{(i)} \cdot \boldsymbol{x}_-\right) - \sum\limits_{i=1}^{k} \sigma\left(\boldsymbol{u}_t^{(i)} \cdot \boldsymbol{x}_-\right)\right)}{\sqrt{\sum\limits_{i=1}^{k} \left|\left|\boldsymbol{w}_t^{(i)}\right|\right|^2 + \left|\left|\boldsymbol{u}_t^{(i)}\right|\right|^2}} \\ \cdots \le \sqrt{k}\alpha v \cdot \max_{i\in[n]} ||\boldsymbol{x}_i||\end{aligned}$$

By Lyu & Li (2020) (Theorem A.7. therein) we reach a contradiction to the network margin assumption again, so $\forall \boldsymbol{x} \in \mathbb{S}_- : \left(\frac{\boldsymbol{u}_t^{(l)}}{\left|\left|\boldsymbol{u}_t^{(l)}\right|\right|}\right) \cdot \boldsymbol{x} \ge \beta$.

To conclude, we have proven so far for all $t > T_{Margin}$:

1. $\forall 1 \le i \le k :, \quad \forall \boldsymbol{x} \in \mathbb{S}_+ : \quad \left(\frac{\boldsymbol{w}_t^{(i)}}{||\boldsymbol{w}_t^{(i)}||}\right) \cdot \boldsymbol{x} \ge \beta.$
2. $\forall 1 \le i \le k :, \quad \forall \boldsymbol{x} \in \mathbb{S}_- : \quad \left(\frac{\boldsymbol{u}_t^{(l)}}{\left|\left|\boldsymbol{u}_t^{(l)}\right|\right|}\right) \cdot \boldsymbol{x} \ge \beta.$

Now, by assumption 4, $\forall \boldsymbol{x} \in \mathbb{S}_- : \quad \left(\frac{\boldsymbol{w}_t^{(i)}}{||\boldsymbol{w}_t^{(i)}||}\right) \cdot \boldsymbol{x} \ngeq \beta$ and similarly $\forall \boldsymbol{x} \in \mathbb{S}_+ : \quad \left(\frac{\boldsymbol{u}_t^{(i)}}{||\boldsymbol{u}_t^{(i)}||}\right) \cdot \boldsymbol{x} \ngeq \beta$. This follows since otherwise $\mathbb{V}_\beta^+(\mathbb{S})$ and $\mathbb{V}_\beta^-(\mathbb{S})$ would not be empty in contradiction to assumption 4.

Next, under the network being in an NAR assumption we have for all $t > T_{Margin}$:

1. $\forall \boldsymbol{x} \in \mathbb{S}_- : \quad \left(\frac{\boldsymbol{w}_t^{(i)}}{||\boldsymbol{w}_t^{(i)}||}\right) \cdot \boldsymbol{x} \le -\beta$
2. $\forall \boldsymbol{x} \in \mathbb{S}_+ : \quad \left(\frac{\boldsymbol{u}_t^{(i)}}{||\boldsymbol{u}_t^{(i)}||}\right) \cdot \boldsymbol{x} \le -\beta$

Thus, for all $t > T_{Margin}$, the network is in PAR(β).

7.2. PAR alignment direction

Now we will find where the parameters converge to when the network is in the PAR(β). By Theorem 6.1, the normalized gradient flow converges to a perfectly clustered solution, i.e., $\lim\limits_{t\to\infty} \hat{\boldsymbol{W}}_t \coloneqq \hat{\boldsymbol{W}}_*$ is of a perfectly clustered form. Formally that means $\exists\beta$ and $\exists\delta$ such that the normalized parameters $\hat{\boldsymbol{W}}$ are of the form $\hat{\boldsymbol{W}}_* = (\beta\tilde{\boldsymbol{w}}, \ldots, \beta\tilde{\boldsymbol{w}}, \delta\tilde{\boldsymbol{u}}, \ldots, \delta\tilde{\boldsymbol{u}}) \in \mathbb{R}^{2kd}$ and WLOG we can assume $||\tilde{\boldsymbol{w}}|| = ||\tilde{\boldsymbol{u}}|| = 1$.

Because the solution is in the PAR(β), the network margins are given as follows for positive points:

$$\begin{aligned}\forall \boldsymbol{x}_i \in \mathbb{S}_+ : q_i(\boldsymbol{W}) = y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i) = y_i ||\boldsymbol{W}|| N_{\hat{\boldsymbol{W}}}(\boldsymbol{x}_i) &= v||\boldsymbol{W}|| \left(\sum_{i=1}^{k} \sigma(\beta\tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_i) - \sigma(\delta\tilde{\boldsymbol{u}} \cdot \boldsymbol{x}_i)\right) \\ &= v||\boldsymbol{W}|| \left(k\beta\tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_i - \alpha k\delta\tilde{\boldsymbol{u}} \cdot \boldsymbol{x}_i\right)\end{aligned}$$

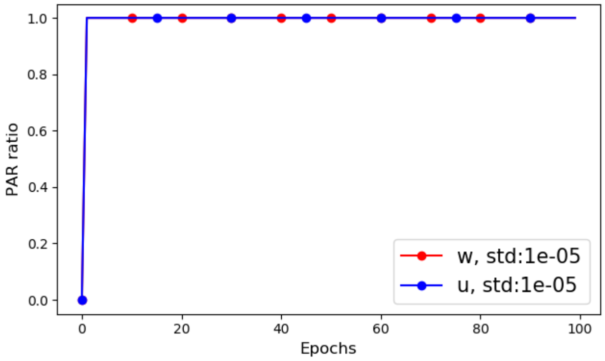


Figure 4: The ratio of neurons from each type in the PAR throughout the training process. We sample 400 data points from two antipodal separable Gaussians (one for each label) in $\mathbb{R}^{50}$. Our network is of 100 neurons (50 of each type) optimized on the data using SGD with batch size 1 with learning rate $\eta = 10^{-3}$.

and negative points:

$$\begin{aligned}\forall \boldsymbol{x}_i \in \mathbb{S}_- : q_i(\boldsymbol{W}) = y_i N_{\boldsymbol{W}}(\boldsymbol{x}_i) = y_i ||\boldsymbol{W}|| N_{\hat{\boldsymbol{W}}}(\boldsymbol{x}_i) &= v||\boldsymbol{W}|| \left(\sum_{i=1}^{k} \sigma(\delta \tilde{\boldsymbol{u}} \cdot \boldsymbol{x}_i) - \sigma(\beta \tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_i) \right) \\ &= v||\boldsymbol{W}|| \left(k\delta \tilde{\boldsymbol{u}} \cdot \boldsymbol{x}_i - \alpha k \beta \tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_i \right)\end{aligned}$$

where we used the fact we know the normalized solution would has a perfectly clustered form. We denote $\tilde{\beta} := ||\boldsymbol{W}|| \cdot \beta$ and similarly $\tilde{\delta} := ||\boldsymbol{W}|| \cdot \delta$

Using the above notations, the max margin problem in Lyu & Li (2020) (Theorem A.8. therein) takes the form:

$$\begin{gathered}\underset{\tilde{\beta}\in\mathbb{R}, \tilde{\delta}\in\mathbb{R}}{\arg\min} \quad k\tilde{\beta}^2 + k\tilde{\delta}^2 = \underset{\tilde{\beta}\in\mathbb{R}, \tilde{\delta}\in\mathbb{R}}{\arg\min} \quad v^2k^2\tilde{\beta}^2 + v^2k^2\tilde{\delta}^2 \\ \forall \boldsymbol{x}_+ \in \mathbb{S}_+ : vk\tilde{\beta}\tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_+ - \alpha v k \tilde{\delta} \tilde{\boldsymbol{u}} \cdot \boldsymbol{x}_+ \geq 1 \\ \forall \boldsymbol{x}_- \in \mathbb{S}_- : vk\tilde{\delta}\tilde{\boldsymbol{u}} \cdot \boldsymbol{x}_- - \alpha v k \tilde{\beta} \tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_- \geq 1\end{gathered}$$

Now we can denote $\boldsymbol{w} := vk\tilde{\beta}\tilde{\boldsymbol{w}}$ and $\boldsymbol{u} := vk\tilde{\delta}\tilde{\boldsymbol{u}}$ and reach the desired formulation:

$$\begin{gathered}\underset{\boldsymbol{w}\in\mathbb{R}^d, \boldsymbol{u}\in\mathbb{R}^d}{arg\min} \quad ||\boldsymbol{w}||^2 + ||\boldsymbol{u}||^2 \\ \forall \boldsymbol{x}_+ \in \mathbb{N}_+ : \boldsymbol{w} \cdot \boldsymbol{x}_+ - \alpha \boldsymbol{u} \cdot \boldsymbol{x}_+ \geq 1 \\ \forall \boldsymbol{x}_- \in \mathbb{N}_- : \boldsymbol{u} \cdot \boldsymbol{x}_- - \alpha \boldsymbol{w} \cdot \boldsymbol{x}_- \geq 1\end{gathered}$$

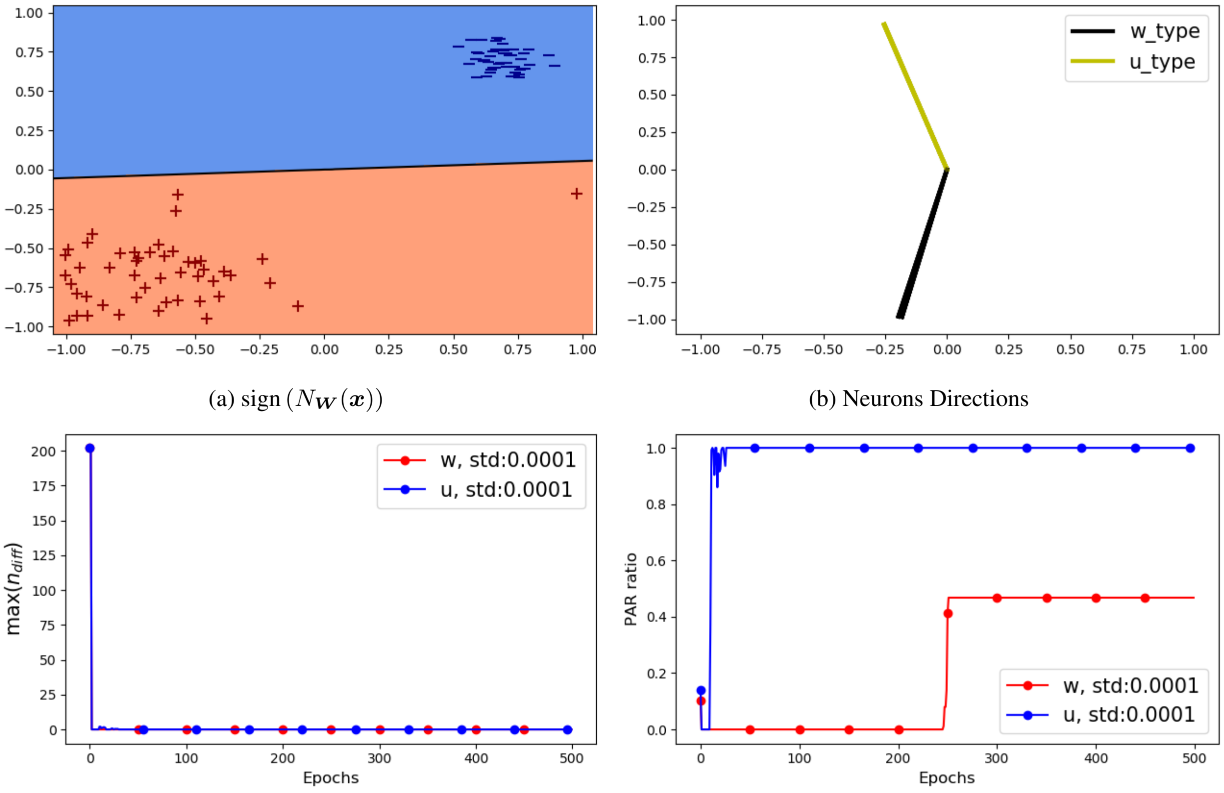
We obtained a reformulation of (P) as an SVM problem with variables $(\boldsymbol{w}, \boldsymbol{u}) \in \mathbb{R}^{2d}$ and with a transformed dataset which is a concatenated version of the original data $\phi(\boldsymbol{x}) = [\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x})\boldsymbol{x}, -\sigma'(-\boldsymbol{w}^* \cdot \boldsymbol{x})\boldsymbol{x}] \in \mathbb{R}^{2d}$, where for $\boldsymbol{x}_+ \in \mathbb{N}_+$, $\phi(\boldsymbol{x}_+) = (\boldsymbol{x}_+, -\alpha\boldsymbol{x}_+) \in \mathbb{R}^{2d}$ and for $\boldsymbol{x}_- \in \mathbb{N}_-$, $\phi(\boldsymbol{x}_-) = (-\alpha\boldsymbol{x}_-, \boldsymbol{x}_-) \in \mathbb{R}^{2d}$.

8. Proof of Lemma 6.1

Assume $\mathbb{V}_\beta^+(\mathbb{S}) \neq \emptyset$, i.e. $\exists \boldsymbol{v} \in \mathbb{S}$, s.t. $\forall \boldsymbol{x} \in \mathbb{S}_+ \hat{\boldsymbol{v}} \cdot \boldsymbol{x} \geq \beta$ and $\exists \boldsymbol{x}_* \in \mathbb{S}_-$ s.t. $\hat{\boldsymbol{v}} \cdot \boldsymbol{x}_* \geq \beta$. This means that $\hat{\boldsymbol{v}} \cdot -\boldsymbol{x}_* \leq -\beta$, because the data is linearly separable $-\boldsymbol{x}_* \in \mathbb{S}$ has to be a positive point and by the definition of $\mathbb{V}_\beta^+(\mathbb{S})$ that would mean $\hat{\boldsymbol{v}} \cdot -\boldsymbol{x}_* \geq \beta$ in contradiction.

By symmetry, if we assume $\mathbb{V}_\beta^-(\mathbb{S}) \neq \emptyset$ by taking the positive point which $\hat{v} \in \mathbb{V}_\beta^-(\mathbb{S})$ mistakenly classifies as a negative one, we'll reach a contradiction again.

Therefore if $\forall \boldsymbol{x} \in \mathbb{S}, -\boldsymbol{x} \in \mathbb{S}$ we have $\mathbb{V}_\beta^+(\mathbb{S}) = \emptyset$ and $\mathbb{V}_\beta^-(\mathbb{S}) = \emptyset$ and Assumption 4 in Theorem 6.2. holds in this case.



(a) sign $(N_{\boldsymbol{W}}(\boldsymbol{x}))$

(b) Neurons Directions

(c) Entrance to NAR

(d) Non Entrance to PAR

Figure 5: The entrance to the NAR of a 100 neurons network.The weights initialization std is 10^{-4}, learning rate is $\eta = 10^{-2}$. Each line in (c) and (d) is averaged over 5 initializations.

9. Entrance to PAR - High Dimensional Gaussians

We will show that the entrance to the PAR indeed happens empirically for two separable Gaussians. We measure the percentage of neurons which are in the PAR of both types. A $\boldsymbol{w}$ type neuron is considered in the PAR if it classifies like the ground truth $\boldsymbol{w}^*$. A $\boldsymbol{u}$ type neuron is considered in the PAR if it classifies like $-\boldsymbol{w}^*$.

The percentage of neurons in the PAR throughout the training process is given in Figure 4. We can see that the network enters the PAR.

10. Entrance to NAR which is not a PAR

In this section we show that learning can enter an NAR which is not a PAR. We sample two antipodal Gaussians and add one outlier positive point. Then for each neuron type ($\boldsymbol{w}$ or $\boldsymbol{u}$) we measure the maximum amount of data points classification disagreements between neurons of the same type denoted $\max(n_{diff})$ and the percentage of neurons which are in the PAR.

In Figure 5a we can see that the network yields 100% prediction accuracy. In Figure 5b we can see the directions of the neurons ($\boldsymbol{w}$ type in black and $\boldsymbol{u}$ type in yellow). In Figure 5c we can see that the maximal number of points which neurons of the same type classified differently goes to zero, therefore all neurons of the same type agree on the classification of the data points. In Figure 5d we can see that the ratio of $\boldsymbol{w}$ type neurons which perfectly classifies the data does not increase to 1 so the network does not enter the PAR.

11. Extension - First Layer Bias Term

In order to extend our results to include a bias term in the first layer, we would just need to reformulate our data points $\mathbb{S}$ to $\mathbb{S}'$ by

$$(\boldsymbol{x}, y) \in \mathbb{S} \subseteq \mathbb{R}^d \times \mathbb{Y} \mapsto ((\boldsymbol{x}, 1), y) \in \mathbb{S}' \subseteq \mathbb{R}^{d+1} \times \mathbb{Y}$$

and extend our neurons to include a bias term:

$$\forall \quad 1 \leq i \leq k \quad \boldsymbol{w}_t^{(i)} \in \mathbb{R}^d \mapsto (\boldsymbol{w}_t^{(i)}, b_w^{(i)}) \in \mathbb{R}^{d+1}, \quad \boldsymbol{u}_t^{(i)} \in \mathbb{R}^d \mapsto (\boldsymbol{u}_t^{(i)}, b_u^{(i)}) \in \mathbb{R}^{d+1}$$

This is equivalent to reformulating the first weights matrix $\boldsymbol{W} \in \mathbb{R}^{2k \times d} \mapsto \boldsymbol{W}' \in \mathbb{R}^{2k \times (d+1)}$.

This reformulation is equivalent to adding a bias term for every neuron in the first layer, and all of the following results would still hold under the above reformulation.

The proofs of Theorem 4.1 and Theorem 5.1 follow exactly if we exchange $\boldsymbol{W}$ with $\boldsymbol{W}'$ while for the proofs of Theorem 6.1 and Theorem 6.2 we use results from (Lyu & Li, 2020) and (Ji & Telgarsky, 2020) that require the model to be homogeneous. Note that if we add a bias in the *first* layer, the model remains homogeneous and the proofs of Theorem 6.1 and Theorem 6.2 still hold for those cases as well.

References

Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. Stochastic subgradient method converges on tame functions, 2018.

Ji, Z. and Telgarsky, M. Directional convergence and alignment in deep learning, 2020.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. *ICLR*, 2020.