A. Proofs of Basin Partition Properties

A.1. Proof of Lemma 1

We will need the following three auxiliary lemmas.

Lemma 3. *Let B be some basin as defined in Definition 2, and define $\mathbf{z}_i = v_i \mathbf{w}_i$. Then*

$$L_S(Z) = L_S(W, \mathbf{v})$$

is convex in $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ on B.

Proof. Restricting ourselves to B, since $\operatorname{sign}(\langle \mathbf{w}_j, \mathbf{x}_t \rangle)$, $\operatorname{sign}(v_j)$ are fixed, we can rewrite our objective as

$$\frac{1}{m} \sum_{t=1}^{m} \ell \left(\sum_{i=1}^{n} \sigma_{i,t} \langle v_i \mathbf{w}_i, \mathbf{x}_t \rangle, y_t \right) = \frac{1}{m} \sum_{t=1}^{m} \ell \left(\sum_{i=1}^{n} \sigma_{i,t} \langle \mathbf{z}_i, \mathbf{x}_t \rangle, y_t \right),$$

where $\sigma_{i,t} \in \{-1, 0, +1\}$ are fixed. This is a linear function composed with a convex loss ℓ, therefore the objective is convex in Z. □

Lemma 4. *Let $(W, \mathbf{v}) \in B_S^{A,\mathbf{b}}$. There exists a continuous path $\left(\tilde{W}^{(\lambda)}, \tilde{\mathbf{v}}^{(\lambda)}\right)$, $\lambda \in [0,1]$ from the initial point $\left(\tilde{W}^{(0)}, \tilde{\mathbf{v}}^{(0)}\right) = (W, \mathbf{v})$, to a point $\left(\tilde{W}^{(1)}, \tilde{\mathbf{v}}^{(1)}\right)$ satisfying $\tilde{\mathbf{v}}^{(1)} = \mathbf{b}$, along which $N_n\left(\tilde{W}^{(\lambda)}, \tilde{\mathbf{v}}^{(\lambda)}\right)$ is constant and $\left(\tilde{W}^{(\lambda)}, \tilde{\mathbf{v}}^{(\lambda)}\right) \in B_S^{A,\mathbf{b}}$ $\forall \lambda \in [0,1]$.*

Proof. If $v_i = 0$ for some $i \in [n]$, then the i^{th} neuron is canceled and we can linearly rescale $\mathbf{w}_i$ to $\mathbf{0}$, and then rescale v_i to b_i, so we may assume without loss of generality that $v_i \neq 0$ for all $i \in [n]$. We have for all $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \succ \mathbf{0}$,

$$\begin{aligned} N_n(W, \mathbf{v})(\mathbf{x}) &= \sum_{i=1}^{n} v_i \left[\langle \mathbf{w}_i, \mathbf{x} \rangle \right]_+ \\ &= \sum_{i=1}^{n} \frac{v_i}{\alpha_i} \alpha_i \left[\langle \mathbf{w}_i, \mathbf{x} \rangle \right]_+ \\ &= \sum_{i=1}^{n} \frac{v_i}{\alpha_i} \left[\langle \alpha_i \mathbf{w}_i, \mathbf{x} \rangle \right]_+ . \end{aligned}$$

Where we used the positive homogeneity of $[\cdot]_+$ in the last equality. So by linearly scaling $\boldsymbol{\alpha}^{(0)} = (1, \ldots, 1)$ to $\boldsymbol{\alpha}^{(1)} = (|v_1|, \ldots, |v_n|)$, i.e. $\boldsymbol{\alpha}^{(\lambda)} = (1 - \lambda + \lambda |v_1|, \ldots, 1 - \lambda + \lambda |v_n|)$, $\lambda \in [0,1]$, we obtain the desired path

$$\tilde{W}^{(\lambda)} = \left(\alpha_1^{(\lambda)} \mathbf{w}_1, \ldots, \alpha_n^{(\lambda)} \mathbf{w}_n, \right),$$

$$\tilde{\mathbf{v}}^{(\lambda)} = \left(\frac{v_1}{\alpha_1^{(\lambda)}}, \ldots, \frac{v_n}{\alpha_n^{(\lambda)}} \right),$$

while noting that $\operatorname{sign}(v_i) = \operatorname{sign}\left(\frac{v_i}{\alpha_i}\right)$ and $\operatorname{sign}(\langle \mathbf{w}_i, \mathbf{x} \rangle) = \operatorname{sign}(\langle \alpha_i \mathbf{w}_i, \mathbf{x} \rangle)$ for all $\boldsymbol{\alpha} \succ 0$, therefore we remain inside $B_S^{A,\mathbf{b}}$. □

Lemma 5. *For $(W, \mathbf{v})$, $\left(\tilde{W}, \tilde{\mathbf{v}}\right) \in B_S^{A,\mathbf{b}}$, define*

$$v_i^{(\lambda)} = \lambda \tilde{v}_i + (1 - \lambda) v_i,$$

$$\mathbf{w}_i^{(\lambda)} = \begin{cases} \lambda \tilde{\mathbf{w}}_i + (1 - \lambda) \mathbf{w}_i & v_i = \tilde{v}_i = 0 \\ \lambda \frac{\tilde{v}_i \tilde{\mathbf{w}}_i}{v_i^{(\lambda)}} + (1 - \lambda) \frac{v_i \mathbf{w}_i}{v_i^{(\lambda)}} & \textit{otherwise} \end{cases}.$$

Then

1. $v_i^{(\lambda)}\mathbf{w}_i^{(\lambda)} = \lambda\tilde{v}_i\tilde{\mathbf{w}}_i + (1-\lambda)\, v_i\mathbf{w}_i \;\; \forall i \in [n]\,, \lambda \in (0,1)\,.$
2. $\left(\mathbf{w}_i^{(\lambda)}, v_i^{(\lambda)}\right) \underset{\lambda\to 0_+}{\longrightarrow} (\mathbf{w}_i, v_i) \;\; \forall i \in [n]\,.$
3. $\left(\mathbf{w}_1^{(\lambda)}, \ldots, \mathbf{w}_n^{(\lambda)}, v_1^{(\lambda)}, \ldots, v_n^{(\lambda)}\right) \in B_S^{A,\mathbf{b}} \;\; \forall \lambda \in (0,1)\,.$

Proof.

1. Can be shown using a straightforward computation.
2. Compute
$$\lim_{\lambda\to 0_+} v_i^{(\lambda)} = \lim_{\lambda\to 0_+} \lambda\tilde{v}_i + (1-\lambda)\, v_i = v_i.$$
Suppose $v_i = \tilde{v}_i = 0$, then
$$\lim_{\lambda\to 0_+} \mathbf{w}_i^{(\lambda)} = \lim_{\lambda\to 0_+} \lambda\tilde{\mathbf{w}}_i + (1-\lambda)\, \mathbf{w}_i = \mathbf{w}_i.$$
Otherwise, $v_i^{(\lambda)} \neq 0 \;\; \forall \lambda \in (0,1)$ since $\operatorname{sign}(v_i) = \operatorname{sign}(\tilde{v}_i)$, and we have
$$\begin{aligned}\lim_{\lambda\to 0_+} \mathbf{w}_i^{(\lambda)} &= \lim_{\lambda\to 0_+} \left(\lambda\frac{\tilde{v}_i\tilde{\mathbf{w}}_i}{v_i^{(\lambda)}} + (1-\lambda)\,\frac{v_i\mathbf{w}_i}{v_i^{(\lambda)}}\right)\\ &= \lim_{\lambda\to 0_+} \frac{\lambda\tilde{v}_i\tilde{\mathbf{w}}_i}{\lambda\tilde{v}_i + (1-\lambda)\, v_i} + \lim_{t\to 0_+} \frac{(1-\lambda)\, v_i\mathbf{w}_i}{\lambda\tilde{v}_i + (1-\lambda)\, v_i}\\ &= 0 + \frac{v_i\mathbf{w}_i}{v_i}\\ &= \mathbf{w}_i.\end{aligned}$$
3. Since $\operatorname{sign}(\tilde{v}_i) = \operatorname{sign}(v_i)$, we have
$$\begin{aligned}\operatorname{sign}\left(v_i^{(\lambda)}\right) &= \operatorname{sign}(\lambda\tilde{v}_i + (1-\lambda)\, v_i)\\ &= \lambda\operatorname{sign}(\tilde{v}_i) + (1-\lambda)\operatorname{sign}(v_i)\,.\end{aligned}$$
Suppose $v_i = \tilde{v}_i = 0$, then since $\operatorname{sign}(\langle\tilde{\mathbf{w}}_i, \mathbf{x}_t\rangle) = \operatorname{sign}(\langle\mathbf{w}_i, \mathbf{x}_t\rangle)$, we have $\forall t \in [m]\,, i \in [n]\,, \lambda \in (0,1)$
$$\begin{aligned}\operatorname{sign}\left(\left\langle\mathbf{w}_i^{(\lambda)}, \mathbf{x}_t\right\rangle\right) &= \operatorname{sign}(\langle\lambda\tilde{\mathbf{w}}_i + (1-\lambda)\, \mathbf{w}_i, \mathbf{x}_t\rangle)\\ &= \operatorname{sign}(\lambda\,\langle\tilde{\mathbf{w}}_i, \mathbf{x}_t\rangle + (1-\lambda)\,\langle\mathbf{w}_i, \mathbf{x}_t\rangle)\\ &= \lambda\operatorname{sign}(\langle\tilde{\mathbf{w}}_i, \mathbf{x}_t\rangle) + (1-\lambda)\operatorname{sign}(\langle\mathbf{w}_i, \mathbf{x}_t\rangle)\,.\end{aligned}$$
Otherwise,
$$\begin{aligned}\operatorname{sign}\left(\left\langle\mathbf{w}_i^{(\lambda)}, \mathbf{x}_t\right\rangle\right) &= \operatorname{sign}\left(\left\langle\lambda\frac{\tilde{v}_i\tilde{\mathbf{w}}_i}{v_i^{(\lambda)}} + (1-\lambda)\,\frac{v_i\mathbf{w}_i}{v_i^{(\lambda)}}, \mathbf{x}_t\right\rangle\right)\\ &= \operatorname{sign}\left(\frac{\tilde{v}_i\lambda}{v_i^{(\lambda)}}\,\langle\tilde{\mathbf{w}}_i, \mathbf{x}_t\rangle + \frac{v_i\cdot(1-\lambda)}{v_i^{(\lambda)}}\,\langle\mathbf{w}_i, \mathbf{x}_t\rangle\right)\\ &= \operatorname{sign}(\langle\mathbf{w}_i, \mathbf{x}_t\rangle)\,.\end{aligned}$$

□

We are now ready to prove Lemma 1.

Proof (of Lemma 1).
Clearly, $B_S^{A,\mathbf{b}}$ is a closed set, and is convex as an intersection of halfspaces.

- $B_S^{A,\mathbf{b}}$ has connected sublevel sets:
  Let $(W,\mathbf{v}),(W',\mathbf{v}') \in B_{\leq\alpha}$. Using Lemma 4 we may assume without loss of generality that $\mathbf{v}=\mathbf{v}' \in \{-1,+1\}^n$. By linearly interpolating W, W', i.e. by taking

  $$W^{(\lambda)} = (1-\lambda)\, W + \lambda W',\ \lambda \in [0,1]\,,$$

  we get a continuous path connecting W, W'. This path remains in the same basin as a result of Lemma 5.3. Moreover, by Lemma 3, the objective is convex in W, so we get for all $\lambda \in [0,1]$

  $$\begin{aligned} E_S\left(W^{(\lambda)},\mathbf{v}\right) &\leq (1-\lambda)\, E_S\left(W,\mathbf{v}\right) + \lambda E_S\left(W',\mathbf{v}\right) \\ &\leq (1-\lambda)\,\alpha + \lambda\alpha \\ &= \alpha. \end{aligned}$$

- Any local minimum in $B_S^{A,\mathbf{b}}$ is global:
  Suppose $(W,\mathbf{v}) = (\mathbf{w}_1,\ldots,\mathbf{w}_n,v_1\ldots,v_n)$ is a local minimum in $B_S^{A,\mathbf{b}}$, let

  $$\left(\tilde{W},\tilde{\mathbf{v}}\right) = (\tilde{\mathbf{w}}_1,\ldots,\tilde{\mathbf{w}}_n,\tilde{v}_1\ldots,\tilde{v}_n) \in B_S^{A,\mathbf{b}}$$

  be arbitrary, and denote

  $$\left(W^{(\lambda)},\mathbf{v}^{(\lambda)}\right) \coloneqq \left(\mathbf{w}_1^{(\lambda)},\ldots,\mathbf{w}_n^{(\lambda)},v_1^{(\lambda)},\ldots,v_n^{(\lambda)}\right).$$

  Then for small enough λ

  $$\begin{aligned} L_S\left(W,\mathbf{v}\right) &\leq L_S\left(W^{(\lambda)},\mathbf{v}^{(\lambda)}\right) \\ &= L_S\left(W^{(\lambda)}\cdot\mathbf{v}^{(\lambda)}\right) \\ &= L_S\left(\lambda\left(\tilde{v}_1\tilde{\mathbf{w}}_1,\ldots,\tilde{v}_n\tilde{\mathbf{w}}_n\right)+(1-\lambda)\left(v_1\mathbf{w}_1,\ldots,v_n\mathbf{w}_n\right)\right) \\ &\leq \lambda L_S\left(\tilde{v}_1\tilde{\mathbf{w}}_1,\ldots,\tilde{v}_n\tilde{\mathbf{w}}_n\right)+(1-\lambda)\, L_S\left(v_1\mathbf{w}_1,\ldots,v_n\mathbf{w}_n\right) \\ &= \lambda L_S\left(\tilde{W},\tilde{\mathbf{v}}\right)+(1-\lambda)\, L_S\left(W,\mathbf{v}\right), \end{aligned}$$

  $$\implies L_S\left(W,\mathbf{v}\right) \leq L_S\left(\tilde{W},\tilde{\mathbf{v}}\right).$$

  Where the first transition comes from $(W,\mathbf{v})$ being a local minimum and Lemma 5.2,5.3, the second and third from Lemma 5.1, and the fourth from Lemma 3.

□

A.2. Proof of Lemma 2

Let $(W^*,\mathbf{v}^*) = (\mathbf{w}_1^*,\ldots,\mathbf{w}_k^*,v_1^*,\ldots,v_k^*) \in \mathbb{R}^{kd+k}$ satisfy

$$\text{Bas}\left(\mathbf{w}_{i_1},\ldots,\mathbf{w}_{i_k},v_{i_1},\ldots,v_{i_k}\right) = L_S\left(N_k\left(W^*,\mathbf{v}^*\right)\right),$$

and let

$$(W',\mathbf{v}') = (\mathbf{w}_1',\ldots,\mathbf{w}_n',v_1',\ldots,v_n') \in \mathbb{R}^{nd+n},$$

where

$$(\mathbf{w}_i',v_i') = \begin{cases} 0 & i \notin I \\ \left(\mathbf{w}_j^*,v_j^*\right) & i = i_j \end{cases}.$$

Then

$$\begin{aligned}\text{Bas}\left(W,\mathbf{v}\right) &\leq L_S\left(N_n\left(W',\mathbf{v}'\right)\right)\\ &= L_S\left(N_n\left(\mathbf{w}'_1,\ldots,\mathbf{w}'_n,v'_1,\ldots,v'_n\right)\right)\\ &= L_S\left(N_k\left(\mathbf{w}^*_1,\ldots,\mathbf{w}^*_k,v^*_1,\ldots,v^*_k\right)\right)\\ &= L_S\left(N_k\left(W^*,\mathbf{v}^*\right)\right)\\ &= \text{Bas}\left(\mathbf{w}_{i_1},\ldots,\mathbf{w}_{i_k},v_{i_1},\ldots,v_{i_k}\right).\end{aligned}$$

Where the first inequality comes from $(W,\mathbf{v}),(W',\mathbf{v}')$ belonging to the same basin, and the second equality comes from both weights computing the same network output for any input $\mathbf{x}\in\mathbb{R}^d$.

B. Proofs of Main Theorems

B.1. Proof of Thm. 1

Before delving into the proof of the theorem, we provide some intuition in the special case of the squared loss, where $L(P(\mathcal{W}))=\frac{1}{m}\sum_{t=1}^{m}(N(\mathcal{W})(\mathbf{x}_t)-y_t)^2$. Fix some $\lambda\in[0,1]$, and consider the objective function along the ray in the parameter space, corresponding to multiplying the last layer weights in $\mathcal{W}^{(\lambda)}$ by some scalar $c\geq 0$. Since the output layer is linear, the objective function (as we vary c) will have the form

$$\frac{1}{m}\sum_{t=1}^{m}(c\cdot N(\mathcal{W}^{(\lambda)})(\mathbf{x}_t)-y_t)^2.$$

Thus, the objective function, as a parameter of c (where $\mathcal{W}^{(\lambda)}$ is fixed) is just a quadratic function. Moreover, by the intermediate value theorem, as long as $N(\mathcal{W}^{(\lambda)})(\mathbf{x}_t)$ is not 0 for all t, then by picking different values of c, we can find points along the ray taking any value between $\frac{1}{m}\sum_{i=1}^{t}y_t^2$ (when $c=0$) and ∞ (as $c\to\infty$). Therefore, as long as we start from a point whose objective value is larger than $\frac{1}{m}\sum_{i=1}^{t}y_t^2$, we can re-scale each $\mathcal{W}^{(\lambda)}$ by some factor c_γ, so that the new path is continuous, as well as monotonically decreasing in value, remaining above $\frac{1}{m}\sum_{i=1}^{t}y_t^2$. When we reach the ray belonging to the endpoint $\mathcal{W}^{(1)}$ of the original path, we simply re-scale back towards $\mathcal{W}^{(1)}$, with the objective function continuing to decrease due to the convex quadratic form of the objective function along the ray.

We now turn to the formal proof in the general setting of Thm. 1. For technical reasons, we will extend the interval $\lambda\in[0,1]$ to a strictly larger interval, and define certain quantities with respect to that larger interval. Specifically, for any $\lambda\in[-1,2]$, define

$$v^{(\lambda)}=\begin{cases}L(P(\mathcal{W}^{(0)}))-\frac{\lambda}{2}\epsilon & \lambda\in[-1,0]\\ \left(1-\frac{\lambda}{3}\right)\cdot L(P(\mathcal{W}^{(0)}))+\frac{\lambda}{3}\cdot\max\{L(\mathbf{0}),L(P(\mathcal{W}^{(1)}))\} & \lambda\in[0,2].\end{cases}$$

and note that it strictly monotonically decreases with λ, and satisfies the chain of inequalities

$$L(P(\mathcal{W}^{(0)}))+\epsilon>v^{(-1)}>v^{(0)}=L(P(\mathcal{W}^{(0)}))>v^{(2)}>\max\{L(\mathbf{0}),L(P(\mathcal{W}^{(1)}))\}.$$

By assumption, for any $\lambda\in[0,1]$, there exists some $c^{(\lambda)}$ such that $L(c^{(\lambda)}\cdot P(\mathcal{W}^{(\lambda)}))\geq L(P(\mathcal{W}^{(0)}))+\epsilon$. Since $L(P(\mathcal{W}^{(0)}))+\epsilon>v^{(\lambda)}$ for any $\lambda\in[-1,2]$, it follows that for any such λ,

$$L(c^{\text{clip}(\lambda)}\cdot P(\mathcal{W}^{\text{clip}(\lambda)}))>v^{(\lambda)},\tag{2}$$

where $\text{clip}(\lambda)=\min\{1,\max\{0,\lambda\}\}$ denote clipping of λ to the interval $[0,1]$. On the other hand, for any $\lambda\in[-1,2]$,

$$L(0\cdot P(\mathcal{W}^{\text{clip}(\lambda)}))=L(\mathbf{0})<v^{(\lambda)}.\tag{3}$$

Since L is convex and real-valued, it is continuous, hence $L(c\cdot P(\mathcal{W}^{\text{clip}(\lambda)}))$ is convex and continuous in c. Combining this with Eq. (2) and Eq. (3), it follows from the intermediate value theorem that

$$\forall\lambda\in[-1,2],\ \ \exists\,\tilde{c}^{(\lambda)}\in(0,c^{\text{clip}(\lambda)})\ \text{ such that }\ L(\tilde{c}^{(\lambda)}\cdot P(\mathcal{W}^{\text{clip}(\lambda)})=v^{(\lambda)}.\tag{4}$$

Moreover, this $\tilde{c}^{(\lambda)}$ is unique: To see why, consider the convex function $f(c) = L(c \cdot P(\mathcal{W}^{\text{clip}(\lambda)})$, and assume there are two distinct values c', c such that $c^{\text{clip}(\lambda)} > c' > c > 0$, and $f(c) = f(c') = v^{(\lambda)}$. Then by Eq. (2) and Eq. (3), we would have the chain of inequalities

$$f(c^{\text{clip}(\lambda)}) > f(c') = f(c) > f(0)$$

which cannot be satisfied by a convex function f.

We now make the following series of observations on $\tilde{c}^{(\lambda)}$, which establish their continuity in λ and that $\tilde{c}^{(0)} = 1$:

1. *For any $\lambda \in (-1, 2)$, there is some open neighborhood of $\tilde{c}^{(\lambda)}$ in which the univariate function $c \mapsto L(c \cdot P(\mathcal{W}^{clip(\lambda)}))$ is one-to-one*: As discussed above, $c \mapsto L(c \cdot P(\mathcal{W}^{\text{clip}(\lambda)}))$ is convex, and does not attain a minimal value at $\tilde{c}^{(\lambda)}$ (since $L(\tilde{c}^{(\lambda)} \cdot P(\mathcal{W}^{\text{clip}(\lambda)})) = v^{(\lambda)} > L(0 \cdot P(\mathcal{W}^{\text{clip}(\lambda)}))$). Therefore, $L(c \cdot P(\mathcal{W}^{\text{clip}(\lambda)}))$ must be strictly increasing or decreasing in some open neighborhood of $\tilde{c}^{(\lambda)}$, and therefore it is locally one-to-one.

2. *$\tilde{c}^{(\lambda)}$ is continuous in $\lambda \in (-1, 2)$*: Consider the function $f(c, \lambda) = L(c \cdot P(\mathcal{W}^{\text{clip}(\lambda)})) - v^{(\lambda)}$, where $(c, \lambda) \in (0, \infty) \times (-1, 2)$. By definition of $\tilde{c}^{(\lambda)}$, we have $f(\tilde{c}^{(\lambda)}, \lambda) = 0$ for all λ. Moreover, f is continuous (since $P(\mathcal{W}^{\text{clip}(\lambda)}), v^{(\lambda)}$ are continuous in λ, and L is convex and hence continuous), and by observation 1 above, $f(\cdot, \lambda)$ is locally one-to-one for any $\lambda \in (-1, 2)$. Applying a version of the implicit function theorem from multivariate calculus for possibly non-differentiable functions (see (Kumagai, 1980)), it follows that there exists some unique and continuous mapping of λ to c in an open neighborhood of $(\tilde{c}^{(\lambda)}, \lambda)$, for which $f(c, \lambda) = 0$. But as discussed earlier, for a given λ, $c = \tilde{c}^{(\lambda)}$ is the unique value for which $f(c, \lambda) = L(c \cdot P(\mathcal{W}^{\text{clip}(\lambda)})) - v^{(\lambda)} = 0$, so this continuous mapping must map λ to $\tilde{c}^{(\lambda)}$ locally at every λ. Since this holds for any λ, the mapping of λ to $\tilde{c}^{(\lambda)}$ is continuous on $\lambda \in (-1, 2)$.

3. $\tilde{c}_0 = 1$: By definition of $\tilde{c}^{(\lambda)}$ and $v^{(\lambda)}$ at $\lambda = 0$, $L(\tilde{c}^{(0)} \cdot P(\mathcal{W}^{(0)})) = v^{(0)} = L(P(\mathcal{W}^{(0)}))$, which is clearly satisfied for $\tilde{c}^{(0)} = 1$, and as discussed earlier, is not satisfied for any other value.

Based on the above observations, we have that $\tilde{c}^{(\lambda)}$, as a function of $\lambda \in [0, 1]$, is continuous, begins at $\tilde{c}_0 = 1$, and satisfies $L(\tilde{c}^{(\lambda)} \cdot P(\mathcal{W}^{(\lambda)})) = v^{(\lambda)}$. Moreover, $v^{(\lambda)}$ is strictly decreasing in λ. Therefore, letting

$$\{\tilde{\mathcal{W}}^{(\lambda)}, \lambda \in [0, 1]\} \tag{5}$$

denote the path in the parameter space, where each $\tilde{\mathcal{W}}^{(\lambda)}$ equals $\mathcal{W}^{(\lambda)}$ with the last layer weights re-scaled by $\tilde{c}^{(\lambda)}$, we have that Eq. (5) indeed defines a continuous path from the initialization point $\mathcal{W}^{(0)}$ in the parameter space, along which the loss $L(P(\tilde{\mathcal{W}}^{(\lambda)})$ is strictly monotonically decreasing.

All that remains now is to argue that from $\tilde{\mathcal{W}}^{(1)}$, we have a strictly monotonically decreasing path to a point whose loss equals $L(P(\mathcal{W}^{(1)}))$. To see this, note that by definition of $\tilde{\mathcal{W}}^{(1)}$ and $v^{(1)}$, we have $L(P(\tilde{\mathcal{W}}^{(1)})) = v^{(1)} > L(P(\mathcal{W}^{(1)}))$. Therefore,

$$L(c \cdot P(\mathcal{W}^{(1)}))$$

is convex in c, equals $L(P(\tilde{\mathcal{W}}^{(1)}))$ at $c = \tilde{c}^{(\lambda)}$, and equals the strictly smaller value $L(P(\mathcal{W}^{(1)}))$ at $c = 1$. Therefore, by re-scaling the last layer parameters of $\tilde{\mathcal{W}}^{(1)}$ to match those of $\mathcal{W}^{(1)}$, we are guaranteed to strictly and monotonically decrease the loss, until we get a loss equal to $L(P(\mathcal{W}^{(1)}))$. Concatenating this with the continuous path $\tilde{\mathcal{W}}^{(\lambda)}, \lambda \in [0, 1]$, the result follows.

B.2. Proof of Proposition 1

It is enough to verify that for both losses, proposition 2 holds with $r = 1$.

For the squared loss, if $P(\mathcal{W}^{(0)}) \neq \mathbf{0}$, then consider the first training example $(\mathbf{x}_i, y_i)$ for which $P(\mathcal{W}^{(0)})(\mathbf{x}_i) \neq \mathbf{0}$. In that case, it is easily verified that $(c \cdot P(\mathcal{W}^{(0)})(\mathbf{x}_i) - y_i)^2$ is strictly convex in c (for any c), and therefore $L(c \cdot P(\mathcal{W}^{(0)})) = \frac{1}{m}\sum_{i=1}^{m}(c \cdot P(\mathcal{W}^{(0)})(\mathbf{x}_i) - y_i)^2$ is also strictly convex, as an average of convex functions where at least one of them is strictly convex. Therefore, strict convexity holds with probability $r = 1$.

For the cross-entropy loss, it is enough to consider the first training example on which the prediction vector $\mathbf{p}$ of $P(\mathcal{W}^{(\lambda)})$ is non-zero, and show strict convexity on that example with probability 1. Since the loss on other examples are convex as well, we get overall strict convexity with probability 1 as required. Specifically, we need to show strict convexity in c of

the function

$$-\log\left(\frac{\exp(c\cdot p_{j_i})}{\sum_j \exp(c\cdot p_j)}\right)=\log\left(\sum_j \exp(c\cdot p_j)\right)-c\cdot p_{j_i}. \tag{6}$$

where j_i is the correct class. To do so, consider the function $f(\mathbf{p})=\log(\sum_j \exp(p_j))$. A straightforward calculation reveals that its Hessian equals

$$\nabla^2 f(\mathbf{p})=\mathrm{diag}(\mathbf{q})-\mathbf{q}\mathbf{q}^\top \text{ where } \mathbf{q}=\frac{1}{\sum_j \exp(p_j)}\left(\exp(p_1),\exp(p_2),\ldots,\exp(p_k)\right),$$

so the second derivative of the function in Eq. (6) w.r.t. c at some value c equals

$$\mathbf{p}^\top\nabla^2 f(c\cdot\mathbf{p})\mathbf{p}. \tag{7}$$

We now argue that this is strictly positive, unless $\mathbf{p}$ is a constant vector $p_1=p_2=\ldots=p_k$, in which case the function in Eq. (6) is indeed strictly convex. To see this, note that the Hessian of f is a rank-1 perturbation of the $k\times k$ positive definite matrix $\mathrm{diag}(\mathbf{q})$, so its rank is at least $k-1$. Thus, there is only a 1-dimensional subspace of vectors $\mathbf{v}$, for which $\mathbf{v}^\top(\mathrm{diag}(\mathbf{q})-\mathbf{q}\mathbf{q}^\top)\mathbf{v}=0$, which can be verified to be exactly the subspace of constant vectors. Thus, Eq. (7) is positive unless $\mathbf{p}$ is a constant vector.

To finish the proof for the cross-entropy loss, it remains to show that the probability that $\mathbf{p}$ is non-constant (conditioned on $P(\mathcal{W}^{(0)})\neq\mathbf{0}$) is 1. To simplify the notation, let NZ be the event that $P(\mathcal{W}^{(0)})\neq\mathbf{0}$, let P be the event that conditioned on NZ, $\mathbf{p}$ (the first non-zero prediction vector over the training examples) is also non-constant. Also, let V be the event that conditioned on NZ, then for the same training example as $\mathbf{p}$, the input vector to the output neurons is non-zero. Then it holds that

$$\mathbb{P}\left[P|NZ\right]=\frac{\mathbb{P}\left[P|V,NZ\right]\mathbb{P}\left[V|NZ\right]}{\mathbb{P}\left[V|P,NZ\right]}. \tag{8}$$

$\mathbb{P}\left[P|V,NZ\right]=1$, since the linear output neurons are initialized independently from a spherically-symmetric distribution supported on non-zero vectors, so given a non-zero input, the probability that some neurons will output different values is 1. Also, $\mathbb{P}\left[V|P,NZ\right]=\mathbb{P}\left[V|NZ\right]=1$, since conditioned on NZ, $\mathbf{p}\neq\mathbf{0}$ by definition, and since the output neurons compute a homogeneous linear mapping, the input to these neurons must also be non-zero. Plugging these observations back into Eq. (8), we get that $\mathbb{P}\left[P|NZ\right]=1$ as required.

B.3. Proof of Proposition 2

We first prove that

$$\mathbb{P}\left[P(\mathcal{W}^{(0)})\neq\mathbf{0}\right]\geq 1-2^{-n_{h-1}}. \tag{9}$$

To see this, consider any neuron in the $(h-1)^{\text{th}}$ layer, computing $\mathbf{x}\mapsto[\langle\mathbf{w},\mathbf{x}\rangle+b]_+$. Since $(\mathbf{w},b)$ is drawn from a spherically symmetric distribution supported on non-zero vectors, it holds for any fixed $\mathbf{x}$ that $\mathbb{P}\left[\langle\mathbf{w},\mathbf{x}\rangle+b>0\right]=\mathbb{P}\left[\langle\mathbf{w},\mathbf{x}\rangle+b<0\right]=\frac{1}{2}$. Therefore, $\mathbb{P}\left[[\langle\mathbf{w},\mathbf{x}\rangle+b]_+\neq 0\right]=\frac{1}{2}$. Since the weights at each neuron are drawn independently, and there are n_{h-1} neurons in the $(h-1)^{\text{th}}$ layer, it follows that a linear output neuron receives a non-zero input with probability at least $1-2^{-n_{h-1}}$. If this event occurs, then the output of the output neuron will be non-zero with probability 1. Since this holds for any fixed network input, it holds in particular for (say) the first training example, in which case $P(\mathcal{W}^{(0)})$ will be non-zero with such a probability. Letting A be the event that $P(\mathcal{W}^{(0)})\neq\mathbf{0}$, as well as $L(c\cdot P(\mathcal{W}^{(0)}))$ being strictly convex in $c\in[-1,+1]$, we have by Eq. (9) and the assumption in the statement that $\mathbb{P}\left[A\right]\geq r\left(1-2^{-n_{h-1}}\right)$

Let W be the realization of the random variable $P(\mathcal{W}^{(0)})$. Since the output neurons are initialized from a spherically symmetric distribution, $\mathbb{P}\left[W\right]=\mathbb{P}\left[-W\right]$ for any W. Moreover, it is easy to verify that for any W, event A occurs for $P(\mathcal{W}^{(0)})=W$ if and only if it occurs for $P(\mathcal{W}^{(0)})=-W$. Therefore,

$$\mathbb{P}\left[W|A\right]=\frac{\mathbb{P}\left[W\right]\mathbb{P}\left[A|W\right]}{\mathbb{P}\left[A\right]}=\frac{\mathbb{P}\left[-W\right]\mathbb{P}\left[A|-W\right]}{\mathbb{P}\left[A\right]}=\mathbb{P}\left[-W|A\right].$$

In other words, conditioned on A, for any realization W, we are equally likely to get W or $-W$. Also, conditioned on A (which implies strict convexity), $\max\{L(W),L(-W)\}\geq\frac{L(W)+L(-W)}{2}>L\left(\frac{W+(-W)}{2}\right)=L(\mathbf{0})$. Therefore, by symmetry, $\mathbb{P}\left[L(P(\mathcal{W}^{(0)}))>L(\mathbf{0})\mid A\right]\geq\frac{1}{2}$. As a result, $\mathbb{P}\left[L(P(\mathcal{W}^{(0)}))>L(\mathbf{0})\right]\geq\frac{1}{2}\mathbb{P}\left[A\right]\geq\frac{r}{2}\left(1-2^{-n_{h-1}}\right)$.

B.4. Proof of Thm. 2

Denote for all $j \in [n]$,

$$S_j^+ = \{\mathbf{x} \in S : x_j > 0\}, S_j^- = \{\mathbf{x} \in S : x_j < 0\},$$

and observe the objective value on S_j^+ satisfies for all $j \in [d]$,

$$\begin{aligned} L_{S_j^+}(W, \mathbf{v}) &= \sum_{t:\mathbf{x}_t \in S_j^+} \ell\left(\sum_{i=1}^n v_i \left[\langle \mathbf{w}_i, \mathbf{x}_t \rangle\right]_+, y_t\right) \\ &= \sum_{t:\mathbf{x}_t \in S_j^+} \ell\left(x_{t,j} \sum_{i=1}^n v_i \left[w_{i,j}\right]_+, y_t\right). \end{aligned}$$

Similarly,

$$L_{S_j^-}(W, \mathbf{v}) = \sum_{t:\mathbf{x}_t \in S_j^-} \ell\left(-x_{t,j} \sum_{i=1}^n v_i \left[-w_{i,j}\right]_+, y_t\right).$$

Since ℓ is convex, $L_{S_j^+}, L_{S_j^-}$ are convex in $\sum_{i=1}^n v_i \left[w_{i,j}\right]_+, \sum_{i=1}^n v_i \left[-w_{i,j}\right]_+$, respectively, so their minimal values are well defined. It is clear that the minimum achievable using a single neuron is lower bounded by the minimum achievable using two-layer nets, α, which in turn is lower bounded by the average of all minimal objective values over the various $S_j^{\pm}$. It then suffices to show that we initialize from a basin achieving such value, which we denote as $\beta \leq \alpha$, with high probability. Moreover, since the objective value on $S_j^{\pm}$ is independent for each $S_j^{\pm}$, it is enough to minimize the objective on each $S_j^{\pm}$ separately.

Since our basins correspond to the partition of our search space to a fixed sign at each coordinate, we have that for the expression $\sum_{i=1}^n v_i \left[w_{i,j}\right]_+$ to take the optimal value p^* in our initialized basin, it suffices that $\operatorname{sign}(w_{i,j}) = 1$ and $\operatorname{sign}(v_i) = \operatorname{sign}(p^*)$ for some $i \in [n]$. Using an analogous argument for S_j^- we have,

- The probability of this condition not to hold for a single neuron is at most $\frac{3}{4}$.
- The probability of this condition not to hold for all neurons (since by Assumption 1 all neurons are independent) is at most $\left(\frac{3}{4}\right)^n$.
- By using the union bound, the probability that exists some $S_j^{\pm}$ such that no neuron can obtain the minimal value over it is at most

$$2d\left(\frac{3}{4}\right)^n.$$

We conclude that when initializing $(W, \mathbf{v})$ using a distribution satisfying Assumption 1 then

$$\mathbb{P}\left[\text{Bas}(W, \mathbf{v}) \leq \beta \leq \alpha\right] \geq 1 - 2d\left(\frac{3}{4}\right)^n.$$

B.5. Proof of Thm. 3

We will need the following two auxiliary lemmas:

Lemma 6. *$N_n(\mathbf{w}_1, \ldots, \mathbf{w}_n, \mathbf{v})(\mathbf{x})$ is $(|v_i| \cdot \|\mathbf{x}\|)$-Lipschitz in each $\mathbf{w}_i$.*

We leave this lemma without proof, and note that it is immediate from the definition of N_n.

The following lemma provides a lower bound on the probability that the i^{th} neuron is initialized from a region with a point of distance at most δ from $\mathbf{w}_i^*$.

Lemma 7. *Let $\delta > 0$, let $(W^*, \mathbf{v}^*)$ satisfying $\|\mathbf{w}_i^*\|_2 = 1,\ \forall i \in [n]$, and let $(W, \mathbf{v})$ be a point on an origin-centered sphere chosen uniformly at random. Then $\forall i \in [n]$*

$$\mathbb{P}\left[\exists \tilde{\mathbf{w}}_i : \|\tilde{\mathbf{w}}_i - \mathbf{w}_i^*\|_2 \leq \delta,\ \operatorname{sign}\left(\langle \tilde{\mathbf{w}}, \mathbf{x}_t\rangle\right) = \operatorname{sign}\left(\langle \mathbf{w}, \mathbf{x}_t\rangle\right)\ \forall t \in [m]\right] \geq \frac{1}{\pi\left(\mathit{rank}\left(X\right) - 1\right)}\left(\delta\sqrt{1 - \frac{\delta^2}{4}}\right)^{\mathit{rank}(X)-1}.$$

Before turning to prove the lemma, we first prove the following auxiliary claim.

Claim 1. *Let $\delta > 0$ and let $\mathbf{a} \in \mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ be a point on the d-dimensional unit sphere. Let $\mathbf{b}$ be a point chosen uniformly at random from $\mathbb{S}^{d-1}$. Then*

$$\mathbb{P}\left[\|\mathbf{a} - \mathbf{b}\|_2 \leq \delta\right] \geq \frac{1}{\pi\left(d-1\right)}\left(\delta\sqrt{1 - \frac{\delta^2}{4}}\right)^{d-1}.$$

This claim suffices for proving a weaker version of Thm. 3 where rank (X) is replaced with d. However, utilizing a simple observation on the structure of the basin partition allows us to prove Lemma 7 which strengthens the result.

Proof. For a point $\mathbf{a} \in \mathbb{S}^{d-1}$, let $\bar{\mathsf{S}}\left(\mathbf{a}, \theta\right) := \left\{\mathbf{b} \in \mathbb{S}^{d-1} : \langle \mathbf{a}, \mathbf{b}\rangle \geq \cos\theta\right\}$ be the closed hyperspherical cap of angle $\theta \in [0, \pi]$. Note that if $\mathbf{a}, \mathbf{b} \in \mathbb{S}^{d-1}$ form an angle of $\theta' \in [0, \theta]$ (i.e. $\mathbf{b} \in \bar{\mathsf{S}}\left(\mathbf{a}, \theta\right)$) then they form an isosceles triangle with apex angle θ' and equal sides of length 1, so the distance between $\mathbf{a}$ and $\mathbf{b}$ satisfies

$$\begin{aligned}\|\mathbf{a} - \mathbf{b}\| &= 2\sin\left(\frac{\theta'}{2}\right)\\ &\leq 2\sin\left(\frac{\theta}{2}\right).\end{aligned}$$

Taking $\delta := 2\sin\left(\frac{\theta}{2}\right)$ we have that $\theta = 2\arcsin\left(\frac{\delta}{2}\right)$, so in order for us to lower bound $\mathbb{P}\left[\|\mathbf{a} - \mathbf{b}\|_2 \leq \delta\right]$, we need to compute the surface area $\nu_{d-1}\left(\theta\right)$ of the hyperspherical cap of angle θ at point $\mathbf{a}$ (independent of $\mathbf{a}$), and normalize this quantity by the area of the hypersphere ω_{d-1}.

The surface area of a hyperspherical cap of radius θ is given by the formula: ((Li, 2011))

$$\nu_{d-1}\left(\theta\right) = \omega_{d-2}\int_0^\theta \left(\sin^{d-2}\xi d\xi\right), \tag{10}$$

where ω_{d-1} denotes the surface area of $\mathbb{S}^{d-1}$. Consider the function $f\left(\theta\right) = \int_0^\theta \left(\sin^{d-2}\xi d\xi\right) - \frac{1}{d-1}\sin^{d-1}\theta$. It is monotonically increasing in $[0, \pi]$ since

$$\begin{aligned}f'\left(\theta\right) &= \frac{\partial}{\partial\theta}\left(\int_0^\theta \left(\sin^{d-2}\xi d\xi\right) - \frac{1}{d-1}\sin^{d-1}\theta\right)\\ &= \sin^{d-2}\theta - \sin^{d-2}\theta\cdot\cos\theta\\ &= \sin^{d-2}\theta\cdot\left(1 - \cos\theta\right)\\ &\geq 0,\end{aligned}$$

where the last inequality holds for all $\theta \in [0, \pi]$. Since $f\left(0\right) = 0$ we have $\forall\theta \in [0, \pi]$ that $\int_0^\theta \left(\sin^{d-2}\xi d\xi\right) \geq \frac{1}{d-1}\sin^{d-1}\theta$.

We compute

$$\begin{aligned}\mathbb{P}\left[A_d\right] &= \frac{\omega_{d-2}}{\omega_{d-1}}\int_0^\theta \left(\sin^{d-2}\xi d\xi\right)\\ &\geq \frac{\omega_{d-2}}{\omega_{d-1}}\cdot\frac{\sin^{d-1}\theta}{d-1}\\ &= \frac{\omega_{d-2}}{\omega_{d-1}}\cdot\frac{\sin^{d-1}\left(2\arcsin\left(\frac{\delta}{2}\right)\right)}{d-1}.\end{aligned}$$

Using the identities $\sin(\arcsin x) = x$, $\cos(\arcsin x) = \sqrt{1-x^2}$ and $\sin 2x = 2\sin x \cdot \cos x$, we have

$$\sin^{d-1}\left(2\arcsin\left(\frac{\delta}{2}\right)\right) = \left(\delta\sqrt{1-\frac{\delta^2}{4}}\right)^{d-1}.$$

Finally, $\frac{\omega_{d-2}}{\omega_{d-1}}$ can be shown to be monotonically increasing for all $d \geq 2$, so $\frac{\omega_{d-2}}{\omega_{d-1}} \geq \frac{\omega_0}{\omega_1} = \frac{1}{\pi}$, thus yielding

$$\mathbb{P}[A_d] \geq \frac{1}{\pi(d-1)}\left(\delta\sqrt{1-\frac{\delta^2}{4}}\right)^{d-1},$$

which concludes the proof of the claim. □

We now turn to prove Lemma 7.

Proof (of Lemma 7). Let $U = \text{span}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, and define $T(\mathbf{x}) \coloneqq \frac{\mathbf{u}}{\|\mathbf{u}\|_2}$ where $\mathbf{x} = \mathbf{u} + \mathbf{u}^\perp$ for $\mathbf{u} \in U, \mathbf{u}^\perp \in U^\perp$. First, we observe that for any initialization of $(W, \mathbf{v})$, that $(W, \mathbf{v})$ and $(T(W), \mathbf{v})$ where $T(W) \coloneqq (T(\mathbf{w}_1), \ldots, T(\mathbf{w}_n))$ both belong to the same basin, since $\forall i \in [n], \forall t \in [m]$

$$\begin{aligned}
\langle \mathbf{x}_t, \mathbf{w}_i \rangle &= \left\langle \mathbf{x}_t, \|\mathbf{w}\|_2 \cdot T(\mathbf{w}_i) + \mathbf{w}_i^\perp \right\rangle \\
&= \langle \mathbf{x}_t, \|\mathbf{w}\|_2 \cdot T(\mathbf{w}_i) \rangle + \left\langle \mathbf{x}_t, \mathbf{w}_i^\perp \right\rangle \\
&= \langle \mathbf{x}_t, \|\mathbf{w}\|_2 \cdot T(\mathbf{w}_i) \rangle \\
&= \|\mathbf{w}\|_2 \cdot \langle \mathbf{x}_t, T(\mathbf{w}_i) \rangle, \\
\implies \text{sign}(\langle \mathbf{x}_t, \mathbf{w}_i \rangle) &= \text{sign}(\langle \mathbf{x}_t, T(\mathbf{w}_i) \rangle).
\end{aligned}$$

Thus both $W, T(W)$ belong to the same basin, achieving the same minimal value. Since any rotation Θ under which $U^\perp$ is invariant commutes with T, we have for any measurable set $A \subseteq U$

$$\sigma_{\text{rank}(X)-1}(A) = \sigma_{d-1}\left(\Theta T^{-1}(A)\right) = \sigma_{d-1}\left(T^{-1}(\Theta A)\right),$$

where $\sigma(k)$ denotes the k-dimensional Lebesgue measure. So initializing uniformly on an origin-centered sphere of dimension d is equivalent to initializing uniformly on an origin-centered sphere of dimension $\text{rank}(X)$ in the sense of the region we initialize from. We complete the proof by invoking Claim 1 with respect to a $(\text{rank}(X))$-dimensional sphere. □

We are now ready to prove Thm. 3.

Proof (of Thm. 3). We first argue that since our initialization distribution satisfies Assumption 1, we may rescale each neuron once initialized to the unit sphere. This is justified since a linear rescaling of the weight of each neuron does not change the basin we initialized from, so the basin value remains the same. For this reason, we assume without loss of generality the distribution where each neuron is distributed uniformly and independently on the unit sphere. Define

$$p_\epsilon = \frac{1}{2\pi(\text{rank}(X)-1)}\left(\frac{\sqrt{\epsilon}}{nB}\cdot\sqrt{1-\frac{\epsilon}{4n^2B^2}}\right)^{\text{rank}(X)-1} = \Omega\left(\left(\frac{\sqrt{\epsilon}}{nB}\right)^{\text{rank}(X)}\right).$$

Using the positive homogeneity of the ReLU, we can rescale each v_i^* to satisfy $|v_i^*| = 1\ \forall i \in [n]$, and rescale $\mathbf{w}_i^*$ accordingly, so we may also assume $|v_i^*| = 1, \|\mathbf{w}_i^*\| \leq B\ \forall i \in [n]$. From Lemma 7 we have

$$\begin{aligned}
&\mathbb{P}\left[\exists \tilde{\mathbf{w}}_i : \|\|\mathbf{w}_i^*\| \cdot \tilde{\mathbf{w}}_i - \mathbf{w}_i^*\|_2 \leq \frac{\sqrt{\epsilon}}{n},\ \text{sign}(\langle \tilde{\mathbf{w}}, \mathbf{x}_t \rangle) = \text{sign}(\langle \mathbf{w}, \mathbf{x}_t \rangle)\ \forall t \in [m]\right] \\
=&\mathbb{P}\left[\exists \tilde{\mathbf{w}}_i : \left\|\tilde{\mathbf{w}}_i - \frac{\mathbf{w}_i^*}{\|\mathbf{w}_i^*\|}\right\|_2 \leq \frac{\sqrt{\epsilon}}{n\|\mathbf{w}_i^*\|},\ \text{sign}(\langle \tilde{\mathbf{w}}, \mathbf{x}_t \rangle) = \text{sign}(\langle \mathbf{w}, \mathbf{x}_t \rangle)\ \forall t \in [m]\right] \\
\geq&\mathbb{P}\left[\exists \tilde{\mathbf{w}}_i : \left\|\tilde{\mathbf{w}}_i - \frac{\mathbf{w}_i^*}{\|\mathbf{w}_i^*\|}\right\|_2 \leq \frac{\sqrt{\epsilon}}{nB},\ \text{sign}(\langle \tilde{\mathbf{w}}, \mathbf{x}_t \rangle) = \text{sign}(\langle \mathbf{w}, \mathbf{x}_t \rangle)\ \forall t \in [m]\right] \\
=&2p_\epsilon,
\end{aligned}$$

and also

$$\mathbb{P}\left[\operatorname{sign}\left(v_i\right)=\operatorname{sign}\left(v_i^*\right)\right]=\frac{1}{2}.$$

Since the two events are independent, we have that both occur w.p. at least p_ϵ. Also, this event is independent for each neuron, so we have w.p. at least p_ϵ for each neuron to initialize 'close' enough to $(\mathbf{w}_i^*, v_i^*)$. In this sense, we can lower bound the number of good initializations from below using $Z \sim B\left(N, p_\epsilon\right)$, where $B\left(N,p\right)$ is the binomial distribution. By using Chernoff's bound we bound the tail of Z as follows

$$\begin{aligned}
&F\left(n; c\left\lceil\frac{n}{p_\epsilon}\right\rceil, p_\epsilon\right)\\
\leq &\exp\left(-\frac{1}{2p_\epsilon}\frac{\left(c\lceil\frac{n}{p_\epsilon}\rceil p_\epsilon - n\right)^2}{c\lceil\frac{n}{p_\epsilon}\rceil}\right)\\
\leq &\exp\left(-\frac{1}{2}\frac{\left(cn-n\right)^2}{cn}\right)\\
\leq &\exp\left(-\frac{1}{4}cn\right).
\end{aligned}$$

Thus with probability $\geq 1-\exp\left(-\frac{1}{4}cn\right)$, we have n neurons initialized in a basin containing a point $\tilde{W} \in \mathbb{R}^{n\times d}$ of distance at most $\frac{\sqrt{\epsilon}}{n}$ from an optimal weight $\mathbf{w}_i^*$ for each $i \in [n]$.

Let $i_1, \ldots, i_n$ be the indices of the well initialized neurons, and let

$$\tilde{W}_i = \left(\mathbf{w}_1^*, \ldots, \mathbf{w}_i^*, \left\|\mathbf{w}_{i+1}^*\right\| \tilde{\mathbf{w}}_{i+1}, \ldots, \left\|\mathbf{w}_n^*\right\| \tilde{\mathbf{w}}_n\right),\ i=0,\ldots,n.$$

We compute the value of the basin corresponding to these neurons as follows:

$$\begin{aligned}
\operatorname{Bas}\left(\mathbf{w}_{i_1}, \ldots, \mathbf{w}_{i_n}, v_{i_1}, \ldots, v_{i_n}\right) &\leq L\left(\tilde{W}, \mathbf{v}^*\right)\\
&= \frac{1}{m}\sum_{t=1}^{m}\left(N_n\left(\tilde{W}_0, \mathbf{v}^*\right)\left(\mathbf{x}_t\right) - y_t\right)^2\\
&= \frac{1}{m}\sum_{t=1}^{m}\left|\sum_{i=1}^{n}\left(N_n\left(\tilde{W}_{i-1}, \mathbf{v}^*\right)\left(\mathbf{x}_t\right) - N_n\left(\tilde{W}_i, \mathbf{v}^*\right)\left(\mathbf{x}_t\right)\right)\right|^2\\
&\leq \frac{1}{m}\sum_{t=1}^{m}\left(\sum_{i=1}^{n}\left|N_n\left(\tilde{W}_{i-1}, \mathbf{v}^*\right)\left(\mathbf{x}_t\right) - N_n\left(\tilde{W}_i, \mathbf{v}^*\right)\left(\mathbf{x}_t\right)\right|\right)^2\\
&\leq \frac{1}{m}\sum_{t=1}^{m}\left|\left|v_i^*\right|\left\|\mathbf{x}_t\right\|\left(\sum_{i=1}^{n}\left\|\tilde{W}_{i-1} - \tilde{W}_i\right\|\right)\right|^2\\
&\leq \left(\sum_{i=1}^{n}\left\|\tilde{W}_{i-1} - \tilde{W}_i\right\|\right)^2\\
&= \left(\sum_{i=1}^{n}\left\|\left\|\mathbf{w}_i^*\right\| \cdot \tilde{\mathbf{w}}_i - \mathbf{w}_i^*\right\|\right)^2\\
&\leq \left(\sum_{i=1}^{n}\frac{\sqrt{\epsilon}}{n}\right)^2\\
&= \epsilon,
\end{aligned}$$

where the second inequality in the triangle inequality and the third inequality is from Lemma 6. We now finish the proof by invoking Lemma 2 to conclude

$$\mathbb{P}\left[\operatorname{Bas}\left(W, \mathbf{v}\right) \leq \operatorname{Bas}\left(\mathbf{w}_{i_1}, \ldots, \mathbf{w}_{i_n}, v_{i_1}, \ldots, v_{i_n}\right) \leq \epsilon\right] \geq 1 - e^{-\frac{1}{4}cn}.$$

□

B.6. Proof of Thm. 4

Denote the initialization point as $W = (\mathbf{w}_1, \ldots, \mathbf{w}_n, v_1, \ldots, v_n)$, and define $(W', \mathbf{v}')$ with $\mathbf{v}' = (\operatorname{sign}(v_1), \ldots, \operatorname{sign}(v_n))$, $\mathbf{w}'_i = \sum_{t'=1}^{m} a_{i,t'} \mathbf{x}_{t'}$ where $a_{i,t'} \in \mathbb{R}$ are to be determined later. Let

$$(\bar{y}_1, \ldots, \bar{y}_m) = \operatorname*{argmin}_{(\bar{y}_1, \ldots, \bar{y}_m) \in \mathbb{R}^m} \frac{1}{m} \sum_{t=1}^{m} \ell(\bar{y}_t, y_t),$$

we want to show that for well chosen values of $a_{i,t'}$, $(W', \mathbf{v}')$ belongs to the same basin as $(W, \mathbf{v})$, and achieves the desired prediction $(\bar{y}_1, \ldots, \bar{y}_m)$ over a certain subset of $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, while achieving a prediction of 0 over the rest of the sample instances, effectively predicting the subset without affecting the prediction over the rest of the sample. By combining enough neurons in this manner, we are able to obtain the minimal objective value over the data. Namely, an objective value of α. Define the vector $\mathbf{y}'_i = \left(y'_{i,1}, \ldots, y'_{i,m}\right)^\top$, where

$$y'_{i,t} = \begin{cases} |\bar{y}_t| & \langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0,\ \forall j < i\ \langle \mathbf{w}_j, \mathbf{x}_t \rangle \leq 0 \wedge v_j \cdot \bar{y}_t < 0 \\ 0 & \text{otherwise} \end{cases}.$$

and choose $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,m})^\top$ such that the equality

$$XX^\top \mathbf{a}_i = \mathbf{y}'_i$$

holds.

We first stress that by our assumption,

$$XX^\top = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \ldots & \langle \mathbf{x}_1, \mathbf{x}_m \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{x}_m, \mathbf{x}_1 \rangle & \ldots & \langle \mathbf{x}_m, \mathbf{x}_m \rangle \end{pmatrix} \in \mathbb{R}^{m \times m}$$

is of rank m, and therefore $\mathbf{a}_i$ exists and is well-defined.

Assuming that for any $t \in [m]$ there exists some neuron i such that $\langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0$ (We will later analyze the probability of this actually happening), we compute the prediction of our network with weights $W' = (\mathbf{w}'_1, \ldots, \mathbf{w}'_n)$ on $\mathbf{x}_t$:

$$\begin{aligned}
N_n(W', \mathbf{v}')(\mathbf{x}_t) &= \sum_{i=1}^{n} v'_i \left[\langle \mathbf{w}'_i, \mathbf{x}_t \rangle\right]_+ \\
&= \sum_{i=1}^{n} v_i \left[\left\langle \sum_{t'=1}^{m} a_{i,t'} \mathbf{x}_{t'}, \mathbf{x}_t \right\rangle\right]_+ \\
&= \sum_{i=1}^{n} v_i \left[\sum_{t'=1}^{m} a_{i,t'} \langle \mathbf{x}_{t'}, \mathbf{x}_t \rangle\right]_+ \\
&= \sum_{i=1}^{n} v_i \left[y'_{i,t}\right]_+ \\
&= \sum_{i=1}^{n} v_i \left[|\bar{y}_t| \cdot \mathbb{1}_{\langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0,\ \forall j < i\ \langle \mathbf{w}_j, \mathbf{x}_t \rangle \leq 0 \wedge v_j \cdot \bar{y}_t < 0}\right]_+ \\
&= \sum_{i=1}^{n} \bar{y}_t \cdot \mathbb{1}_{\langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0,\ \forall j < i\ \langle \mathbf{w}_j, \mathbf{x}_t \rangle \leq 0 \wedge v_j \cdot \bar{y}_t < 0} \\
&= \bar{y}_t.
\end{aligned}$$

Where the last equality comes from our assumption that there exists some neuron i s.t. $\langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0$, and from the definition of $y'_{i,t}$ which asserts that at most a single neuron will predict $\mathbf{x}_t$. Thus, we have

$$\forall t \in [m]\ \ N_n\left(W', \mathbf{v}\right)(\mathbf{x}_t) = \bar{y}_t$$

$$\implies L_S\left(W', \mathbf{v}\right) = \alpha.$$

To put this result in different words, if $\mathbf{x}_t$ is positive on the hyperplane induced by $\mathbf{w}_i$ and if v_i has the same sign as $\bar{y}_t$, then $\mathbf{w}'_i$ predicts $\mathbf{x}_t$ correctly, given that $\mathbf{x}_t$ was not previously predicted by a neuron $\mathbf{w}'_j$ where $j < i$.

We now assert that $(W, \mathbf{v})$ and $(W', \mathbf{v}')$ indeed belong to the same basin with respect to S. For $\mathbf{v}, \mathbf{v}'$ this is clear by definition, and note that for $\mathbf{w}_i, \mathbf{w}'_i$ we require that $\operatorname{sign}\left(\langle \mathbf{w}_i, \mathbf{x}_t \rangle\right) \cdot \operatorname{sign}\left(\langle \mathbf{w}'_i, \mathbf{x}_t \rangle\right) \geq 0,\ \forall i \in [n], t \in [m]$. Thus we compute:
If $\langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0,\ \forall j < i\ \langle \mathbf{w}_j, \mathbf{x}_t \rangle \leq 0 \wedge v_j \cdot \bar{y}_t < 0$ all hold, then we have

$$\operatorname{sign}\left(\langle \mathbf{w}_i, \mathbf{x}_t \rangle\right) = 1,$$

and

$$\begin{aligned}
\operatorname{sign}\left(\langle \mathbf{w}'_i, \mathbf{x}_t \rangle\right) &= \operatorname{sign}\left(\langle \mathbf{w}'_i, \mathbf{x}_t \rangle\right) \\
&= \operatorname{sign}\left(\left\langle \sum_{t'=1}^{m} a_{i,t'} \mathbf{x}_{t'}, \mathbf{x}_t \right\rangle\right) \\
&= \operatorname{sign}\left(\sum_{t'=1}^{m} a_{i,t'} \left\langle \mathbf{x}_{t'}, \mathbf{x}_t \right\rangle\right) \\
&= \operatorname{sign}\left(y'_{i,t}\right) \\
&= \operatorname{sign}\left(|\bar{y}_t|\right) \\
&\geq 0.
\end{aligned}$$

Otherwise, we have $\operatorname{sign}\left(\langle \mathbf{w}_i, \mathbf{x}_t \rangle\right) \leq 0$ and

$$\operatorname{sign}\left(\langle \mathbf{w}'_i, \mathbf{x}_t \rangle\right) = \operatorname{sign}\left(y'_{i,t}\right) = 0.$$

Finally, we define the event $A_i^t \coloneqq \langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0,\ v_i \cdot \bar{y}_t \geq 0$, i.e. the i^{th} neuron is able to predict $\mathbf{x}_t$ correctly. Since $v_i, \mathbf{w}_i$ are independent, and since $\mathbf{w}_i$ is drawn from a spherically symmetric distribution for all $i \in [n]$, we have

$$\mathbb{P}\left[A_i^t\right] = \mathbb{P}\left[\langle \mathbf{w}_i, \mathbf{x}_t \rangle > 0\right] \cdot \mathbb{P}\left[v_i \cdot, \bar{y}_t \geq 0\right] \geq \frac{1}{4}$$

$$\implies \mathbb{P}\left[\overline{A_i^t}\right] \leq \frac{3}{4}.$$

Since the neurons are independent, and since $\left(\operatorname{sign}\left(v_1\right), \ldots, \operatorname{sign}\left(v_n\right)\right)$ is uniformly distributed on the Boolean cube, we have

$$\mathbb{P}\left[\bigcap_{i=1}^{n} \overline{A_i^t}\right] \leq \left(\frac{3}{4}\right)^n.$$

Using the union bound on $\bigcap_{i=1}^{n} \overline{A_i^t}$ for $t = 1, \ldots, m$ we get

$$\mathbb{P}\left[\exists t \text{ s.t. no neuron predicts } \mathbf{x}_t\right] \leq m \left(\frac{3}{4}\right)^n.$$

Thus, the probability of initializing from a basin achieving a global minimum with value α is at least

$$1 - m \left(\frac{3}{4}\right)^n.$$

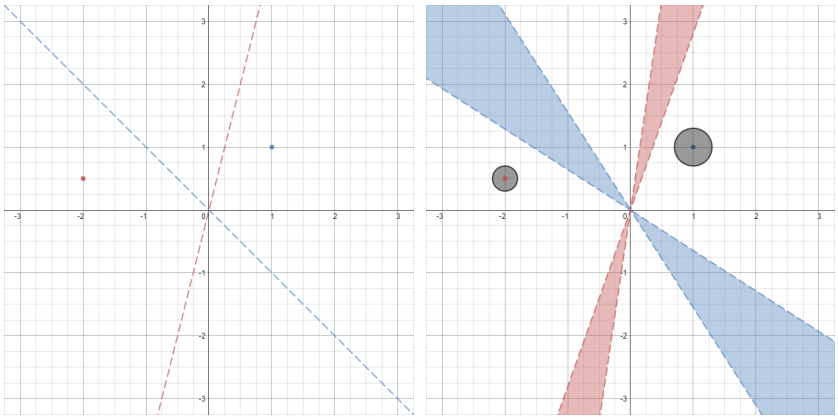

Figure 1. The partition of $\mathbb{R}^2$ into regions by the instances $\mathbf{c}_1 = (1,1), \mathbf{c}_2 = (-2, 0.5)$, and the corresponding partition by clustered instances with centers $\mathbf{c}_1, \mathbf{c}_2$. The noisy regions are depicted by the light blue and light red.

B.7. Proof of Thm. 5

The idea behind the proof is comprised of two parts. The first is that by predicting the clusters' centers well, we are able to obtain a good objective value over the data. The second is that the basin partition of the clustered data is similar to the basin partition of the clusters' centers. So by approximating a good solution for the clusters' centers, we are able to reach a good objective value.

Recall that by Definition 2, we partition the parameter space $\mathbb{R}^{n\times d}$ of the first layer into sets where $\operatorname{sign}(\langle \mathbf{w}_i, \mathbf{x}_t \rangle)$ is fixed for all $i \in [n], t \in [m]$. This restricts the possible weight vector of each neuron in the first layer to a subset of $\mathbb{R}^d$. Referring to these subsets as *regions*, we observe that their structure varies slightly when changing δ from 0 (where a cluster contains a single point) to a small positive quantity, where the new regions introduced by the clusters are referred to as *noisy regions* (see Fig. 1).

To approximate a good solution for the clusters' centers, we need to initialize from a basin where such an approximation exists. Note that if $\delta = 0$, then the result will hold as a corollary of Thm. 4. Alternatively, if δ is small enough, then we would expect such an approximation to exist in the basins comprised of non-noisy regions, as these vary slightly when δ is small. Therefore, we would like to assert that we initialize from these basins to guarantee the existence of a good solution.

Before delving into the proof of Thm. 5, we first prove two auxiliary lemmas (Lemma 8 and Lemma 9). The following lemma provides an upper bound on initializing a single neuron from a noisy region, for distributions satisfying Assumption 1.

Lemma 8. *Define the set of noisy regions with respect to the j^{th} cluster,*

$$A_j = \left\{ \mathbf{x} : \|\mathbf{x}\|_2 = 1,\ \exists \mathbf{y} : \|\mathbf{c}_j - \mathbf{y}\|_2 \le \delta_j,\ \langle \mathbf{x}, \mathbf{y} \rangle = 0 \right\}.$$

Then under the assumptions in Thm. 5, its complement with respect to the d-dimensional unit sphere $A_j^c = \mathbb{S}^{d-1} \backslash A_j$ satisfies

$$\frac{\sigma_{d-1}\left(A_j^c\right)}{\omega_{d-1}} \ge 1 - \frac{1}{4d}.$$

Where σ_{d-1} is the $(d-1)$-dimensional Lebesgue measure, and ω_{d-1} is the surface area of the d-dimensional unit sphere.

To prove the lemma we will need two auxiliary claims.

Claim 2. *Let $\mathsf{S}(\mathbf{a}, \theta) := \left\{ \mathbf{b} \in \mathbb{S}^{d-1} : \langle \mathbf{a}, \mathbf{b} \rangle > \cos\theta \right\}$ denote the open hyperspherical cap of spherical radius θ and center $\mathbf{x}$. Then*

$$\mathsf{S}\left(\mathbf{c}_j, \frac{\pi}{2} - 2\arcsin\frac{\delta_j}{2\|\mathbf{c}_j\|}\right) \dot{\cup} \mathsf{S}\left(-\mathbf{c}_j, \frac{\pi}{2} - 2\arcsin\frac{\delta_j}{2\|\mathbf{c}_j\|}\right) \subseteq A_j^c.$$

Proof. Clearly, the two open hyperspherical caps are disjoint, as they are of spherical radius $\le \frac{\pi}{2}$ and the two originate in two diametrically opposite points. Assume $\mathbf{x} \in \mathsf{S}\left(\mathbf{c}_j, \frac{\pi}{2} - 2\arcsin\frac{\delta_j}{2\|\mathbf{c}_j\|}\right)$, then the projection of $\left\{ \mathbf{z} : \|\mathbf{c}_j - \mathbf{z}\|_2 \le \delta_j \right\}$

onto $\mathbb{S}^{d-1}$, denoted P_j, is a hyperspherical cap of spherical radius $\theta \coloneqq 2\arcsin\frac{\delta_j}{2\|\mathbf{c}_j\|}$. Since the dot product is a bi-linear operation, it suffices to show that $\forall \mathbf{y} \in P_j \ \langle \tilde{\mathbf{x}}, \mathbf{y}\rangle \neq 0$, where $\tilde{\mathbf{x}} \in \mathbb{S}^{d-1}$ is the projection of $\mathbf{x}$ onto $\mathbb{S}^{d-1}$.

Let $\mathbf{y} \in P_j$, using the fact that $\mathsf{s} : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}_+$, the spherical distance function defined by $\mathsf{s}(\mathbf{a}, \mathbf{b}) \coloneqq \arccos(\langle \mathbf{a}, \mathbf{b}\rangle)$, satisfies the triangle inequality we have

$$\begin{aligned}\mathsf{s}(\tilde{\mathbf{x}}, \mathbf{y}) &\leq \mathsf{s}(\tilde{\mathbf{x}}, \mathbf{c}_j) + \mathsf{s}(\mathbf{c}_j, \mathbf{y}) \\ &< \frac{\pi}{2} - \theta + \theta \\ &= \frac{\pi}{2}, \\ &\implies \langle \tilde{\mathbf{x}}, \mathbf{y}\rangle \neq 0.\end{aligned}$$

Where the same argument works for $\mathbf{x} \in \mathsf{S}\left(-\mathbf{c}_j, \frac{\pi}{2} - 2\arcsin\frac{\delta_j}{2\|\mathbf{c}_j\|}\right)$ and $-P_j$. □

Moving on to our next auxiliary claim.

Claim 3. *$\forall \theta \geq 0$ we have*

$$\int_0^{\frac{\pi}{2}-\theta} \sin^{d-2}\xi d\xi \geq \frac{\omega_{d-1}}{2\omega_{d-2}} - \theta.$$

Proof. Consider the function $f(\theta) = \left(\int_0^{\frac{\pi}{2}-\theta} \sin^{d-2}\xi d\xi\right) - \left(\frac{\omega_{d-1}}{2\omega_{d-2}} - \theta\right)$, it is monotonically increasing in $[0, \infty)$ since

$$\begin{aligned}f'(\theta) &= \frac{\partial}{\partial\theta}\left(\left(\int_0^{\frac{\pi}{2}-\theta} \sin^{d-2}\xi d\xi\right) - \left(\frac{\omega_{d-1}}{2\omega_{d-2}} - \theta\right)\right) \\ &= -\sin^{d-2}\left(\frac{\pi}{2} - \theta\right) + 1 \\ &\geq 0.\end{aligned}$$

And since $f(0) = 0$ we have $\forall \theta \in [0, \infty)$ that $\int_0^{\frac{\pi}{2}-\theta} \sin^{d-2}\xi d\xi \geq \frac{\omega_{d-1}}{2\omega_{d-2}} - \theta$. □

We now turn to prove Lemma 8.

Proof (of Lemma 8). Using Claims 2 and 3, Eq. (10) and the fact that $\forall d \geq 2 \ \frac{\omega_{d-1}}{\omega_d} \leq \sqrt{\frac{d}{2\pi}}$ ((Leopardi, 2007), Lemma 2.3.20), we have the following:

$$\begin{aligned}\frac{\sigma_{d-1}(A_j^c)}{\omega_{d-1}} &\geq \frac{\sigma_{d-1}\left(\mathsf{S}\left(\mathbf{c}_j, \frac{\pi}{2} - \theta\right)\right) + \sigma_{d-1}\left(\mathsf{S}\left(-\mathbf{c}_j, \frac{\pi}{2} - \theta\right)\right)}{\omega_{d-1}} \\ &= \frac{2\nu_{d-1}\left(\frac{\pi}{2} - \theta\right)}{\omega_{d-1}} \\ &= 2\frac{\omega_{d-2}}{\omega_{d-1}}\int_0^{\frac{\pi}{2}-\theta} \sin^{d-2}\xi d\xi \\ &\geq 2\frac{\omega_{d-2}}{\omega_{d-1}}\left(\frac{\omega_{d-1}}{2\omega_{d-2}} - \theta\right) \\ &= 1 - \frac{2\omega_{d-2}\theta}{\omega_{d-1}} \\ &\geq 1 - 4\sqrt{\frac{d}{2\pi}} \cdot \arcsin\frac{\delta_j}{2\|\mathbf{c}_j\|} \\ &\geq 1 - 4\sqrt{\frac{d}{2\pi}} \cdot \arcsin\left(\sin\left(\frac{\sqrt{2\pi}}{16d\sqrt{d}}\right)\right) \\ &= 1 - \frac{1}{4d}.\end{aligned}$$

□

Now that we can bound the probability of initializing from a noisy region $A_j, j \in [k]$, we turn to show that with high probability, a solution with $\mathcal{O}\left(\delta^2\right)$ value can be found. Let C be the matrix with rows $\mathbf{c}_1, \ldots, \mathbf{c}_k$, then by Thm. 4 we know that with high probability there exists some $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ which achieves a value of 0 on the dataset $\{\mathbf{c}_j, \hat{y}_j\}_{j=1}^k$, and since the cluster target values are γ-Lipschitz, this $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ will also perform well on S. Unfortunately, we cannot guarantee that $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ resides in the basin we initialized from, as this guarantee can only be given on the basin partition where $\delta = 0$. Instead, we take a surrogate $(W', \mathbf{v}')$ which approximates $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ well, and then show that the value it achieves is also of magnitude $\mathcal{O}\left(\delta^2\right)$. More formally, we have the following lemma.

Lemma 9. *Let C be a matrix with rows $\mathbf{c}_1, \ldots, \mathbf{c}_k$, satisfying rank $(C) = k$. Let $(W, \mathbf{v}) \in B_S^{A,\mathbf{b}}$ satisfy $\forall j \in [k]\,,\ \exists i \in [n]:\ \mathbf{w}_i \notin \cup_r A_r, \langle \mathbf{w}_i, \mathbf{c}_j \rangle > 0, v_i \cdot \hat{y}_j \geq 0$. Then exist $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ and $(W', \mathbf{v}')$ where the following holds:*

1. *$\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ predicts y_t well:*

$$\left|N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right) - y_t\right| \leq \delta\left(n\frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2 + 2\gamma\right).$$

2. *$(W', \mathbf{v}') \in B_S^{A,\mathbf{b}}$ approximates $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ well:*

$$\left|N_n\left(W', \mathbf{v}'\right) - N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)\right| \leq nB \cdot \frac{\delta}{c}\frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2.$$

Before proving the lemma, we state and prove the following two auxiliary claims.

Claim 4. *Let $\tilde{\mathbf{w}}_i \coloneqq \sum_{j=1}^k a_{i,j}\mathbf{c}_j = C^\top \mathbf{a}_i$, where $\mathbf{a}_i$ satisfies the equality $CC^\top \mathbf{a}_i = \mathbf{y}_i'$ as in Appendix B.6. Then for all $i \in [n]$,*

$$\left\|\tilde{\mathbf{w}}_i\right\|_2 \leq \frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2.$$

Proof. We derive a bound on $\left\|\tilde{\mathbf{w}}_i\right\|_2$ as follows:

$$\begin{aligned}
CC^\top \mathbf{a}_i &= \mathbf{y}_i', \\
\implies \mathbf{a}_i &= \left(CC^\top\right)^{-1}\mathbf{y}_i', \\
\implies C^\top \mathbf{a}_i &= C^\top\left(CC^\top\right)^{-1}\mathbf{y}_i', \\
\implies \left\|\tilde{\mathbf{w}}_i\right\|_2 &= \left\|C^\top\left(CC^\top\right)^{-1}\mathbf{y}_i'\right\|_2 \\
&\leq \left\|C^\top\right\|_{\text{op}}\left\|\left(CC^\top\right)^{-1}\right\|_{\text{op}}\left\|\mathbf{y}_i'\right\|_2 \\
&= \sigma_{\max}\left(C^\top\right) \cdot \frac{1}{\sigma_{\min}^2\left(C^\top\right)}\left\|\mathbf{y}_i'\right\|_2 \\
&\leq \frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2.
\end{aligned}$$

□

Claim 5. *$N_n\left(\mathbf{w}_1, \ldots, \mathbf{w}_n, \mathbf{v}\right)(\mathbf{x})$ is $\left(\sum_{i=1}^n |v_i| \cdot \|\mathbf{w}_i\|\right)$-Lipschitz in $\mathbf{x}$.*

The proof of this claim follows the same idea behind Lemma 6, and is therefore omitted.

We are now ready to prove Lemma 9.

Proof (of Lemma 9). We first define $\left(\tilde{W}, \tilde{\mathbf{v}}\right)$ as the point satisfying $E_{S'}\left(\tilde{W}, \tilde{\mathbf{v}}\right) = 0$, as demonstrated in Appendix B.6. Defining $(W', \mathbf{v}')$, we let $\mathbf{v}' = \tilde{\mathbf{v}} \in \{-1, +1\}^n$. If $(\tilde{\mathbf{w}}_i, \mathbf{w}'_i)$ both belong to the same region with respect to S, then take $\mathbf{w}'_i = \tilde{\mathbf{w}}_i$. Otherwise, we approximate $\tilde{\mathbf{w}}_i$ in the $\|\cdot\|_2$ sense, by taking $\mathbf{w}'_i$ in the region we initialized from which is closest to $\tilde{\mathbf{w}}_i$.

1. We compute using Claims 4 and 5,

$$\begin{aligned}
&\left|N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t) - y_t\right| \\
=&\left|N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t) - N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathsf{c}(\mathbf{x}_t)) + N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathsf{c}(\mathbf{x}_t)) - y_t\right| \\
\leq&\left|N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t) - N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathsf{c}(\mathbf{x}_t))\right| + \left|N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathsf{c}(\mathbf{x}_t)) - y_t\right| \\
\leq&\sum_{i=1}^{n} \|\tilde{\mathbf{w}}_i\| \cdot \|\mathbf{x}_t - \mathsf{c}(\mathbf{x}_t)\| + |\hat{y}_t - y_t| \\
\leq&\delta\left(n\frac{\sigma_{\max}\left(C^\top\right)}{\sigma^2_{\min}\left(C^\top\right)}\|\hat{\mathbf{y}}\|_2 + 2\gamma\right).
\end{aligned}$$

   where the last inequality comes from $y_t, \hat{y}_t$ being the target values of points belonging to a ball of diameter at most 2δ and the target values being γ-Lipschitz.

2. Note that by definition we have $(W', \mathbf{v}') \in B_S^{A,\mathbf{b}}$. Denote the origin as O. In the worst case, $\tilde{\mathbf{w}}$ is on the line connecting O and $\mathbf{c}_i$, so assume this is the case. Denote the point at which the line connecting O and $\tilde{\mathbf{w}}_i$ is tangent to the i^{th} cluster by H_i, then the vertices $O, \mathbf{w}'_i, \tilde{\mathbf{w}}_i$ and $O, H_i, \mathbf{c}_i$ form similar triangles, and we have

$$\|\mathbf{w}'_i - \tilde{\mathbf{w}}_i\|_2 = \frac{\delta_i}{\|\mathbf{c}_i\|_2}\|\tilde{\mathbf{w}}_i\|_2 \leq \frac{\delta}{c}\|\tilde{\mathbf{w}}_i\|_2.$$

   Now, using Claim 4 and Lemma 6,

$$\begin{aligned}
&\left|N_n\left(W', \mathbf{v}'\right)(\mathbf{x}_t) - N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t)\right| \\
=&\left|N_n\left(W', \mathbf{v}'\right)(\mathbf{x}_t) - N_n\left(\tilde{W}, \mathbf{v}'\right)(\mathbf{x}_t)\right| \\
\leq&\sum_{i=1}^{n} |v_i| \cdot \|\mathbf{x}_t\|_2 \cdot \|\mathbf{w}'_i - \tilde{\mathbf{w}}_i\|_2 \\
\leq&\sum_{i=1}^{n} B \cdot \frac{\delta}{c}\|\tilde{\mathbf{w}}_i\|_2 \\
\leq&\sum_{i=1}^{n} B \cdot \frac{\delta}{c}\frac{\sigma_{\max}\left(C^\top\right)}{\sigma^2_{\min}\left(C^\top\right)}\|\hat{\mathbf{y}}\|_2 \\
=&nB \cdot \frac{\delta}{c}\frac{\sigma_{\max}\left(C^\top\right)}{\sigma^2_{\min}\left(C^\top\right)}\|\hat{\mathbf{y}}\|_2.
\end{aligned}$$

□

Equipped with the above lemmas, we are now ready to prove Thm. 5.

Proof (of Thm. 5). Using Lemma 8, we have for all $j \in [k]$

$$\frac{\sigma_{d-1}\left(A_j^c\right)}{\omega_{d-1}} \geq 1 - \frac{1}{4d}.$$

Applying the union bound to the $k \leq d$ events where we initialize from A_j, we have that we don't initialize a single neuron from a noisy region w.p. at least $\frac{3}{4}$. For a given $j \in [k]$, using the union bound again, the probability of initializing from a non-noisy region in which any internal point $\mathbf{w} \in \mathbb{R}^d$ satisfies $\langle \mathbf{w}, \mathbf{c}_j \rangle > 0$ is at least $\frac{1}{4}$, and finally, since v_i has the correct sign w.p. $\frac{1}{2}$ and is independent of where we initialize $\mathbf{w}_i$ from, we are unable to predict $\mathbf{c}_j$ w.p. at most $\frac{7}{8}$. Using the union bound once more in the same manner as we did in Appendix B.6 gives that we initialize "properly" w.p. at least

$$1 - k\left(\frac{7}{8}\right)^n \geq 1 - d\left(\frac{7}{8}\right)^n.$$

We stress that by using Lemma 2, for the purpose of analyzing the objective value, we can ignore initializations made from noisy regions, as we may just consider the neurons that were properly initialized. By our assumption that the clusters' centers are in general position, namely that the matrix C with rows $\mathbf{c}_1, \ldots, \mathbf{c}_k$ satisfies $\sigma_{\min}\left(C^\top\right) > 0$, we have that it is in particular of rank k, and the conditions in Lemma 9 are met, so we compute

$$\begin{aligned}
L_S\left(W', \mathbf{v}'\right) &= \frac{1}{m}\sum_{t=1}^{m}\left(N_n\left(W', \mathbf{v}'\right)(\mathbf{x}_t) - \hat{y}_t\right)^2 \\
&= \frac{1}{m}\sum_{t=1}^{m}\left|N_n\left(W', \mathbf{v}'\right)(\mathbf{x}_t) - N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t) + N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t) - \hat{y}_t\right|^2 \\
&\leq \frac{1}{m}\sum_{t=1}^{m}\left(\left|N_n\left(W', \mathbf{v}'\right)(\mathbf{x}_t) - N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t)\right| + \left|N_n\left(\tilde{W}, \tilde{\mathbf{v}}\right)(\mathbf{x}_t) - \hat{y}_t\right|\right)^2 \\
&\leq \frac{1}{m}\sum_{t=1}^{m}\left(nB \cdot \frac{\delta}{c}\frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2 + \delta n\frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2 + 2\gamma\delta\right)^2 \\
&= \delta^2\left(\left(1 + \frac{B}{c}\right)n\frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2 + 2\gamma\right)^2.
\end{aligned}$$

Thus we conclude that when $(W, \mathbf{v})$ is initialized using a distribution satisfying Assumption 1, we have

$$\mathbb{P}\left[\text{Bas}\left(W, \mathbf{v}\right) \leq \delta^2\left(\left(1 + \frac{B}{c}\right)n\frac{\sigma_{\max}\left(C^\top\right)}{\sigma_{\min}^2\left(C^\top\right)}\left\|\hat{\mathbf{y}}\right\|_2 + 2\gamma\right)^2\right] \geq 1 - d\left(\frac{7}{8}\right)^n.$$

□

C. Poor Basin Structure for Single Neurons

In this appendix, we prove a hardness result for initializing ReLU single neuron nets with convex losses from a basin (as will shortly be defined for the single neuron architecture context) with a good basin value, and then provide an explicit construction for the squared loss.

For a single neuron, our objective function with respect to a ReLU activation and convex loss function ℓ is

$$L_S\left(\mathbf{w}\right) = \frac{1}{m}\sum_{t=1}^{m}\ell\left(\left[\langle \mathbf{w}, \mathbf{x}_t\rangle\right]_+, y_t\right),$$

corresponding to the parameter space $\mathbb{R}^d$. As done in Sec. 4 for two-layer networks, we can partition the parameter space according to the signs of $\langle \mathbf{w}, \mathbf{x}_t\rangle$ on each training instance $\mathbf{x}_t$. Each region in this partition corresponds to an intersection of halfspaces, in which our objective $L_S(\mathbf{w})$ can easily be shown to be convex. Thus, each such region corresponds to basin (as defined in Definition 1), and we can consider the probability of initializing in a basin with low minimal value.

C.1. Exponentially Many Poor Local Minima

Building on the work of (Auer et al., 1996), we provide a construction of a dataset which results in exponentially many poor local minima in the dimension. Moreover, we provide in subsection Appendix C.2 an explicit construction for the

squared loss. The results extend those of (Auer et al., 1996) by showing that they hold for a single neuron with the ReLU activation function (for which the technical conditions assumed in (Auer et al., 1996) do not apply).

From an optimization point of view, having exponentially many local minima is not necessarily problematic as many of which may obtain good objective values. However, following our initialization scheme throughout this work, we modify the result obtained in (Auer et al., 1996) to satisfy that when the weight vector of the neuron is initialized from a distribution satisfying Assumption 1, then the distribution of the minimal value in the basin we initialize from is strongly concentrated around a sub-optimal value as the dimension increases. More formally, we have the following Theorem.

Theorem 6. *Consider a ReLU single neuron neural net, with a convex and symmetric loss function ℓ satisfying $\ell(a,b)=0$ if and only if $a=b$. Then for all $\epsilon>0$ there exists a sample S such that $L_S(\mathbf{w}^*)=\epsilon$ for some $\mathbf{w}^*\in\mathbb{R}^d$, and a constant $c\in\mathbb{R}$ which depends only on ℓ, such that the objective value over the sample L_S contains 2^d strict local minima, and*

$$\mathbb{P}\left[Bas\left(\mathbf{w}\right)\leq\frac{c}{4}\right]\leq e^{-\frac{d}{16}}.$$

Where $\mathbf{w}$ *is initialized according to Assumption 1.*

In other words, we have exponentially many local minima, where the probability of initializing from a sub-optimal basin converges exponentially fast (in the dimension) to 1, yet there exists a solution which obtains a value of ϵ.

Proof. Let $L_0=\ell(0,1)$. Since $[0]_+=0\neq 1=[1]_+$ we have $L_0>0$. We are interested in a construction where ϵ is small enough, therefore assume $\epsilon<\frac{L_0}{2}$. By the continuity of ℓ as a convex function, we can find $\delta\in(0,1)$ small enough such that $\ell(0,\delta)=2\epsilon$.

Consider the sample

$$S=\left\{(x_1=\delta,y_1=\delta),(x_2=-1,y_2=1)\right\}.$$

We compute

$$\ell\left(\left[wx_1\right]_+,y_1\right)=\begin{cases}2\epsilon & w\in(-\infty,0]\\ 0 & w=1\end{cases},$$

$$\ell\left(\left[wx_2\right]_+,y_2\right)=\begin{cases}0 & w=-1\\ L_0 & w\in[0,\infty)\end{cases}.$$

Therefore the objective value over S, $L_S(w)=\frac{1}{2}\left(\ell\left(\left[wx_1\right]_+,y_1\right)+\ell\left(\left[wx_2\right]_+,y_2\right)\right)$ satisfies

$$L_S(w)=\begin{cases}\epsilon & w=-1\\ \frac{L_0}{2} & w=1\\ >\frac{L_0}{2} & w\in[0,1)\cup(1,\infty)\end{cases}.$$

But since ℓ is convex, we have that L_S is convex in $(-\infty,0]$ and in $[0,\infty)$, so L_S has exactly two local minima, one is $L_S(-1)=\epsilon$ and the other is $L_S(1)=\frac{L_0}{2}$.

We now extend our sample to be d-dimensional in a similar manner as did the authors in (Auer et al., 1996) as follows: For $t=1,2$ and $j\in[d]$, we use the mapping $x_t(j)\mapsto(0,\ldots,0,x_t,0,\ldots,0)$ where the non-zero coordinate is the j^{th} coordinate. It is straightforward to show that the partial derivative $\frac{\partial}{\partial w_j}L_S$ is 0 for $x_t(k)$ with $j\neq k$, so the geometry of the surface of the objective function L_S is independent for each coordinate. Now, every Cartesian product of local minima in the one-dimensional setting form a d-dimensional local minimum. Since we have exactly two local minima, a good and another bad one in each coordinate, this combines into 2^d local minima, where each minimum's value would be the average of the one-dimensional minima forming it. Note that the combination of all good minima forms the global minimum with value ϵ. Following standard convention, we say that the data in this case is ϵ-*realizable* using a single neuron architecture. We stress that an important property of this initialization scheme is that the signs of the coordinates of the initialization point is uniformly distributed on the Boolean cube, as it implies that on each coordinate, independently, we have a probability 0.5 of reaching a bad basin, hence the number of bad basins we initialize from is distributed according to a Binomial distribution $B(d,0.5)$. Letting $c=\frac{L_0}{2}$, we have from Chernoff's bound that

$$\mathbb{P}\left[\text{Bas}\left(\mathbf{w}\right)\leq\frac{c}{4}\right]\leq e^{-\frac{d}{16}},$$

which concludes the proof of the theorem. □

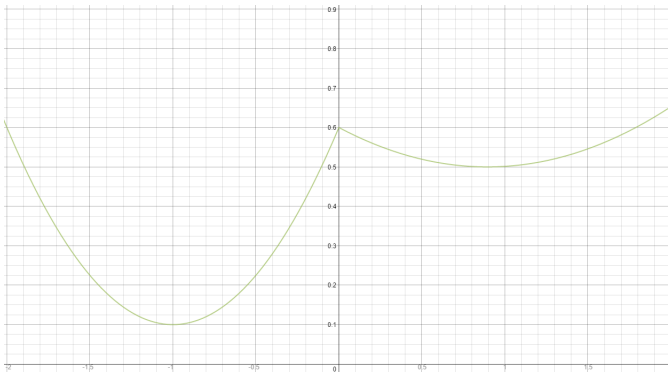

Figure 2. Plot of $L_S(w)$ for $\epsilon = 0.1$.

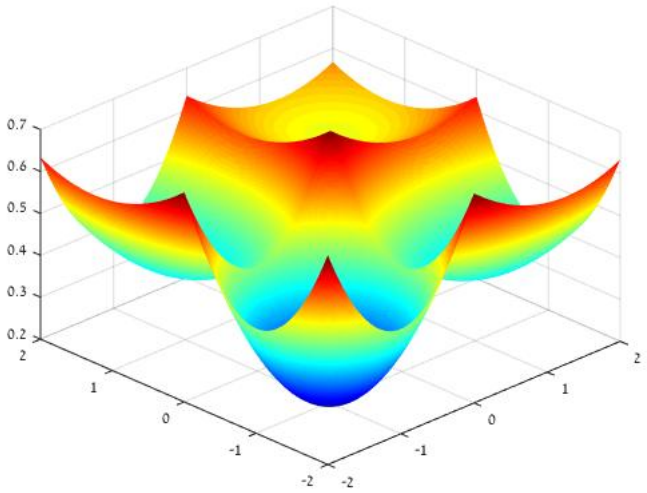

Figure 3. Plot of $L_S(w)$ after extending the sample to 2 dimensions. The surface contains one optimal minimum, another bad minimum and 2 average valued minima.

C.2. An Explicit Construction With the Squared Loss

We illustrate a specific construction of Thm. 6, for ReLU paired with the squared loss.

Define

$$\ell(y, y') = (y - y')^2.$$

Given $\epsilon > 0$, consider the following sample:

$$S = \left\{ \left(x_1 = \frac{1}{2}, y_1 = \sqrt{2\epsilon} \right), (x_2 = -1, y_2 = 1) \right\}.$$

Define for $i = 1, 2$

$$\ell_i(w) = \left([w x_i]_+ - y_i \right)^2,$$

and denote

$$L_S(w) = \frac{1}{2} \left(\ell_1(w) + \ell_2(w) \right).$$

Note that

$$L_S(-1) = \epsilon,$$

$$L_S\left(2\sqrt{2\epsilon}\right) = \frac{1}{2}$$

are both local minima, and thus S is ϵ-realizable. As evident in Fig. 2 and Fig. 3, if we are using a distribution corresponding to Assumption 1, then we have a 50% chance to initialize from the bad basin.

Extending the sample into a d-dimensional one as we did in Thm. 6, we have an ϵ-realizable dataset S with 2^d local minima. Furthermore, we have that

$$\mathbb{P}\left[\text{Bas}\left(\mathbf{w}\right) \leq \frac{1}{8}\right] \leq e^{-\frac{d}{16}}.$$