

A. Formal Statement of Results for General Piecewise Linear Activations

In §5, we stated our results in the case of ReLU activation, and now frame these results for a general piecewise linear non-linearity. We fix some notation. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous piecewise linear function with T breakpoints $\xi_0 = -\infty < \xi_1 < \xi_2 < \dots < \xi_T < \xi_{T+1} = \infty$. That is, there exist $p_j, q_j \in \mathbb{R}$ so that

$$t \in [\xi_j, \xi_{j+1}] \Rightarrow \phi(t) = q_j t + p_j, \quad q_j \neq q_{j+1}. \quad (11)$$

The analog of Theorem 3 for general ϕ is the following.

Theorem 6. *Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous piecewise linear function with T breakpoints $\xi_1 < \dots < \xi_T$ as in (11). Suppose \mathcal{N} is a fully connected network with input dimension n_{in} , output dimension 1, random weights and biases satisfying A1 and A2 above, and non-linearity ϕ .*

Let J_{z_1, \dots, z_k} be the $k \times n_{\text{in}}$ Jacobian of the map $x \mapsto (z_1(x), \dots, z_k(x))$,

$$\|J_{z_1, \dots, z_k}(x)\| := \det \left(J_{z_1, \dots, z_k}(x) (J_{z_1, \dots, z_k}(x))^T \right)^{1/2},$$

and write $\rho_{b_{z_1}, \dots, b_{z_k}}$ for the density of the joint distribution of the biases b_{z_1}, \dots, b_{z_k} . We say a neuron z is good at x if there exists a path of neurons from z to the output in the computational graph of \mathcal{N} so that each neuron \hat{z} along this path is open at x (i.e. $\phi'(\hat{z}(x) - b_{\hat{z}}) \neq 0$).

Then, for any bounded, measurable set $K \subseteq \mathbb{R}^{n_{\text{in}}}$ and any $k = 1, \dots, n_{\text{in}}$, the average $(n_{\text{in}} - k)$ -dimensional volume

$$\mathbb{E} [\text{vol}_{n_{\text{in}}-k}(\mathcal{B}_{\mathcal{N},k} \cap K)]$$

of $\mathcal{B}_{\mathcal{N},k}$ inside K is, in the notation of (6),

$$\sum_{\substack{\text{distinct neurons} \\ z_1, \dots, z_k \text{ in } \mathcal{N}}} \sum_{i_1, \dots, i_k=1}^T \int_K \mathbb{E} [Y_{z_1, \dots, z_k}^{(\xi_{i_1}, \dots, \xi_{i_k})}(x)] dx, \quad (12)$$

where $Y_{z_1, \dots, z_k}^{(\xi_{i_1}, \dots, \xi_{i_k})}(x)$ equals

$$\|J_{z_1, \dots, z_k}(x)\| \rho_{b_{z_1}, \dots, b_{z_k}}(z_1(x) - \xi_{i_1}, \dots, z_k(x) - \xi_{i_k}) \quad (13)$$

multiplied by the indicator function of the event that z_j is good at x for every j .

Note that if in the definition (11) of ϕ we have that the possible values $\phi'(t) \in \{q_0, \dots, q_T\}$ do not include 0, then we may ignore the event that z_j are good at x in the definition of $Y_{z_1, \dots, z_k}^{(\xi_{i_1}, \dots, \xi_{i_k})}$.

Corollary 7. *With the notation and assumptions of Theorem 6, suppose in addition that the weights and biases are independent. Fix $k \in \{1, \dots, n_{\text{in}}\}$ and suppose that for*

every collection of distinct neurons z_1, \dots, z_k , the average magnitude of the product of gradients is uniformly bounded:

$$\sup_{\substack{\text{neurons } z_1, \dots, z_k \\ \text{inputs } x}} \mathbb{E} \left[\prod_{j=1}^k \|\nabla z_j(x)\| \right] \leq C_{\text{grad}}^k. \quad (14)$$

Then we have the following upper bounds

$$\begin{aligned} & \frac{\mathbb{E} [\text{vol}_{n_{\text{in}}-k}(\mathcal{B}_{\mathcal{N},k} \cap K)]}{\text{vol}_{n_{\text{in}}}(K)} \\ & \leq \binom{\#\{\text{neurons}\}}{k} (T \cdot 2C_{\text{grad}}C_{\text{bias}})^k, \end{aligned} \quad (15)$$

where T is the number of breakpoints in the non-linearity ϕ of \mathcal{N} (see (11)) and

$$C_{\text{bias}} = \sup_z \sup_{b \in \mathbb{R}} \rho_{b_z}(b).$$

We prove Corollary 7 in §D and state a final corollary of Theorem 3:

Corollary 8. *Suppose \mathcal{N} is as in Theorem 3 and satisfies the hypothesis (14) in Corollary 7 with constants C_{bias} , C_{grad} . Then, for any compact set $K \subset \mathbb{R}^{n_{\text{in}}}$ let x be a uniform point in K . There exists $c > 0$ independent of K so that*

$$\mathbb{E} [\text{distance}(x, \mathcal{B}_{\mathcal{N}})] \geq \frac{c T}{C_{\text{bias}} C_{\text{grad}} \#\{\text{neurons}\}},$$

where, as before, T is the number of breakpoints in the non-linearity ϕ of \mathcal{N} .

We prove Corollary 8 in §E. The basic idea is simple. For every $\epsilon > 0$, we have

$$\mathbb{E} [\text{distance}(x, \mathcal{B}_{\mathcal{N}})] \geq \epsilon \mathbb{P} (\text{distance}(x, \mathcal{B}_{\mathcal{N}}) > \epsilon),$$

with the probability on the right hand side scaling like

$$1 - \text{vol}_{n_{\text{in}}}(T_{\epsilon}(\mathcal{B}_{\mathcal{N}}) \cap K) / \text{vol}_{n_{\text{in}}}(K),$$

where $T_{\epsilon}(\mathcal{B}_{\mathcal{N}})$ is the tube of radius ϵ around $\mathcal{B}_{\mathcal{N}}$. We expect that its volume like $\epsilon \text{vol}_{n_{\text{in}}-1}(\mathcal{B}_{\mathcal{N}})$. Taking $\epsilon = c / \#\{\text{neurons}\}$ yields the conclusion of Corollary 8.

B. Outline of Proof of Theorem 6

The purpose of this section is to give an intuitive explanation of the proof of Theorem 3. We fix a non-linearity $\phi : \mathbb{R} \rightarrow \mathbb{R}$ with breakpoints $\xi_1 < \dots < \xi_T$ (as in (11)) and consider a fully connected network \mathcal{N} with input dimension $n_{\text{in}} \geq 1$, output dimension 1, and non-linearity ϕ . For each neuron z in \mathcal{N} , we write

$$\ell(z) := \text{layer index of } z \quad (16)$$

and set

$$S_z := \{x \in \mathbb{R}^{n_{\text{in}}} \mid z(x) - b_z \in \{\xi_1, \dots, \xi_T\}\}. \quad (17)$$

We further

$$\tilde{S}_z := S_z \cap \mathcal{O}, \quad (18)$$

where

$$\mathcal{O} := \left\{ x \in \mathbb{R}^{n_{\text{in}}} \mid \forall j=1, \dots, d \exists \text{ neuron } z \text{ with } \ell(z)=j \text{ s.t. } \phi'(z(x)-b_z) \neq 0 \right\}.$$

Intuitively, the set S_z is the collection of inputs for which the neuron z turns from on to off. In contrast, the set \mathcal{O} is the collection of inputs $x \in \mathbb{R}^{n_{\text{in}}}$ for which \mathcal{N} is open in the sense that there is a path from the input to the output of \mathcal{N} so that all neurons along this path compute are not constant in a neighborhood x . Thus, \tilde{S}_z is the set of inputs at which neuron z switches between its linear regions and at which the output of neuron z actually affects the function computed by \mathcal{N} .

We remark here that $\mathcal{O} = \emptyset$ if in the non-linearity ϕ there are no linear pieces at which the slopes on ϕ equals 0 (i.e. $q_j \neq 0$ for all j in the definition (11) of ϕ). If, for example, ϕ is ReLU, then \mathcal{O} need not be empty.

The overall proof of Theorem 3 can be divided into several steps. The first gives the following representation of $\mathcal{B}_{\mathcal{N}}$.

Proposition 9. *Under Assumptions A1 and A2 of Theorem 3, we have, with probability 1,*

$$\mathcal{B}_{\mathcal{N}} = \bigcup_{\text{neurons } z} \tilde{S}_z.$$

The precise proof of Proposition 9 can be found in §C.1 below. The basic idea is that if for all y near a fixed input $x \in \mathbb{R}^{n_{\text{in}}}$, none of the pre-activations $z(y) - b_z$ cross the boundary of a linear region for ϕ , then $x \notin \mathcal{B}_{\mathcal{N}}$. Thus, $\mathcal{B}_{\mathcal{N}} \subset \bigcup_z S_z$. Moreover, if a neuron z satisfies $z(x) - b_z = S_i$ for some i but there are no open paths from z to the output of \mathcal{N} for inputs near x , then z is dead at x and hence does not influence \mathcal{N} at x . Thus, we expect the more refined inclusion $\mathcal{B}_{\mathcal{N}} \subset \bigcup_z \tilde{S}_z$. Finally, if $x \in \tilde{S}_z$ for some z then $x \in \mathcal{B}_{\mathcal{N}}$ unless the contribution from other neurons to $\nabla \mathcal{N}(y)$ for y near x exactly cancels the discontinuity in $\nabla z(x)$. This happens with probability 0.

The next step in proving Theorem 3 is to identify the portions of $\mathcal{B}_{\mathcal{N}}$ of each dimension. To do this, we write for any distinct neurons z_1, \dots, z_k ,

$$\tilde{S}_{z_1, \dots, z_k} := \bigcap_{j=1}^k \tilde{S}_{z_j}.$$

The set $\tilde{S}_{z_1, \dots, z_k}$ is, intuitively, the collection of inputs at which $z_j(x) - b_{z_j}$ switches between linear regions for ϕ and

at which the output of \mathcal{N} is affected by the post-activations of these neurons. Proposition 9 shows that we may represent $\mathcal{B}_{\mathcal{N}}$ as a disjoint union

$$\mathcal{B}_{\mathcal{N}} = \bigcup_{k=1}^{n_{\text{in}}} \mathcal{B}_{\mathcal{N},k},$$

where

$$\mathcal{B}_{\mathcal{N},k} := \bigcup_{\substack{\text{distinct neurons} \\ z_1, \dots, z_k}} \tilde{S}_{z_1, \dots, z_k} \cap \left(\bigcup_{z \neq z_1, \dots, z_k} \tilde{S}_z \right)^c.$$

In words, $\mathcal{B}_{\mathcal{N},k}$ is the collection of inputs in \mathcal{O} at which exactly k neurons turn from on to off. The following Proposition shows that $\mathcal{B}_{\mathcal{N},k}$ is precisely the “ $(n_{\text{in}} - k)$ -dimensional piece of $\mathcal{B}_{\mathcal{N}}$ ” (see (5)).

Proposition 10. *Fix $k = 1, \dots, n_{\text{in}}$, and k distinct neurons z_1, \dots, z_k in \mathcal{N} . Then, with probability 1, for every $x \in \mathcal{B}_{\mathcal{N},k}$ there exists a neighborhood in which $\mathcal{B}_{\mathcal{N},k}$ coincides with a $(n_{\text{in}} - k)$ -dimensional hyperplane.*

We prove Proposition 10 in §C.2. The idea is that each $\tilde{S}_{z_1, \dots, z_k}$ is piecewise linear and, with probability 1, at every point at which exactly the neurons z_1, \dots, z_k contribute to $\mathcal{B}_{\mathcal{N}}$, its co-dimension is the number of linear conditions needed to define it. Observe that with probability 1, the bias vector $(b_{z_1}, \dots, b_{z_{k+1}})$ for any collection z_1, \dots, z_{k+1} of distinct neurons is a regular value for $x \mapsto (z_1(x), \dots, z_{k+1}(x))$. Hence,

$$\text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_1, \dots, z_{k+1}}) = 0.$$

Proposition 10 thus implies that, with probability 1,

$$\text{vol}_{n_{\text{in}}-k}(\mathcal{B}_{\mathcal{N},k}) = \sum_{\substack{\text{distinct neurons} \\ z_1, \dots, z_k}} \text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_1, \dots, z_k}).$$

The final step in the proof of Theorem 3 is therefore to prove the following result.

Proposition 11. *Let z_1, \dots, z_k be distinct neurons in \mathcal{N} . Then, for any bounded, measurable $K \subset \mathbb{R}^{n_{\text{in}}}$,*

$$\begin{aligned} & \mathbb{E} \left[\text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_1, \dots, z_k}) \right] \\ &= \int_K \sum_{i_1, \dots, i_k=1}^T \mathbb{E} \left[Y_{z_1, \dots, z_k}^{(S_{i_1}, \dots, S_{i_k})}(x) \right] dx, \end{aligned}$$

where $Y_{z_1, \dots, z_k}^{(S_{i_1}, \dots, S_{i_k})}$ is defined as in (13).

We provide a detailed proof of Proposition 11 in §C.3. The intuition is that the image of the volume element dx under $x \mapsto z(x) - S_i$ is the volume element

$$\|J_{z_1, \dots, z_k}(x)\| dx$$

from (13). The probability of an infinitesimal neighborhood dx of x belonging to a $(n_{\text{in}} - k)$ -dimensional piece of $\mathcal{B}_{\mathcal{N}}$ is therefore the probability

$$\rho_{b_{z_1}, \dots, b_{z_k}}(z_1(x) - S_{i_1}, \dots, z_k(x) - S_{i_k}) \times \|J_{z_1, \dots, z_k}(x)\| dx$$

that the vector of biases $(b_{z_j}, j = 1, \dots, k)$ belongs to the image of dx under map $(z_j(x) - S_{i_j}, j = 1, \dots, k)$ for some collection of breakpoints S_{i_j} . The formal argument uses the co-area formula (see (29) and (30)).

C. Proof of Theorem 3

C.1. Proof of Proposition 9

Recall that the non-linearity $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is continuous and piecewise linear with T breakpoints $\xi_1 < \dots < \xi_T$, so that, with $\xi_0 = -\infty$, $\xi_{T+1} = \infty$, we have

$$t \in (\xi_i, \xi_{i+1}) \Rightarrow \phi(t) = q_i t + p_i$$

with $q_i \neq q_{i+1}$. For each $x \in \mathbb{R}^{n_{\text{in}}}$, write

$$\begin{aligned} Z_x^+ &:= \{z \mid z(x) - b_z \in (\xi_i, \xi_{i+1}) \text{ and } q_i \neq 0 \text{ for some } i\} \\ Z_x^- &:= \{z \mid z(x) - b_z \in (\xi_i, \xi_{i+1}) \text{ and } q_i = 0 \text{ for some } i\} \\ Z_x^0 &:= \{z \mid z(x) - b_z = \xi_i \text{ for some } i\} \end{aligned}$$

Intuitively, Z_x^+ are the neurons that, at the input x are open (i.e. contribute to the gradient of the output $\mathcal{N}(x)$) but do not change their contribution in a neighborhood of x , Z_x^- are the neurons that are closed, and Z_x^0 are the neurons that, at x , produce a discontinuity in the derivative of \mathcal{N} . Thus, for example, if $\phi = \text{ReLU}$, then

$$Z_x^* := \{z \mid \text{sgn}(z(x) - b_z) = *\}, \quad * \in \{+, -, 0\}.$$

We begin by proving that $\mathcal{B}_{\mathcal{N}} \subseteq \bigcup_z \tilde{S}_z$ by checking the contrapositive

$$\left(\bigcup_z \tilde{S}_z \right)^c \subseteq \mathcal{B}_{\mathcal{N}}^c. \quad (19)$$

Fix $x \in \left(\bigcup_z \tilde{S}_z \right)^c$. Note that Z_x^\pm are locally constant in the sense that there exists $\varepsilon > 0$ so that for all y with $\|y - x\| < \varepsilon$, we have

$$Z_x^- \subseteq Z_y^-, \quad Z_x^+ \subseteq Z_y^+, \quad Z_y^+ \cup Z_y^0 \subseteq Z_x^+ \cup Z_x^0. \quad (20)$$

Moreover, observe that if in the definition (11) of ϕ none of the slopes q_i equal 0, then $Z_y^- = \emptyset$ for every y . To prove (19), consider any path γ from the input to the output in the computational graph of \mathcal{N} . Such a path consists of $d + 1$ neurons, one in each layer:

$$\gamma = (z_\gamma^{(0)}, \dots, z_\gamma^{(d)}), \quad \ell(z_\gamma^{(j)}) = j.$$

To each path we may associate a sequence of weights:

$$w_\gamma^{(j)} := \text{weight connecting } z_\gamma^{(j-1)} \text{ to } z_\gamma^{(j)}, \quad j = 1, \dots, d.$$

We will also define

$$q_\gamma^{(j)}(x) := \sum_{i=0}^T q_i \mathbf{1}_{\{z_\gamma^{(x)} - b_{z_\gamma^{(j)}} \in (\xi_i, \xi_{i+1}]\}}.$$

For instance, if $\phi = \text{ReLU}$, then

$$q_\gamma^{(j)}(x) = \mathbf{1}_{\{z_\gamma^{(j)}(x) - b_{z_\gamma^{(j)}} \geq 0\}},$$

and in general only one term in the definition of $q_\gamma^{(j)}(x)$ is non-zero for each z . We may write

$$\mathcal{N}(y) = \sum_{i=1}^{n_{\text{in}}} y_i \sum_{\text{paths } \gamma: i \rightarrow \text{out}} \prod_{j=1}^d q_\gamma^{(j)}(y) w_\gamma^{(j)} + \text{constant}, \quad (21)$$

Note that if $x \in \left(\bigcup_z \tilde{S}_z \right)^c$, then for any path γ through a neuron $z \in Z_x^0$, we have

$$\exists j \text{ s.t. } z_\gamma^{(j)} \in Z_x^-.$$

This is an open condition in light of (20), and hence for all y in a neighborhood of x and for any path γ through a neuron $z \in Z_x^0$ we also have that

$$\exists j \text{ s.t. } z_\gamma^{(j)} \in Z_y^-.$$

Thus, since the summand in (21) vanishes identically if $\gamma \cap Z_y^- \neq \emptyset$, we find that for y in a neighborhood of any $x \in \left(\bigcup_z \tilde{S}_z \right)^c$ we may write

$$\mathcal{N}(y) = \sum_{i=1}^{n_{\text{in}}} y_i \sum_{\substack{\text{paths } \gamma: i \rightarrow \text{out} \\ \gamma \subset Z_x^+}} \prod_{j=1}^d q_\gamma^{(j)}(y) w_\gamma^{(j)} + \text{constant}. \quad (22)$$

But, again by (20), for any fixed x , all y in a neighborhood of x and each $z \in Z_x^+$, we have $z \in Z_y^+$ as well. Thus, in particular,

$$z(x) - b_z \in (\xi_i, \xi_{i+1}) \Rightarrow z(y) - b_z \in (\xi_i, \xi_{i+1}).$$

Thus, for y sufficiently close to x , we have for every path in the sum (22) that

$$q_\gamma^{(j)}(y) = q_\gamma^{(j)}(x).$$

Therefore, the partial derivatives $(\partial \mathcal{N} / \partial y_i)(y)$ are independent of y in a neighborhood of x and hence continuous at x . This proves (19). Let us now prove the reverse inclusion:

$$\bigcup_z \tilde{S}_z \subseteq \mathcal{B}_{\mathcal{N}} \quad (23)$$

Note that, with probability 1, we have

$$\text{vol}_{n_{\text{in}}-1}(S_{z_1} \cap S_{z_2}) = 0$$

for any pair of distinct neurons z_1, z_2 . Note also that since $x \mapsto \mathcal{N}(x)$ is continuous and piecewise linear, the set $\mathcal{B}_{\mathcal{N}}$ is closed. Thus, it is enough to show the slightly weaker inclusion

$$\bigcup_z \left(\tilde{S}_z \setminus \bigcup_{\hat{z} \neq z} S_{\hat{z}} \right) \subseteq \mathcal{B}_{\mathcal{N}} \quad (24)$$

since the closure of $\tilde{S}_z \setminus \bigcup_{\hat{z} \neq z} S_{\hat{z}}$ equals \tilde{S}_z . Fix a neuron z and suppose $x \in \tilde{S}_z \setminus \bigcup_{\hat{z} \neq z} S_{\hat{z}}$. By definition, we have that for every neuron $\hat{z} \neq z$, either

$$\hat{z} \in Z_x^+ \quad \text{or} \quad \hat{z} \in Z_x^-.$$

This has two consequences. First, by (20), the map $y \mapsto z(y)$ is linear in a neighborhood of x . Second, in a neighborhood of x , the set \tilde{S}_z coincides with S_z . Hence, combining these facts, near x the set \tilde{S}_z coincides with the hyperplane

$$\{x \mid z(x) - b_z = \xi_i\}, \quad \text{for some } i. \quad (25)$$

We may take two sequences of inputs y_n^+, y_n^- on opposite sides of this hyperplane so that

$$\lim_{n \rightarrow \infty} y_n^+ = \lim_{n \rightarrow \infty} y_n^- = x$$

and

$$\phi'(z(y_n^+) - b_z) = q_i, \quad \phi'(z(y_n^-) - b_z) = q_{i-1}, \quad \forall n,$$

where the index i the same as the one that defines the hyperplane (25). Further, since $\mathcal{B}_{\mathcal{N}}$ has co-dimension 1 (it is contained in the piecewise linear co-dimension 1 set $\bigcup_z S_z$, for example), we may also assume that $y_n^+, y_n^- \notin \mathcal{B}_{\mathcal{N}}$. Consider any path γ from the input to the output of the computational graph of \mathcal{N} passing through z (so that $z = z_\gamma^{(j)} \in \gamma$). By construction, for every n , we have

$$q_\gamma^{(j)}(y_n^+) \neq q_\gamma^{(j)}(y_n^-),$$

and hence, after passing to a subsequence, we may assume that the symmetric difference

$$Z_{y_n^+}^+ \Delta Z_{y_n^-}^+ \neq \emptyset \quad (26)$$

of the paths that contribute to the representation (21) for y_n^+, y_n^- is fixed and non-empty (the latter since it always contains z). For any $y \notin \mathcal{B}_{\mathcal{N}}$, we may write, for each $i = 1, \dots, n_{\text{in}}$

$$\frac{\partial \mathcal{N}}{\partial y_i}(y) = \sum_{\substack{\text{paths } \gamma: i \rightarrow \text{out} \\ \gamma \subset Z_y^+}} \prod_{j=1}^d q_\gamma^{(j)}(y) w_\gamma^{(j)}. \quad (27)$$

Substituting into this expression $y = y_n^\pm$, we find that there exists a non-empty collection Γ of paths from the input to the output of \mathcal{N} so that

$$\frac{\partial \mathcal{N}}{\partial y_i}(y_n^+) - \frac{\partial \mathcal{N}}{\partial y_i}(y_n^-) = \sum_{\gamma \in \Gamma} a_\gamma \prod_{j=1}^d c_\gamma^{(j)} w_\gamma^{(j)}$$

where

$$a_\gamma \in \{-1, 1\}, \quad c_\gamma^{(j)} \in \{q_0, \dots, q_T\}.$$

Note that the expression above is a polynomial in the weights of \mathcal{N} . Note also that, by construction, this polynomial is not identically zero due to the condition (26). There are only finitely many such polynomials since both a_γ and $c_\gamma^{(j)}$ range over a finite alphabet. For each such non-zero polynomial, the set of weights at which it vanishes has co-dimension 1. Hence, with probability 1, the difference $\frac{\partial \mathcal{N}}{\partial y_i}(y_n^+) - \frac{\partial \mathcal{N}}{\partial y_i}(y_n^-)$ is non-zero. This shows that the partial derivatives $\frac{\partial \mathcal{N}}{\partial y_i}$ are not continuous at x and hence that $x \in \mathcal{B}_{\mathcal{N}}$. \square

C.2. Proof of Proposition 10

Fix distinct neurons z_1, \dots, z_k and suppose $x \in \tilde{S}_{z_1, \dots, z_k}$ but not in \tilde{S}_z for any $z \neq z_1, \dots, z_k$. After relabeling, we may assume that they are ordered by layer index:

$$\ell(z_1) \leq \dots \leq \ell(z_k).$$

Since $x \in \mathcal{O}$, we also have that $x \notin S_z$ for any $z \neq z_1, \dots, z_k$. Thus, there exists a neighborhood U of x so $S_z \cap U = \emptyset$ for every $z \neq z_1, \dots, z_k$. Thus, there exists a neighborhood of x on which $y \mapsto z_1(y)$ is linear.

Hence, as explained near (25) above, \tilde{S}_{z_1} is a hyperplane near x . We now restrict our inputs to this hyperplane and repeat this reasoning to see that, near x , the set \tilde{S}_{z_1, z_2} is a hyperplane inside \tilde{S}_{z_1} and hence, near x , is the intersection of two hyperplanes in $\mathbb{R}^{n_{\text{in}}}$. Continuing in this way shows that in a neighborhood of x , the set $\tilde{S}_{z_1, \dots, z_k}$ is equal to the intersection of k hyperplanes in $\mathbb{R}^{n_{\text{in}}}$. Thus, $\tilde{S}_{z_1, \dots, z_k} \setminus \left(\bigcup_{z \neq z_1, \dots, z_k} \tilde{S}_z \right)^c$ is precisely the intersection of k hyperplanes in a neighborhood of each of its points. \square

C.3. Proof of Proposition 11

Let z_1, \dots, z_k be distinct neurons in \mathcal{N} , and fix a compact set $K \subset \mathbb{R}^{n_{\text{in}}}$. We seek to compute the mean of $\text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_1, \dots, z_k} \cap K)$, which we may rewrite as

$$\begin{aligned} & \int_{S_{z_1, \dots, z_k} \cap K} \mathbf{1}_{\left\{ \begin{smallmatrix} z_j \text{ is good at } x \\ j=1, \dots, k \end{smallmatrix} \right\}} d\text{vol}_{n_{\text{in}}-k}(x) \\ &= \sum_{i_1, \dots, i_k=1}^T \int_{S_{z_{i_1}, \dots, z_{i_k}} \cap K} \mathbf{1}_{\left\{ \begin{smallmatrix} z_j \text{ is good at } x \\ j=1, \dots, k \end{smallmatrix} \right\}} d\text{vol}_{n_{\text{in}}-k}(x), \end{aligned} \quad (28)$$

where we've set

$$S_{z_1, \dots, z_k}^{(\xi_{i_1}, \dots, \xi_{i_k})} = \{x \mid z_j(x) - b_{z_j} = \xi_{i_j}, \quad j = 1, \dots, k\}.$$

Note that the map $x \mapsto (z_1(x), \dots, z_k(x))$ is Lipschitz, and recall the co-area formula, which says that if $\psi \in L^1(\mathbb{R}^n)$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $m \leq n$ is Lipschitz, then

$$\int_{\mathbb{R}^m} \int_{g^{-1}(t)} \psi(x) \, \text{dvol}_{n-m}(x) dt \quad (29)$$

equals

$$\int_{\mathbb{R}^n} \psi(x) \|Jg(x)\| \, \text{dvol}_n(x), \quad (30)$$

where Jg is the $m \times n$ Jacobian of g and

$$\|Jg(x)\| = \det((Jg(x))(Jg(x))^T)^{1/2}.$$

We assumed that the biases b_{z_1}, \dots, b_{z_k} have a joint conditional density

$$\rho_{\mathbf{b}_z} = \rho_{b_{z_1}, \dots, b_{z_k}}$$

given all other weights and biases. The mean of the term in (28) corresponding to a fixed $\xi = (\xi_{i_1}, \dots, \xi_{i_k})$ over the conditional distribution of b_{z_1}, \dots, b_{z_k} is therefore

$$\int_{\mathbb{R}^k} d\mathbf{b} \rho_{\mathbf{b}_z}(\mathbf{b}) \int_{\{\mathbf{z}=\mathbf{b}\} \cap K} \mathbf{1}_{\left\{ \begin{smallmatrix} z_j \text{ is good at } x \\ j=1, \dots, k \end{smallmatrix} \right\}} \, \text{dvol}_{n_{\text{in}}-k}(x),$$

where we've abbreviated $\mathbf{b} = (b_1, \dots, b_k)$ as well as $\mathbf{z}(x) = (z_1(x), \dots, z_k(x))$. This can be rewritten as

$$\int_{\mathbb{R}^k} d\mathbf{b} \int_{\{\mathbf{z}=\mathbf{b}\} \cap K} \rho_{\mathbf{b}_z}(\mathbf{z}(x)-\xi) \mathbf{1}_{\left\{ \begin{smallmatrix} z_j \text{ is good at } x \\ j=1, \dots, k \end{smallmatrix} \right\}} \, \text{dvol}_{n_0-k}(x).$$

Thus, applying the co-area formula (29), (30) shows that the average of (28) over the conditional distribution of b_{z_1}, \dots, b_{z_k} is precisely

$$\int_K Y_{z_1, \dots, z_k}(x) \, dx.$$

Taking the average over the remaining weights and biases, we may commute the expectation $\mathbb{E}[\cdot]$ with the dx integral since the integrand is non-negative. This completes the proof of Proposition 11. \square

D. Proof of Corollary 7

We begin by proving the upper bound in (15). By Theorem 3, $\mathbb{E}[\text{vol}(\mathcal{B}_{\mathcal{N},k} \cap K)]$ equals

$$\sum_{\text{distinct neurons } z_1, \dots, z_k} \sum_{i_1, \dots, i_k=1}^T \int_K \mathbb{E} \left[Y_{z_1, \dots, z_k}^{(\xi_{i_1}, \dots, \xi_{i_k})}(x) \right] (x) dx,$$

where, as in (13), $Y_{z_1, \dots, z_k}^{(\xi_{i_1}, \dots, \xi_{i_k})}(x)$ is

$$\|J_{z_1, \dots, z_k}(x)\| \rho_{b_{z_1}, \dots, b_{z_k}}(z_1(x) - \xi_{i_1}, \dots, z_k(x) - \xi_{i_k})$$

times the indicator function of the event that z_j is good at x for every j . When the weights and biases of \mathcal{N} are independent, we may write $\rho_{b_{z_1}, \dots, b_{z_k}}(b_1, \dots, b_k)$ as

$$\prod_{j=1}^k \rho_{b_{z_j}}(b_j) \leq \left(\sup_{\text{neurons } z} \sup_{b \in \mathbb{R}} \rho_{b_z}(b) \right)^k = C_{\text{bias}}^k.$$

Hence,

$$Y_{z_1, \dots, z_k}(x) \leq C_{\text{bias}}^k \left(\det \left(J_{z_1, \dots, z_k}(x) (J_{z_1, \dots, z_k}(x))^T \right) \right)^{1/2}.$$

Note that

$$J_{z_1, \dots, z_k}(x) (J_{z_1, \dots, z_k}(x))^T = \text{Gram}(\nabla z_1(x), \dots, \nabla z_k(x)),$$

where for any $v_i \in \mathbb{R}^n$

$$\text{Gram}(v_1, \dots, v_k)_{i,j} = \langle v_i, v_j \rangle$$

is the associated Gram matrix. The Gram identity says that $\det \left(J_{z_1, \dots, z_k}(x) (J_{z_1, \dots, z_k}(x))^T \right)^{1/2}$ equals

$$\|\nabla z_1(x) \wedge \dots \wedge \nabla z_k(x)\|,$$

which is the k -dimensional volume of the paralleliped in $\mathbb{R}^{n_{\text{in}}}$ spanned by $\{\nabla z_j(x), j = 1, \dots, k\}$. We thus have

$$\det \left(J_{z_1, \dots, z_k}(x) (J_{z_1, \dots, z_k}(x))^T \right)^{1/2} \leq \prod_{j=1}^k \|\nabla z_j(x)\|.$$

The estimate (14) proves the upper bound (15). For the special case of $\phi = \text{ReLU}$ we use the AM-GM inequality and Jensen's inequality to write

$$\begin{aligned} \mathbb{E} \left[\prod_{j=1}^k \|\nabla z_j(x)\| \right] &\leq \mathbb{E} \left[\left(\frac{1}{k} \sum_{j=1}^k \|\nabla z_j(x)\| \right)^k \right] \\ &\leq \frac{1}{k} \sum_{j=1}^k \mathbb{E} \left[\|\nabla z_j\|^k \right]. \end{aligned}$$

Therefore, by Theorem 1 of Hanin & Nica (2018), there exist $C_1, C_2 > 0$ so that

$$\mathbb{E} \left[\prod_{j=1}^k \|\nabla z_j(x)\| \right] \leq \left(C_1 e^{C_2 \sum_{j=1}^d \frac{1}{n_j}} \right)^k.$$

This completes the proof of the upper bound in (15). To prove the power bound, lower bound in (15) we must argue in a different way. Namely, we will induct on k and use the following facts to prove the base case $k = 1$:

1. At initialization, for each fixed input x , the random variables $\{\mathbf{1}_{\{z(x) > b_z\}}\}$ are independent Bernoulli random variables with parameter $1/2$. This fact is proved in Proposition 2 of Hanin & Nica (2018). In particular, the event $\{z \text{ is good at } x\}$, which occurs when there exists a layer $j \in \ell(z) + 1, \dots, d$ in which $z(x) \leq b_z$ for every neuron, is independent of $\{z(x), b_z\}$ and satisfies

$$\mathbb{P}(z \text{ is good at } x) \geq 1 - \sum_{j=1}^d 2^{-n_j}. \quad (31)$$

2. At initialization, for each fixed input x , we have

$$\frac{1}{2} \mathbb{E}[z(x)^2] = \frac{\|x\|^2}{n_{\text{in}}} + \sum_{j=1}^{\ell(z)} \sigma_{b_j}^2, \quad (32)$$

where $\sigma_{b_j}^2 := \text{Var}[\text{biases at layer } j]$. This is Equation (11) in the proof of Theorem 5 from Hanin & Rolnick (2018).

3. At initialization, for every neuron z and each input x , we have

$$\mathbb{E}[\|\nabla z(x)\|^2] = 2. \quad (33)$$

This follows easily from Theorem 1 of Hanin (2018).

4. At initialization, for each $1 \leq j \leq n_{\text{in}}$ and every $x \in \mathbb{R}^{n_{\text{in}}}$

$$\mathbb{E} \left[\log \left(n_{\text{in}} \left(\frac{\partial z}{\partial x_j}(x) \right)^2 \right) \right] = -\frac{5}{2} \sum_{j=1}^{\ell(z)} \frac{1}{n_j} \quad (34)$$

plus $O\left(\sum_{j=1}^{\ell(z)} \frac{1}{n_j^2}\right)$, where n_j is the width of the j^{th} hidden layer and the implied constant depends only on the 4th moment of the measure μ according to which weights are distributed. This estimate follows immediately by combining Corollary 26 and Proposition 28 in Hanin & Nica (2018).

We begin by proving the lower bound in (15) when $k = 1$. We use (31) to see that $\mathbb{E}[\text{vol}_{n_{\text{in}}-1}(\mathcal{B}_{\mathcal{N}} \cap K)]$ is bounded below by

$$\left(1 - \sum_{j=1}^d 2^{-n_j}\right) \sum_{\text{neurons } z} \int_K \mathbb{E}[\|\nabla z(x)\| \rho_{b_z}(z(x))] dx.$$

Next, we bound the integrand. Fix $x \in \mathbb{R}^{n_{\text{in}}}$ and a parameter $\eta > 0$ to be chosen later. The integrand $\mathbb{E}[\|\nabla z(x)\| \rho_{b_z}(z(x))]$ is bounded below by

$$\begin{aligned} & \mathbb{E}[\|\nabla z(x)\| \rho_{b_z}(z(x)) \mathbf{1}_{\{|z(x)| \leq \eta\}}] \\ & \geq \left[\inf_{|b| \leq \eta} \rho_{b_z}(b) \right] \mathbb{E}[\|\nabla z(x)\| \mathbf{1}_{\{|z(x)| \leq \eta\}}], \end{aligned}$$

which is bounded below by

$$\left[\inf_{|b| \leq \eta} \rho_{b_z}(b) \right] \left[\mathbb{E}[\|\nabla z(x)\|] - \mathbb{E}[\|\nabla z(x)\| \mathbf{1}_{\{|z(x)| > \eta\}}] \right].$$

Using Cauchy-Schwarz, the term $\mathbb{E}[\|\nabla z(x)\| \mathbf{1}_{\{|z(x)| > \eta\}}]$ is bounded above by

$$\left(\mathbb{E}[\|\nabla z(x)\|^2] \mathbb{P}(|z(x)| > \eta) \right)^{1/2},$$

which using (33) and (32) together with Markov's inequality, is bounded above by

$$\frac{2}{\eta^{1/2}} \left(\frac{\|x\|^2}{n_{\text{in}}} + \sum_{j=1}^{\ell(z)} \sigma_{b_j}^2 \right)^{1/2}.$$

Next, using Jensen's inequality twice, we write

$$\begin{aligned} \log \mathbb{E}[\|\nabla z(x)\|] & \geq \frac{1}{2} \mathbb{E} \left[\log \left(\|\nabla z(x)\|^2 \right) \right] \\ & = \frac{1}{2} \mathbb{E} \left[\log \left(\sum_{j=1}^{n_{\text{in}}} \left(\frac{\partial z}{\partial x_j}(x) \right)^2 \right) \right] \\ & \geq \frac{1}{2} \mathbb{E} \left[\log \left(n_{\text{in}}^{1/2} \frac{\partial z}{\partial x_j}(x) \right)^2 \right] \\ & = -\frac{5}{4} \sum_{j=1}^{\ell(z)} \frac{1}{n_j} + O \left(\sum_{j=1}^{\ell(z)} \frac{1}{n_j^2} \right), \end{aligned}$$

where in the last inequality we applied (34). Putting this all together, we find that exists $c > 0$ so that

$$\mathbb{E}[\|\nabla z(x)\| \rho_{b_z}(z(x))] \geq c \left[\inf_{|b| \leq \eta} \rho_{b_z}(b) \right],$$

where

$$\eta \geq 4 \left(\frac{\|x\|^2}{n_{\text{in}}} + \sum_{j=1}^d \sigma_{b_j}^2 \right) e^{\frac{5}{4} \sum_{j=1}^d \frac{1}{n_j} + O\left(\sum_{j=1}^{\ell(z)} \frac{1}{n_j^2}\right)}.$$

In particular, we may take

$$\eta = \left(\frac{\sup_{x \in K} \|x\|^2}{n_{\text{in}}} + \sum_{j=1}^d \sigma_{b_j}^2 \right) e^{C \sum_{j=1}^d \frac{1}{n_j}}$$

for C sufficiently large. This completes the proof of the lower bound in (15) when $k = 1$. To complete the proof of Corollary 7, suppose we have proved the lower bound in (15) for all ReLU networks \mathcal{N} and all collections of $k - 1$ distinct neurons. We may assume after relabeling that the neurons z_1, \dots, z_k are ordered by layer index:

$$\ell(z_1) \leq \dots \leq \ell(z_k).$$

With probability 1, the set $S_{z_1} \subset \mathbb{R}^{n_{\text{in}}}$ is piecewise linear, co-dimension 1 with finitely many pieces, which we denote by P_α . We may therefore rewrite $\text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_1, \dots, z_k} \cap K)$ as

$$\sum_{\alpha} \text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_2, \dots, z_k} \cap P_\alpha \cap K).$$

We now define a new neural network \mathcal{N}_α , obtained by restricting \mathcal{N} to P_α . The input dimension for \mathcal{N}_α equals $n_{\text{in}} - 1$, and the weights and biases of \mathcal{N}_α satisfy all the assumptions of Corollary 7. We can now apply our inductive hypothesis to the $k - 1$ neurons z_2, \dots, z_k in \mathcal{N}_α and to the set $K \cap P_\alpha$. This gives

$$\begin{aligned} \mathbb{E} \left[\sum_{\alpha} \text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_2, \dots, z_k} \cap P_\alpha \cap K) \right] \\ \geq \left(\inf_z \inf_{|b| \leq \eta} \rho_{b_z}(b) \right)^{k-1} \mathbb{E}[\text{vol}_{n_{\text{in}}-1}(P_\alpha \cap K)]. \end{aligned}$$

Summing this lower bound over α yields

$$\begin{aligned} \mathbb{E}[\text{vol}_{n_{\text{in}}-k}(\tilde{S}_{z_1, \dots, z_k} \cap K)] \\ \geq \left(\inf_z \inf_{|b| \leq \eta} \rho_{b_z}(b) \right)^{k-1} \mathbb{E}[\text{vol}_{n_{\text{in}}-1}(\tilde{S}_{z_1} \cap K)]. \end{aligned}$$

Applying the inductive hypothesis once more completes the proof. \square

E. Proof of Corollary 8

We will need the following observation.

Lemma 12. *Fix a positive integer $n \geq 1$, and let $S \subseteq \mathbb{R}^n$ be a compact continuous piecewise linear submanifold with finitely many pieces. Define $S_0 = \emptyset$ and let S_k be the union of the interiors of all k -dimensional pieces of $S \setminus (S_0 \cup \dots \cup S_{k-1})$. Denote by $T_\varepsilon(X)$ the ε -tubular neighborhood of any $X \subset \mathbb{R}^n$. We have*

$$\text{vol}_n(T_\varepsilon(S)) \leq \sum_{k=0}^n \omega_{n-k} \varepsilon^{n-k} \text{vol}_k(S_k),$$

where $\omega_d := \text{volume of ball of radius 1 in } \mathbb{R}^d$.

Proof. Define d to be the maximal dimension of the linear pieces in S . Let $x \in T_\varepsilon(S)$. Suppose $x \notin T_\varepsilon(S_k)$ for all $k = 0, \dots, d-1$. Then the intersection of the ball of radius ε around s with S is a ball inside $S_d \cong U \subset \mathbb{R}^d$. Using the convexity of this ball, there exists a point y in S_d so that the vector $x - y$ is parallel to the normal vector to S_d at y . Hence, x belong to the normal ε -ball bundle $B_\varepsilon(N^*(S_d))$ (i.e. the union of the fiber-wise ε -balls in the normal bundle to S_d). Therefore, we have

$$\text{vol}_n(T_\varepsilon(S)) \leq \text{vol}_n(B_\varepsilon(N^*(S_d))) + \text{vol}_n(T_\varepsilon(S_{\leq d-1})),$$

where we abbreviated $S_{\leq d-1} := \bigcup_{k=0}^{d-1} S_k$. Using that

$$\begin{aligned} \text{vol}_n(B_\varepsilon(N^*(S_d))) &= \text{vol}_d(S_d) \text{vol}_{n-d}(B_\varepsilon(\mathbb{R}^{n-d})) \\ &= \text{vol}_d(S_d) \varepsilon^{n-d} \omega_{n-d} \end{aligned}$$

and repeating this argument $d - 1$ times completes the proof. \square

We are now ready to prove Corollary 2. Let $x \in K = [0, 1]^{n_{\text{in}}}$ be uniformly chosen. Then, for any $\varepsilon > 0$, using Markov's inequality and Lemma 12, we have

$$\begin{aligned} \mathbb{E}[\text{distance}(x, \mathcal{B}_\mathcal{N})] &\geq \varepsilon \mathbb{P}(\text{distance}(x, \mathcal{B}_\mathcal{N}) > \varepsilon) \\ &= \varepsilon (1 - \mathbb{P}(\text{distance}(x, \mathcal{B}_\mathcal{N}) \leq \varepsilon)) \\ &= \varepsilon (1 - \mathbb{E}[\text{vol}_{n_{\text{in}}}(T_\varepsilon(\mathcal{B}_\mathcal{N}))]) \\ &\geq \varepsilon \left(1 - \sum_{k=1}^{n_{\text{in}}} \omega_{n_{\text{in}}-k} \varepsilon^{n_{\text{in}}-k} \mathbb{E}[\text{vol}_{n_{\text{in}}-k}(\mathcal{B}_{\mathcal{N},k})] \right) \\ &\geq \varepsilon \left(1 - \sum_{k=1}^{n_{\text{in}}} (C_{\text{grad}} C_{\text{bias}} \varepsilon \#\{\text{neurons}\})^k \right) \\ &\geq \varepsilon (1 - C' C_{\text{grad}} C_{\text{bias}} \varepsilon \#\{\text{neurons}\}) \end{aligned}$$

for some $C' > 0$. Taking ε to be a small constant times $1/(C_{\text{grad}} \#\{\text{neurons}\})$ completes the proof. \square