Appendix for Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances

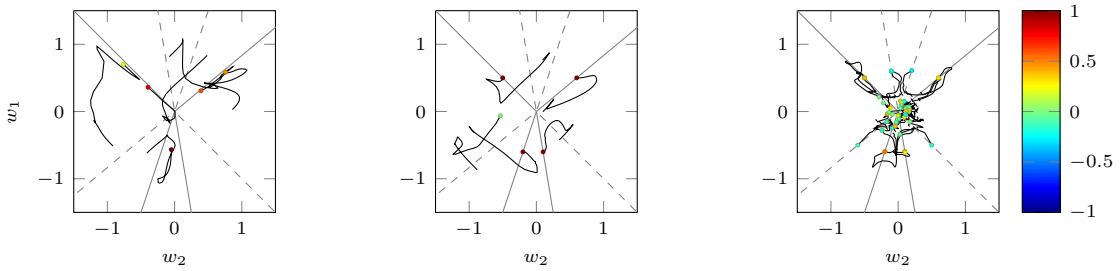
We organize the Appendix as follows:

- In Section A, we discuss the experimental details presented in the main text (and in the Appendix).
  - In Subsection A.1, we present numerical experiments on two-layer ANNs with various widths trained to implement the MNIST task. We investigate whether the gradient trajectories approach a saddle or not.
  - In Subsection A.2, we present a detailed numerical analysis of the number of critical subspaces G, the number of global minima subspaces T, and their ratio G/T.
  - In Subsection A.3, we present a catalog of toy symmetric loss landscapes.
- In Section B, we present proofs of the theorems and propositions stated in the main text.
  - In Subsection B.1, we present further properties of symmetric loss landscapes. In particular, we prove Lemma 2.1.
  - In Subsection B.2, we present the expansion manifold in two-layer ANNs. In particular, we prove the Theorem 3.1 in the main.
  - In Subsection B.3, we present a case where there is no new global minimum outside of the expansion manifold for some smooth activation functions. In particular we prove Theorem 4.2 in the main and discuss the implications for standard activation functions such as sigmoid and tanh.
  - In Subsection B.4, we present the symmetry-induced critical points. In particular, we prove the Proposition 4.3 and Proposition 4.4 in the main.
  - In Subsection B.5, we present the combinatorial analysis which is used to derive the closed-form formulas and the limiting behavior of the numbers T and G. In particular, we prove the Proposition 4.5, Lemma 4.6, and we present the calculations in the Subsection 4.3 in the main.
  - In Subsection B.6, we present some generalizations of the two-layer ANN results for the multi-layer ANNs.

A. Further Experimental Results

The code is available at `https://github.com/jbrea/SymmetrySaddles.jl`. We first present an extension of the Figure 1 in the main below.

Experimental details for the Figure 1 and 5 in the main. The input of the training data consists of 1681 two-dimensional points on a regular grid $\{(x_1, x_2)|4x_1 = -20, \dots, 20, 4x_2 = -20, \dots, 20\}$ and target values $y = \sum_{i=1}^{4} \sigma(\sum_{j=1}^{2} w_{ij} x_j)$ with $w_{11} = 0.6, w_{12} = 0.5, w_{21} = -0.5, w_{22} = 0.5, w_{31} = -0.2, w_{32} = -0.6, w_{41} = 0.1, w_{42} = -0.6$. Student networks were initialised with the Glorot uniform initialisation (Glorot & Bengio, 2010), trained with Adam (Kingma & Ba, 2014), and gradients always computed on the full dataset, until reaching a loss below 10^{-7}. To reach efficiently the local minimum closest to the point found with Adam, we continued optimizing the parameters of the student networks with the sequential quadratic programming algorithm SLSQP of the NLopt package (Johnson) for a maximal duration of 1000 seconds. The final loss values of all students that converged to a good solution was below 10^{-15} for every random seed considered. To obtain a non-trivial teacher network with 3 hidden layers (Fig. 1d-e), we fitted a network with widths 4-4-4 to the function $f(x_1, x_2) = \sin(2x_1) + x_1 + \cos(3x_2) - 0.4(x_2 - 1)^2$ evaluated on the same two-dimensional grid as above. The teacher network does not reach zero loss on this data set. To obtain target values for the student networks we evaluated this teacher network on the two-dimensional grid; hence there exist zero loss configurations for the student networks.

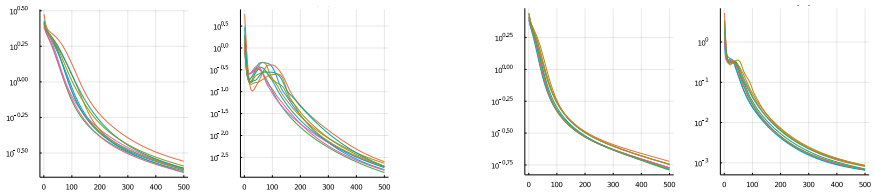


(a) **(b)** **(c)**

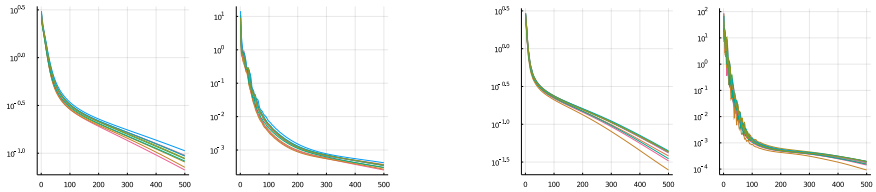
Figure 1. A student-teacher regression setting with 2D input and a two-layer teacher network with $r^* = 4$ sigmoid neurons with incoming weight vectors shown as solid black lines and output weights set to 1. black lines: trajectories of the incoming weight vectors with dots marking their position at convergence, color: output weights at convergence. For mild overparameterization, the algorithm may get stuck at a local minimum (a), or may find a global minimum (b); whereas for vast overparameterization it always converges to a global minimum (c). **(a&b)** Mildly overparameterized networks with width 5 do not reliably find the global minimum. **(c)** Vastly overparameterized networks with width 45 find a global minimum by setting some of the output weights to zero or matching the incoming weight vectors with that of the teacher's up to a $\pm$ factor.

A.1. MNIST Experiments for Two-Layer ANNs with Various Widths

Experimental details for the Figure 2 in the appendix. The training set consisted of the standard MNIST test set, i.e. 10'000 grayscale images of 28x28 pixels with corresponding labels. The networks had a single hidden layer of width N with the softplus non-linearity $g(x) = log(exp(x) + 1)$. The networks were initialised with the Glorot uniform initialisation (Glorot & Bengio, 2010) and trained on the cross-entropy loss with Adam and gradients always computed on the full dataset. We measured the squared norm of the gradients and the squared norm of the parameter updates.



(a) $m = 10$ **(b)** $m = 20$



(c) $m = 100$ **(d)** $m = 1000$

Figure 2. Network width m impacts whether gradient trajectories approach a saddle or not. For all a-b-c-d, the loss curves are demonstrated on the left and the norm of the gradient is demonstrated on the right. We observe that the norm of the gradient decreases and then increases in narrow networks **(a-b)**, indicating an approach to a saddle and then escaping it. We do not observe a sharp non-monotonicity in the norm of the gradient for wider networks **(c-d)**. Instead we observe short decrease and increase periods in the norm of the gradient (see the zigzag) **(d)**, which indicates that the gradient trajectories move from one saddle to the next in this regime, yet without getting very close these saddles.

For the MNIST experiments, we observe that the gradient trajectories visit a saddle in a narrow network and the duration of the visit to the saddles becomes shorter as we increase the width (i.e. in (a), we see a longer plateau in the loss curve compare to (b)). In the excessive overparameterization regime, we observe another behavior change, i.e. we observe a zigzag behavior on the norm of the gradient, possibly indicating many short visits to the saddles.

A.2. A Detailed Analysis of the Number of Critical Subspaces and the Number of Global Minima Subspaces

In this section, first we present a detailed numerical analysis of the numbers T and G (see Figure 3) and then we present additional figures for various minimal widths (see Figure 4), expanding the Figure 6 in the main.

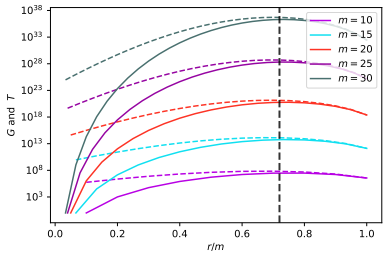
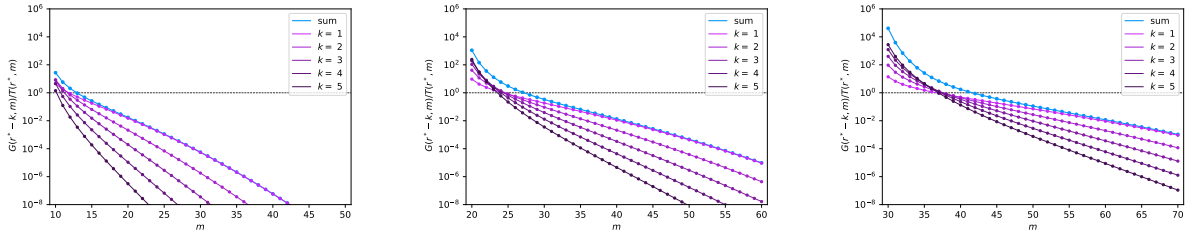


Figure 3. Comparison of the number of critical subspaces G (solid) with the number of global minima subspaces T (dashed) for $m = 10, 15, 20, 25, 30$ where x-axis denotes r/m. Each solid line indicates $G(r, m)$, and the dashed lines indicate $T(r, m)$ for $r = 1, \dots, m$. We observe that the maximum $G(r, m)$ is achieved at $r \approx 0.72m$.

In Figure 3, we observe an interesting linear relationship between r and m, i.e. for fixed m, $G(r, m)$ is maximized for $r \approx 0.7m$. A refined analysis of these numbers can be useful for studying how much overparameterization is needed to converge to a global minimum efficiently.



(a) $r^* = 10$ **(b)** $r^* = 20$ **(c)** $r^* = 30$

Figure 4. The ratio of the number of critical subspaces $G(r^* - k, m)$ to global minima subspaces $T(r^*, m)$ as the width m of the overparameterized network increases. Plotted for various minimal widths **(a)** $r^* = 10$, **(b)** $r^* = 20$, and **(c)** $r^* = 30$ (as shown in the main). The ratio of all critical subspaces to the global minima subspaces $\sum_{k=1}^{r^*-1} G(r^* - k, m)/T(r^*, m)$ is shown in blue.

In Figure 4, we observe that the rate of decay to zero is faster is smaller minimal widths (see for example $r^* = 10$). This is consistent with out mathematical analysis, since $\frac{r^*-1}{r^*}$ increases as r^* increases, yielding a slower decay to zero (see the blue curves). We note that the exact implementation of the numbers becomes unstable for $r^* > 35$ in our numerical experiments. Therefore for wider minimal widths, an approximation of the numbers G and T is needed.

A.3. Symmetric Loss Landscape Examples

We present some example symmetric losses $\mathbb{R}^2 \to \mathbb{R}$ in Figure 5, expanding Figure 2 in the main. We observe that in between two partner global minima (red points), there may be more than one saddles emerging on the symmetry subspaces.

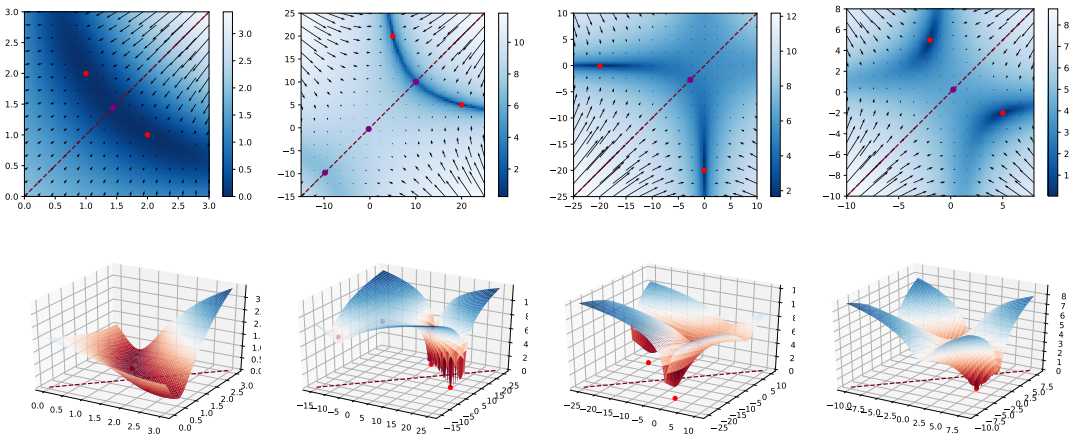


Figure 5. The gradient flow and the landscape and of a permutation-symmetric loss $L(w_1, w_2) = \log(\frac{1}{2}((w_1 + w_2 - a)^2 + (w_1 w_2 - b)^2) + 1)$. Red dots: global minima, purple dots: non-global stationary points. Dashed lines represent the symmetry hyperplanes. **1st.** $a = 3$ and $b = 2$, the global minima at $(2, 1)$ and $(1, 2)$. **2nd.** $a = 25$ and $b = 100$, the global minima at $(20, 5)$ and $(5, 20)$. **3rd.** $a = -20.1$ and $b = 2$, the global minima at $(-20, -0.1)$ and $(-0.1, -20)$. **4th.** $a = 3$ and $b = -10$, the global minima at $(5, -2)$ and $(-2, 5)$.

B. Proofs and Further Discussions

B.1. Further Properties of Symmetric Losses

The most well known property of symmetric losses is the $m!$ multiplicity of the critical points: for a critical point $\boldsymbol{\theta}^* = (\vartheta_1^*, \vartheta_2^*, \dots, \vartheta_m^*)$ with distinct units $\vartheta_i^* \neq \vartheta_j^*$ for all $i \neq j$, there are $m!$ equivalent critical points induced by permutations $\pi \in S_m$. Similarly, every point $\boldsymbol{\theta}$ with distinct units has $m! - 1$ partner points with equal loss. For a symmetric loss function, a fundamental region

$$\mathcal{R}_0 := \{(\vartheta_1, \dots, \vartheta_m) \in \mathbb{R}^{Dm} : \vartheta_1 \geq \dots \geq \vartheta_m\}$$

has $m!-1$ partner regions where the landscape of the loss is the same up to permutations. Note that above and elsewhere we use the lexicographic order: for two units $\vartheta, \vartheta' \in \mathbb{R}^D$, we write $\vartheta > \vartheta'$ if there exists $j \in [D]$ such that $\vartheta_i = \vartheta'_i$ for all $i \in [j-1]$ and $\vartheta_j > \vartheta'_j$; and $\vartheta = \vartheta'$, if $\vartheta_i = \vartheta'_i$ for all $i \in [D]$.

Definition B.1. *For a permutation $\pi \in S_m$, a* **replicant region** *$\mathcal{R}_\pi$ is defined by*

$$\mathcal{R}_\pi := \{(\vartheta_1, \dots, \vartheta_m) \in \mathbb{R}^{Dm} : \vartheta_{\pi(1)} \geq \dots \geq \vartheta_{\pi(m)}\}. \tag{1}$$

We denote by $\mathring{\mathcal{R}}_\pi$ the interior of the replicant region.

Any two partner points $\boldsymbol{\theta}_\pi \in \mathcal{R}_\pi$ and $\boldsymbol{\theta}_{\pi'} \in \mathcal{R}_{\pi'}$ have the same loss $L^m(\boldsymbol{\theta}_\pi) = L^m(\boldsymbol{\theta}_{\pi'})$ and they are linked with a permutation matrix $\mathcal{P}_{\pi' \circ \pi^{-1}} : \mathcal{P}_{\pi' \circ \pi^{-1}} \boldsymbol{\theta}_\pi = \boldsymbol{\theta}_{\pi'}$.

Note that the lexicographic order is a total order thus it allows to compare any two D-dimensional units. Therefore every point $\boldsymbol{\theta} \in \mathbb{R}^{Dm}$ falls in at least one replicant region, i.e.

$$\mathbb{R}^{Dm} = \cup_{\pi \in S_m} \mathcal{R}_\pi.$$

The intersection of all these regions $\mathcal{R}_\pi$ corresponds to the D-dimensional linear subspace $\vartheta_1 = \vartheta_2 = \dots = \vartheta_m$; more generally intersections of replicant regions define symmetry subspaces.

As each constraint $\vartheta_i = \vartheta_j$ suppresses D degrees of freedom, we have $\dim(\mathcal{H}_{i_1, \dots, i_k}) = D(m - k + 1)$. Observe that the largest symmetry subspaces are $\mathcal{H}_{i,j}$'s since any other symmetry subspace is included in one of these $\binom{m}{2}$ subspaces.

For $D = 1$, the largest symmetry subspaces have codimension 1. As a result, any path from $\mathcal{R}_\pi$ to any another replicant region has to cross a symmetry subspace (see Figure 6). However, for $D > 1$, the symmetry subspaces have codimension at least D; thus there exist paths connecting replicant regions without crossing symmetry subspaces.

Lemma B.1 (Lemma 2.1 in the main)**.** *Let $L^m : \mathbb{R}^{Dm} \to \mathbb{R}$ be a symmetric loss on m units thus a C^1 function and let $\boldsymbol{\rho} : \mathbb{R}_{\geq 0} \to \mathbb{R}^{Dm}$ be its gradient flow. If $\boldsymbol{\rho}(0) \in \mathcal{H}_{i_1,\dots,i_k}$, the gradient flow stays inside the symmetry subspace, i.e. $\boldsymbol{\rho}(t) \in \mathcal{H}_{i_1,\dots,i_k}$ for all $t > 0$. If $\boldsymbol{\rho}(0) \notin \mathcal{H}_{i,j}$ for all $i \neq j \in [m]$, that is outside of all symmetry subspaces, the gradient flow does not visit any symmetry subspace in finite time.*

Proof. We will use the identity that comes from chain rule $\nabla L^m(\mathcal{P}_\pi \boldsymbol{\theta}) = \mathcal{P}_\pi \nabla L^m(\boldsymbol{\theta})$. We will show that if $\boldsymbol{\theta} = (\vartheta_1, \cdots, \vartheta_m) \in \mathcal{H}_{i_1,\dots,i_k}$ where $\vartheta_{i_1} = \cdots = \vartheta_{i_k}$, its gradient satisfies $\nabla_{i_1} L^m(\boldsymbol{\theta}) = \ldots = \nabla_{i_k} L^m(\boldsymbol{\theta})$ therefore the gradient flow remains on the symmetry subspace for all times.

We denote a transposition by $(i, j) \in S_m$, which is a permutation that only swaps the units i and j. Assume $\boldsymbol{\theta} \in \mathcal{H}_{i,j}$, that is $\boldsymbol{\theta} = \mathcal{P}_{(i,j)}\boldsymbol{\theta}$, and thus

$$\nabla L^m(\boldsymbol{\theta}) = \nabla L^m(\mathcal{P}_{(i,j)}\boldsymbol{\theta}) = \mathcal{P}_{(i,j)} \nabla L^m(\boldsymbol{\theta}),$$

and in particular $\nabla_i L^m(\boldsymbol{\theta}) = \nabla_j L^m(\boldsymbol{\theta})$. This entails that for $\boldsymbol{\theta} \in \mathcal{H}_{i_1,\dots,i_k}$, we have $\nabla L^m(\boldsymbol{\theta}) \in \mathcal{H}_{i_1,\dots,i_k}$ as well, which completes the first part of the proof.

We now prove the second part of the claim by contradiction. Suppose now that $\boldsymbol{\gamma}(0) \notin \mathcal{H}_{i,j}$ for any $i \neq j \in [k]$ and $t_0 < \infty$ be the first time such that $\boldsymbol{\gamma}(t_0) \in \mathcal{H}_{i',j'}$ for some $i' \neq j' \in [k]$. Let $\widetilde{\boldsymbol{\gamma}}(t) = \mathcal{P}_{(i',j')}\boldsymbol{\gamma}(t)$, that is the symmetric path with respect to $\mathcal{H}_{i',j'}$. Then one sees that $\boldsymbol{\gamma}$ and $\widetilde{\boldsymbol{\gamma}}$ intersect for the first time at t_0 on $\mathcal{H}_{i',j'}$ and then $\boldsymbol{\gamma}(t) = \widetilde{\boldsymbol{\gamma}}(t) \in \mathcal{H}_{i',j'}$ for all $t > t_0$, as we showed in the first part of the proof. Since ∇L^m is continuous, Picard-Lindelöf Theorem applies on a neighbourhood of $\boldsymbol{\gamma}(t_0)$, which ensures the unicity of the gradient flow on $[t_0 - \epsilon, t_0]$ for some $\epsilon > 0$. Thus, $\boldsymbol{\gamma}(t_0 - \epsilon) = \widetilde{\boldsymbol{\gamma}}(t_0 - \epsilon)$, which contradicts the fact that t_0 is the first time when $\boldsymbol{\gamma}$ intersects $\widetilde{\boldsymbol{\gamma}}$.

□

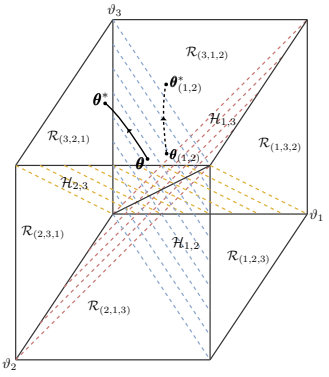


Figure 6. Replicant regions $\mathcal{R}_\pi$ and symmetry subspaces $\mathcal{H}_{i,j}$ for the 3-dimensional parameter space $\mathbb{R}^3$. An example gradient flow trajectory starting at $\boldsymbol{\theta} \in \mathcal{R}_{(3,2,1)}$ and arriving at a minimum $\boldsymbol{\theta}^*$ (solid curve) and its partner trajectory starting at a partner point $\boldsymbol{\theta}_{(1,2)} \in R_{(3,1,2)}$ thus arriving at a partner minimum $\boldsymbol{\theta}^*_{(1,2)}$ (dashed curve) are shown.

We write the gradient of L^m in the block form

$$\nabla L^m(\boldsymbol{\theta}) = (\nabla_1 L^m(\boldsymbol{\theta}), \dots, \nabla_m L^m(\boldsymbol{\theta}))$$

where for all $j \in [m]$,

$$\nabla_j L^m(\boldsymbol{\theta}) = (\partial_{D(j-1)+1} L^m(\boldsymbol{\theta}), \dots, \partial_{D(j-1)+D} L^m(\boldsymbol{\theta}))$$

is a D-dimensional vector.

Remark B.1. *Let $\boldsymbol{\rho}(0) \in \mathcal{R}_\pi$ for some $\pi \in S_m$. In the case of 1-dimensional units, $D = 1$, we have $\boldsymbol{\rho}(t) \in \mathcal{R}_\pi$ for all $t \in \mathbb{R}_+$. Hence, in this case, the gradient flow can only be affected by the critical points of a single replicant region.*

Proof. Indeed, assume that $\boldsymbol{\rho}(0) = (\vartheta_1(0), \dots, \vartheta_m(0)) \in \mathcal{R}_\pi$, i.e. $\vartheta_{\pi_1}(0) \geq \cdots \geq \vartheta_{\pi_m}(0)$ and $\boldsymbol{\rho}(1) = (\vartheta_1(1), \dots, \vartheta_m(1)) \in \mathcal{R}_{\pi'}$ for another permutation π', i.e. $\vartheta_{\pi'_1}(0) \geq \cdots \geq \vartheta_{\pi'_m}(0)$. Since $\pi \neq \pi'$, there exists a pair (i, j) such that $\vartheta_i(0) \geq \vartheta_j(0)$ and $\vartheta_j(1) \geq \vartheta_i(1)$. Thus we have

$$(\vartheta_i - \vartheta_j)(0) \geq 0 \geq (\vartheta_i - \vartheta_j)(1).$$

Because the gradient flow $\boldsymbol{\rho}$ is continuous (since L^m is C^1) there exists a time t_0 such that $(\vartheta_i - \vartheta_j)(t_0) = 0$, i.e. $\boldsymbol{\rho}(t_0) \in \mathcal{H}_{i,j}$, which yields a contradiction. □

Remark B.2. *In the case of 1-dimensional units, $D = 1$, if $\boldsymbol{\rho}(0) \in \mathcal{R}_\pi$ for some $\pi \in S_m$, we have $\boldsymbol{\rho}(t) \in \mathcal{R}_\pi$ for all $t \in \mathbb{R}_+$. Hence, in this case, the gradient flow $\boldsymbol{\rho}$ can only be affected by the critical points of a single replicant region.*

B.2. The Expansion Manifold in Two-Layer ANNs

Theorem B.2 (Theorem 3.1 in the main)**.** *For $m \geq r$, the expansion manifold $\Theta_{r\to m}(\boldsymbol{\theta}^r)$ of an irreducible point $\boldsymbol{\theta}^r$ consists of exactly*[1]

$$T(r,m) := \sum_{j=0}^{m-r} \sum_{\substack{sum(s)=m \\ k_i\geq 1, b_i \geq 1}} \binom{m}{k_1,...,k_r,b_1,...,b_j} \frac{1}{c_1!...c_{m-r}!}$$

distinct affine subspaces (none is including another one) of dimension at least $\min(d_{in}, d_{out})(m-r)$, where c_i is the number of occurences of i among $(b_1,...,b_j)$.

For $m > r$, $\Theta_{r\to m}(\boldsymbol{\theta}^r)$ is connected: any pair of distinct points $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{r\to m}(\boldsymbol{\theta}^r)$ is connected via a union of line segments $\boldsymbol{\gamma} : [0,1] \to \Theta_{r\to m}(\boldsymbol{\theta}^r)$ such that $\boldsymbol{\gamma}(0) = \boldsymbol{\theta}$ and $\boldsymbol{\gamma}(1) = \boldsymbol{\theta}'$.

Proof. The proof of this theorem is divided in two parts. In Proposition B.3, we count the number of affine subspaces in $\Theta_{r\to m}(\boldsymbol{\theta}^r)$ and in Theorem B.4, we prove the connectivity of the $r \to m$ expansion manifold for $m > r$. □

Proposition B.3. *For $m \geq r$, $\Theta_{r\to m}(\boldsymbol{\theta}^r)$ has exactly*

$$T(r,m) := \sum_{j=0}^{m-r} \sum_{\substack{sum(s)=m \\ k_i\geq 1, b_i \geq 1}} \binom{m}{k_1,\ldots,k_r,b_1,\ldots,b_j} \frac{1}{c_b}$$

distinct affine subspaces (none is including another one) of dimension at least $\min(d_{in}, d_{out})(m-r)$. Here $c_b := c_1!c_2!\cdots c_{m-r}!$ is a normalization factor where c_i is the number of occurence of i among $(b_1,\ldots,b_j)$.

Proof. The dimension of the subspace $\Gamma_s(\boldsymbol{\theta}^r)$ is

$$\sum_{t=1}^{r}(k_t-1)d_{\text{out}} + \sum_{t=1}^{j}(b_t-1)d_{\text{out}} + jd_{\text{in}} = (m-r-j)\,d_{\text{out}} + j\,d_{\text{in}} \geq \min(d_{\text{in}}, d_{\text{out}})(m-r).$$

It is enough to count the distinct configurations of the incoming weight vectors

$$(\underbrace{w_1,\ldots,w_1}_{k_1},\ldots,\underbrace{w_r,\ldots,w_r}_{k_r},\underbrace{w'_1,\ldots,w'_1}_{b_1},\ldots,\underbrace{w'_j,\ldots,w'_j}_{b_j})$$

since the outgoing weight vectors configuration follows that of the incoming ones. For this particular tuple, the number of configurations is $\frac{m!}{k_1!\cdots k_r! b_1!\cdots b_j!}\frac{1}{c_b}$ where the normalization factor $c_b = c_1!\cdots c_{m-r}!$ comes from the following sub-configurations: if $b_1 = b_2$, then we need to divide by 2 since in that case one can swap w'_1 with w'_2. More generally, if $b_{\ell_1} = b_{\ell_2} = \ldots = b_{\ell_{c_i}} = i$, we need to divide by the number of permutations between the groups of i incoming weight vectors

$$\underbrace{(\underbrace{w'_{\ell_1},\ldots,w'_{\ell_1}}_{i},\ldots,\underbrace{w'_{\ell_{c_i}},\ldots,w'_{\ell_{c_i}}}_{i})}_{c_i}.$$

There are c_i groups with the repetition of i zero-type incoming weight vectors (such that their summation is fixed at zero) thus we have to cancel out the recounting coming from these groups via a division by $1/c_i!$. Summing over all possible tuples $(k_1,\ldots,k_r,b_1,\ldots,b_j)$, we find the formula. □

Theorem B.4. *For $m > r$, $\Theta_{r\to m}(\boldsymbol{\theta}^r)$ is connected: any pair of distinct points $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{r\to m}(\boldsymbol{\theta}^r)$ is connected via a union of line segments $\boldsymbol{\gamma} : [0,1] \to \Theta_{r\to m}(\boldsymbol{\theta}^r)$ such that $\boldsymbol{\gamma}(0) = \boldsymbol{\theta}$ and $\boldsymbol{\gamma}(1) = \boldsymbol{\theta}'$.*

[1] $\binom{n_1+\ldots+n_r}{n_1,\ldots,n_r}$ denotes the coefficient $\frac{(n_1+\ldots+n_r)!}{n_1!\ldots n_r!}$.

Proof. We first prove the case $m = r + 1$. Let $\boldsymbol{\theta}^r = (w_1, \ldots, w_r, a_1, \ldots, a_r)$ and consider the following set of points

$$\widetilde{\Theta}_{r\to r+1}(\boldsymbol{\theta}^r) := \{\mathcal{P}_\pi \boldsymbol{\theta}^{r+1} : \boldsymbol{\theta}^{r+1} = (w_0, w_1, w_2, \ldots, w_r, 0, a_1, a_2, \ldots, a_r);\ \pi \in S_r,\ w_0 \in \mathbb{R}^{d_{\text{in}}}\}$$

which is a subset of the expansion manifold $\Theta_{r\to r+1}(\boldsymbol{\theta}^r)$. We will show that by construction that a point $\boldsymbol{\theta}_0 \in \widetilde{\Theta}_{r\to r+1}(\boldsymbol{\theta}^r)$ such that $\boldsymbol{\theta}_0 = (w_0, w_1, w_2, \ldots, w_r, 0, a_1, a_2, \ldots, a_r)$ is connected to any other point $\widetilde{\boldsymbol{\theta}} = \mathcal{P}_\pi \boldsymbol{\theta}_0 \in \widetilde{\Theta}_{r\to r+1}(\boldsymbol{\theta}^r)$ via a path in $\Theta_{r\to r+1}(\boldsymbol{\theta}^r)$. To do so we first show that a neighbor where the neuron $\vartheta_0 = (w_0, 0)$ is swapped with $\vartheta_i = (w_i, a_i)$

$$\boldsymbol{\theta}_1 = (w_i, w_1, \ldots, w_{i-1}, w_0, w_{i+1}, \ldots, w_r, a_i, a_1, \ldots, a_{i-1}, 0, a_{i+1}, \ldots, a_r)$$

can be reached in three steps using the following line segments $\boldsymbol{v}_1^{(1)}, \boldsymbol{v}_2^{(1)}, \boldsymbol{v}_3^{(1)} : [0, 1] \to \Theta_{r\to r+1}(\boldsymbol{\theta}^r)$

$$\begin{aligned}
\boldsymbol{v}_1^{(1)}(\alpha) &= (\alpha(w_i - w_0) + w_0, w_1, w_2, \ldots, w_r, 0, a_1, a_2, \ldots, a_r)\\
\boldsymbol{v}_2^{(1)}(\alpha) &= (w_i, w_1, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_r, \alpha a_i, a_1, \ldots, a_{i-1}, (1-\alpha)a_i, a_{i+1}, \ldots, a_r)\\
\boldsymbol{v}_3^{(1)}(\alpha) &= (w_i, w_1, \ldots, w_{i-1}, \alpha(w_0 - w_i) + w_i, w_{i+1}, \ldots, w_r, a_i, a_1, \ldots, a_{i-1}, 0, a_{i+1}, \ldots, a_r)
\end{aligned}$$

where we have $\boldsymbol{v}_1^{(1)}(0) = \boldsymbol{\theta}_0$, $\boldsymbol{v}_1^{(1)}(1) = \boldsymbol{v}_2^{(1)}(0)$, $\boldsymbol{v}_2^{(1)}(1) = \boldsymbol{v}_3^{(1)}(0)$, and $\boldsymbol{v}_3^{(1)}(1) = \boldsymbol{\theta}_1$. In particular, we constructed a path $\boldsymbol{\gamma}^{(1)}$ by glueing three line segments at their end points

$$\boldsymbol{\gamma}^{(1)}(t) = \boldsymbol{v}_1^{(1)}(3t)\mathbb{1}_{t\in[0,1/3)} + \boldsymbol{v}_2^{(1)}(3(t-1/3))\mathbb{1}_{t\in[1/3,2/3)} + \boldsymbol{v}_3^{(1)}(3(t-2/3))\mathbb{1}_{t\in[2/3,1]}$$

where $\boldsymbol{\gamma}^{(1)}(0) = \boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}^{(1)}(1) = \boldsymbol{\theta}_1$. Note that going from $\boldsymbol{\theta}_0 \to \boldsymbol{\theta}_1$, we swapped the neurons ϑ_0 and ϑ_i. Moreover, it is well known that any permutation can be written as a composition of transpositions (permutations leaving all elements unchanged but two) and that $(i\ j) = (0\ j) \circ (0\ i) \circ (0\ j)$. In particular, we can reach $\widetilde{\boldsymbol{\theta}}$ only by swapping ϑ_0 with other neurons, which corresponds to some other paths $\boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(r)}$ made of three line segments. Glueing these paths, we observe that $\widetilde{\Theta}_{r\to r+1}(\boldsymbol{\theta}^r)$ is connected via paths in $\Theta_{r\to r+1}(\boldsymbol{\theta}^r)$. To finish the case for $m = r + 1$, it is enough to show that any point $\boldsymbol{\theta} \in \Theta_{r\to r+1}(\boldsymbol{\theta}^r) \setminus \widetilde{\Theta}_{r\to r+1}(\boldsymbol{\theta}^r)$

$$\boldsymbol{\theta} = \mathcal{P}_\pi(w_i, w_i, w_1, \ldots, w_r, \alpha a_i, (1-\alpha)a_i, a_1, \ldots, a_r)$$

is connected (via a line segment) to a point in $\widetilde{\Theta}_{r\to r+1}(\boldsymbol{\theta}^r)$ which is simply

$$\widetilde{\boldsymbol{\theta}} = \mathcal{P}_\pi(w_0, w_i, w_1, \ldots, w_r, 0, a_i, a_1, \ldots, a_r).$$

Next we will prove for the general case $m \geq r + 1$ by induction. We assume that $\Theta_{r\to m}(\boldsymbol{\theta}^r)$ is connected and we will show that $\Theta_{r\to m+1}(\boldsymbol{\theta}^r)$ is also connected. First we show the connectivity of the points in the following set

$$\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r) := \{\mathcal{P}_\pi \boldsymbol{\theta}^{m+1} : \boldsymbol{\theta}^{m+1} = (\underbrace{w_1, \ldots, w_1}_{k_1}, \ldots, \underbrace{w_r, \ldots, w_r}_{k_r}, \underbrace{w_1', \ldots, w_j'}_{j+1}, \underbrace{a_1^1, \ldots a_1^{k_1}}_{k_1}, \ldots, \underbrace{a_r^1, \ldots a_r^{k_r}}_{k_r}, \underbrace{0, \ldots, 0}_{j+1})$$
$$\text{where } k_i \geq 1, j \geq 0, k_1 + \ldots + k_r + j = m, \sum_{i=1}^{k_j} a_j^i = a_j, \text{and } \pi \in S_{m+1}\}$$

which is a subset of $\Theta_{r\to m+1}(\boldsymbol{\theta}^r)$. From the induction hypothesis, we have the connectivity of the manifold $\Theta_{r\to m}(\boldsymbol{\theta}^r)$. An element $\widetilde{\boldsymbol{\theta}} \in \widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$ can be written as

$$\widetilde{\boldsymbol{\theta}} = \mathcal{P}_{\tilde{\pi}}(\underbrace{w_1, \ldots, w_1}_{k_1}, \ldots, \underbrace{w_r, \ldots, w_r}_{k_r}, \underbrace{w_1', \ldots, w_j'}_{j}, \underbrace{w_0}_{1}, \underbrace{a_1^1, \ldots a_1^{k_1}}_{k_1}, \ldots, \underbrace{a_r^1, \ldots a_r^{k_r}}_{k_r}, \underbrace{0, \ldots, 0}_{j+1}),$$

for some $j \geq 0$ and $\tilde{\pi} \in S_{m+1}$. For a fixed w_0 at a fixed position, there is a bijection $\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r) \to \Theta_{r\to m}(\boldsymbol{\theta}^r)$ that sends $\widetilde{\boldsymbol{\theta}}$ to

$$\boldsymbol{\theta} = \mathcal{P}_\pi(\underbrace{w_1, \ldots, w_1}_{k_1}, \ldots, \underbrace{w_r, \ldots, w_r}_{k_r}, \underbrace{w_1', \ldots, w_j'}_{j}, \underbrace{a_1^1, \ldots a_1^{k_1}}_{k_1}, \ldots, \underbrace{a_r^1, \ldots a_r^{k_r}}_{k_r}, \underbrace{0, \ldots, 0}_{j})$$

for some $\pi \in S_m$, i.e. $\widetilde{\boldsymbol{\theta}}$ where w_0 and its associated 0 outgoing weight vector have been dropped. In particular, any two points of $\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$ with the same w_0 component at the same position are connected as a consequence of this correspondence and the connectivity of $\Theta_{r\to m}(\boldsymbol{\theta}^r)$. Moreover, we note that $\widetilde{\boldsymbol{\theta}} \in \widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$ is connected via a line segment in $\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$ to every other point in $\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$ whose components are the same as $\widetilde{\boldsymbol{\theta}}$ except for w_0. This straightforwardly generalizes for different positions of w_0 and this establishes the connectivity of $\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$.

Finally, we pick a point $\boldsymbol{\theta} \in \Theta_{r\to m+1}(\boldsymbol{\theta}^r)$ that is

$$\boldsymbol{\theta} = \mathcal{P}_\pi(\underbrace{w_1,\dots,w_1}_{k_1},\dots,\underbrace{w_r,\dots,w_r}_{k_r},\underbrace{w'_1,\dots,w'_1}_{b_1},\dots,\underbrace{w'_j,\dots,w'_j}_{b_j},\underbrace{a^1_1,\dots a^{k_1}_1}_{k_1},\dots,\underbrace{a^1_r,\dots a^{k_r}_r}_{k_r},\underbrace{\alpha^1_1,\dots,\alpha^{b_1}_1}_{b_1},\dots,\underbrace{\alpha^1_j,\dots,\alpha^{b_j}_j}_{b_j}).$$

for some $\pi \in S_{m+1}$. Note that $\boldsymbol{\theta}$ is connected to

$$\widetilde{\boldsymbol{\theta}} = \mathcal{P}_\pi(\underbrace{w_1,\dots,w_1}_{k_1},\dots,\underbrace{w_r,\dots,w_r}_{k_r},\underbrace{w'_1,\dots,w'_1}_{b_1},\dots,\underbrace{w'_j,\dots,w'_j}_{b_j},\underbrace{a^1_1,\dots a^{k_1}_1}_{k_1},\dots,\underbrace{a^1_r,\dots a^{k_r}_r}_{k_r},\underbrace{0,\dots,0}_{b_1},\dots,\underbrace{0,\dots,0}_{b_j}),$$

which is in $\widetilde{\Theta}_{r\to m+1}(\boldsymbol{\theta}^r)$. We have shown that all points in $\Theta_{r\to m+1}(\boldsymbol{\theta}^r)$ are connected, which completes the induction step thus the proof.

□

B.3. No New Global Minimum

The following assumption, only made in Theorem **??**, ensures that the activation function σ has no specificity that yields other invariances than the symmetries between units, e.g. σ cannot be even or odd.

Assumption A. *Let σ be a smooth activation function. We suppose that $\sigma(0) \neq 0$, that $\sigma^{(n)}(0) \neq 0$ for infinitely many even and odd values of $n \geq 0$, where $\sigma^{(n)}$ denotes the n-th derivative.*

The next lemma contains the main argument to prove that when considering an overparametrized 2-layers neural network, no new global minima are created besides those coming from invariances.

Lemma B.5. *Suppose that the activation function σ satisfies the Assumption A. If for some pairwise distinct nonzero $\beta_1,\dots,\beta_k \in \mathbb{R}$ and some constant $c \in \mathbb{R}$ we have $g(\alpha) := \sum_{\ell=1}^k a_\ell \sigma(\alpha\beta_\ell) = c$ for all $\alpha \in \mathbb{R}$, then $a_\ell = 0$ for all $\ell \in [k]$.*

Proof. We reorder the indices such that for all $\ell \in [k-1]$, either $|\beta_\ell| > |\beta_{\ell+1}|$, or $\beta_\ell = -\beta_{\ell+1}$ such that $|a_\ell| \geq |a_{\ell+1}|$ (if the equality holds the labelling between the two is not important). We distinguish the four following cases:

1. $|\beta_1| > |\beta_2|$,
2. $\beta_1 = -\beta_2$ and $|a_1| > |a_2|$,
3. $\beta_1 = -\beta_2$ and $a_1 = a_2$,
4. $\beta_1 = -\beta_2$ and $a_1 = -a_2$.

Note that there cannot be more that two indices ℓ with same $|\beta_\ell|$ and that 1. 2. 3. and 4. above are disjoint and cover all the possible cases.

Suppose that 1. holds. Note that

$$g^{(n)}(0) = \sum_{\ell=1}^k a_\ell \beta_\ell^n \sigma^{(n)}(0) = 0,$$

for all $n \geq 1$, by assumption. On the other hand, the triangle inequality yields that

$$|g^{(n)}(0)| \geq \left(|a_1\beta_1^n| - \left|\sum_{\ell\neq 1} a_\ell\beta_\ell^n\right|\right)|\sigma^{(n)}(0)| \geq \left(|a_1\beta_1^n| - |\beta_2^n|\sum_{\ell\neq 1}|a_\ell|\right)|\sigma^{(n)}(0)|.$$

One can always choose $n_0 \geq 1$ large enough such that $\sigma^{(n_0)}(0) \neq 0$ and

$$|\beta_1| > |a_1|^{-1/n_0}|\beta_2| \left(\sum_{\ell \neq \ell_1} |a_\ell| \right)^{1/n_0},$$

so that $|g^{(n)}(0)| > 0$, which is a contradiction with the fact that $g \equiv c$. Hence $a_1 = 0$. This shows the claim in the particular situation where all $|\beta_\ell|$'s are distinct.

One can deal with case 2. using that $|a_1| > |a_2|$, writing

$$|g^{(n)}(0)| \geq \left((|a_1| - |a_2|)|\beta_1^n| - |\beta_3| \sum_{\ell \neq 1,2} |a_\ell| \right) |\sigma^{(n)}(0)|.$$

The reasoning is then identical to 1.

In the case 3., since σ has infinitely many non-zero even derivatives at 0, we use that $a_1\beta_1^{2n} + a_2\beta_2^{2n} = 2a_1\beta_1^{2n}$ to write

$$|g^{(2n)}(0)| \geq \left((2|a_1|)|\beta_1^{2n}| - \sum_{\ell \neq 1,2} |a_\ell \beta_\ell^{2n}| \right) |\sigma^{(2n)}(0)|,$$

then choose n large enough to argue as above that $a_1 = a_2 = 0$. We can thus eliminates these terms from the definition of g and go on with the argument.

In the case 4., if σ has infinitely many non-zero odd derivatives at 0, we apply the same reasoning as in 3. to show that $a_1 = a_2 = 0$.

Since σ has infinitely many even and infinitely many odd non-zero derivatives at 0, we can iterate the argument and the proof is over since the four cases above cover all possible cases. □

When σ does not satisfy Assumption A, the proof above allows us to derive the following results:

Lemma B.6. *If σ is analytic such that $\sigma^{(n)}(0) \neq 0$ for infinitely many even $n \geq 0$ but only finitely many odd $n \geq 1$, then the function g in Lemma B.5 can be written as*

$$g(\alpha) = \sum_{\ell=1}^{\widetilde{k}} \widetilde{a}_\ell \widetilde{\sigma}(\alpha \widetilde{\beta}_\ell),$$

where $\widetilde{\sigma}$ is an odd polynomial, the $\widetilde{a}_\ell$'s are nonzero and the $|\beta_\ell|$'s are pairwise distinct.

Similarly, if $\sigma^{(n)}(0) \neq 0$ for infinitely many odd $n \geq 1$ but only finitely many even $n \geq 0$, then the function g in Lemma B.5 can be written as

$$g(\alpha) = \sum_{\ell=1}^{\widetilde{k}} \widetilde{a}_\ell \widetilde{\sigma}(\alpha \widetilde{\beta}_\ell),$$

where $\widetilde{\sigma}$ is an even polynomial, the $\widetilde{a}_\ell$'s are nonzero and the $|\beta_\ell|$'s are pairwise distinct.

Proof. Suppose that $\sigma^{(2n+1)}(0) \neq 0$ for only finitely many $n \geq 0$. In the proof of Lemma B.5, the only problematic situation is 4., that is $\beta_1 = -\beta_2$ and $a_1 = -a_2$. In particular, they cancel out in the even derivatives of g, that is

$$g^{(2n)}(0) = \sigma^{(2n)}(0) \sum_{\ell \neq 1,2} a_\ell \beta_\ell^{2n}.$$

If $\beta_3, a_3, \beta_4, a_4$ do not fall into case 4. from the proof of Lemma B.5, then one can show with the same argument therein that $a_3 = a_4 = 0$. Therefore, the problem reduces to the situation where k is even, $\beta_{2\ell-1} = -\beta_{2\ell}$ and $a_{2\ell+1} = -a_{2\ell+2}$ for all $\ell \in [k/2]$. We can then rewrite g as

$$g(\alpha) = \sum_{\ell=1}^{\widetilde{k}} \widetilde{a}_\ell \widetilde{\sigma}(\alpha\widetilde{\beta}_\ell),$$

where $\widetilde{k} \leq k/2$, $\widetilde{a}_\ell := a_{2\ell-1}$, $\widetilde{\beta}_\ell := \beta_{2\ell-1}$ and $\widetilde{\sigma}(x) := \sigma(x) - \sigma(-x)$. The function $\widetilde{\sigma}$ is analytic and locally polynomial around 0, therefore is a polynomial on $\mathbb{R}$ and the $|\widetilde{\beta}_\ell|$'s are pairwise distinct.

When the even derivatives eventually vanish at 0 instead, then the problematic situation is the 3. from Lemma B.5 and the function becomes

$$g(\alpha) = \sum_{\ell=1}^{k/2} \widetilde{a}_\ell \widetilde{\sigma}(\alpha\widetilde{\beta}_\ell),$$

where $\widetilde{a}_\ell := a_{2\ell-1}$, $\widetilde{\beta}_\ell := \beta_{2\ell-1}$ and $\widetilde{\sigma}(x) := \sigma(x) + \sigma(-x)$ with $\widetilde{\sigma}$ polynomial as above. □

The case of the sigmoid activation $\sigma(x) = 1/(1+e^{-x})$. In this case, $\sigma(x) = 1/2 + \tanh(x)$ and $\tanh$ is an odd function, i.e. $\sigma^{(2n)}(0) = 0$ for all $n \geq 1$. Hence, $\widetilde{\sigma}(x) = \sigma(x) + \sigma(-x) = 1$ for all $x \in \mathbb{R}$ and one can construct the null function with already four β's satisfying the constraints: $a_1\sigma(\beta_1 x) + a_1\sigma(-\beta_1 x) + a_3\sigma(\beta_3 x) + a_3\sigma(-\beta_3 x) = 0$ as soon as $a_1 = -a_3$, such that $|\beta_1| \neq |\beta_3|$. (One could then also achieve this for any even $p \geq 4$ such functions by tuning the a_ℓ's.)

The case of the softplus activation $\sigma(x) = \ln(1+e^x)$. The Softplus function is the primitive of the sigmoid such that $\sigma(x) = \int_{-\infty}^{x} \frac{1}{1+e^{-u}} \mathrm{d}u$. Therefore, $\sigma^{(2n+1)}(0) = 0$ when $n \geq 1$. In particular, $\widetilde{\sigma}(x) = \sigma(x) - \sigma(-x) = x$ for all $x \in \mathbb{R}$. One can thus obtain the null function with four (or a strictly greater even number) β's satisfying the constraints: $a_1\sigma(\beta_1 x) - a_1\sigma(-\beta_1 x) + a_3\sigma(\beta_3 x) - a_3\sigma(-\beta_3 x) = 0$, as soon as $a_1\beta_1 + a_3\beta_3 = 0$, where $|\beta_1| \neq |\beta_3|$ are pairwise distinct.

The case of the tanh activation function $\sigma(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Since σ is an odd function, $\widetilde{\sigma}(x) = \sigma(x) + \sigma(-x) = 0$ for all $x \in \mathbb{R}$ and therefore one can achieve the null function with two (or a strictly greater even number) β's satisfying the constraints: $a_1\sigma(\beta_1 x) - a_1\sigma(-\beta_1 x)$.

We stress that for the three functions above, there is no other way to obtain the null function (i.e. the coefficients β_ℓ's and a_ℓ's have to be all in case 3. or case 4. depicted in the proof of Lemma B.5, according to the derivatives of σ).

Recall that we consider the loss L_μ^m where μ is an input data distribution with support $\mathbb{R}^{d_{\text{in}}}$.

Theorem B.7 (Theorem 4.2 in the main)**.** *Suppose that the activation function σ satisfies the Assumption A. For $m > r^*$, let $\boldsymbol{\theta}$ be an m-neuron point, and $\boldsymbol{\theta}_*$ be a unique r^*-neuron global minimum up to permutation, i.e. $L_\mu^{r^*}(\boldsymbol{\theta}_*) = 0$. If $L_\mu^m(\boldsymbol{\theta}) = 0$, then $\boldsymbol{\theta} \in \Theta_{r^* \to m}(\boldsymbol{\theta}_*)$.*

Proof. For $x \in \mathbb{R}^{d_{\text{in}}}$, let $h(x) := \sum_{j=1}^{m} a_j\sigma(w_j \cdot x) - \sum_{j=1}^{m^*} a_j^*\sigma(w_j^* \cdot x)$ and note that this function is zero on $\mathbb{R}^{d_{\text{in}}}$. Since $\boldsymbol{\theta}_*$ is irreducible, we know that the w_j^*'s are pairwise distinct, and the a_j^*'s are nonzero. We can always group terms such that, wlog, the w_j's are nonzero, pairwise distinct and the a_j's are nonzero, and we remain in the expansion manifold, as we now argue: we have that

$$h(x) = \sum_{j=1}^{m+m^*} a_j\sigma(w_j \cdot x),$$

where we set $a_j = -a_{j-m}^*$ and $w_j = w_{j-m}^*$ for $j \in \{m+1, \ldots, m+m^*\}$. If some of the w_j's appear several times, we group them together and if some are zero vectors, we summarize them in a constant $c \in \mathbb{R}$ and arrive at

$$h(x) = \sum_{j=1}^{M} A_j\sigma(W_j \cdot x) = c,$$

with $M \leq m + m^*$, such that $W_i \neq W_j$ for all $i \neq j \in [M]$ with $W_j \neq (0, \dots, 0)^{\mathrm{T}}$. Proving the claim, i.e. that $\boldsymbol{\theta} \in \Theta_{r^* \to m}(\boldsymbol{\theta}_*)$, is now equivalent to showing that $A_j = 0$ for all $j \in M$.

If $d_{\text{in}} = 1$, we simply apply Lemma B.5 which shows that $A_j = 0$ for all $j \in [M]$.

Suppose now that $d_{\text{in}} > 1$. Let $\epsilon > 0$ and let $t_\epsilon = (1, \epsilon, \epsilon^2, \dots, \epsilon^M)^{\mathrm{T}}$. We define

$$h_\epsilon(\alpha) := \sum_{j=1}^{M} A_j \sigma(\alpha W_j \cdot t_\epsilon), \qquad \alpha \in \mathbb{R}.$$

We claim that Lemma B.5 applies to h_ϵ, that is, the elements in $\{W_j \cdot t_\epsilon; j \in [M]\}$ are pairwise distinct for all $\epsilon > 0$ small enough. Indeed, by contradiction, suppose that there exists a positive decreasing sequence $(\epsilon_n)_{n \geq 1}$ such that $\lim_{n \to \infty} \epsilon_n = 0$ and $W_1 \cdot t_{\epsilon_n} = W_2 \cdot t_{\epsilon_n}$. Then $(W_1)_1 + \mathcal{O}(\epsilon_n) = (W_2)_1 + \mathcal{O}(\epsilon_n)$ where $(W_j)_k$ denotes the k-th component of W_j. Choosing n large enough enforces $(W_1)_1 = (W_2)_1$. It suffices then to explicit the terms of order ϵ_n in the identity and to reason identically since the rest is $\mathcal{O}(\epsilon_n^2)$. This implies that $W_1 = W_2$, which is a contradiction with the assumption that the vectors W_j are pairwise distinct.

Hence, by Lemma B.5 applied on h_ϵ, we have that $A_j = 0$ for all $j \in [M]$, which concludes the proof. □

Remark B.3. *The theorem above does not apply to the sigmoid, the softplus and the* $\tanh$ *activation functions, since none of these satisfy Assumption A. Nonetheless, we discussed above the theorem how to reconstruct a neural network function with these activations, with parameters that have to satisfy some explicit constraints depending on the activation (in particular, every w' in the bigger network has to be either equal to w or $-w$ of the smaller network). By considering the extended expansion manifolds of these activation functions, comprised of the classical expansion manifold and these new points, Theorem B.7 holds true, that is, the extended expansion manifold is exactly the set of global minima.*

B.4. Symmetry-Induced Critical Points

We will prove the Propositions 4.3 and 4.4 in the main. Recall that $\boldsymbol{\theta}_*^r = (w_1^*, \dots, w_r^*, a_1^*, \dots, a_r^*)$ denotes an irreducible critical point of L^r.

Proposition B.8 (Proposition 4.3 in the main)**.** *The expansion manifold $\bar{\Theta}_{r \to m}(\boldsymbol{\theta}_*^r)$ is a union of*

$$G(r, m) := \sum_{\substack{k_1 + \dots + k_r = m \\ k_i \geq 1}} \binom{m}{k_1, \dots, k_r}$$

distinct non-intersecting affine subspaces of dimension $(m - r)$ and all points therein are critical points of L^m.

Proof. First, we show that $\bar{\Theta}_{r \to m}(\boldsymbol{\theta}_*^r)$ contains $G(r, m)$ non-intersecting affine subspaces of dimension $(m - r)$. Recall that by definition, we have

$$\bar{\Theta}_{r \to m}(\boldsymbol{\theta}_*^r) = \bigcup_{\substack{s = (k_1, \dots, k_r) \\ \pi \in S_m}} \mathcal{P}_\pi \bar{\Gamma}_s(\boldsymbol{\theta}_*^r)$$

where $\bar{\Gamma}_s(\boldsymbol{\theta}_*^r)$ contains the points in the set

$$\{(\underbrace{w_1^*, \dots, w_1^*}_{k_1}, \dots, \underbrace{w_r^*, \dots, w_r^*}_{k_r}, \underbrace{\beta_1^1 a_1^*, \dots, \beta_1^{k_1} a_1^*}_{k_1}, \dots, \underbrace{\beta_r^1 a_r^*, \dots, \beta_r^{k_r} a_r^*}_{k_r}) : \sum_{i=1}^{k_t} \beta_t^i = 1 \text{ for } t \in [r]\}.$$

Observe that this is an affine subspace. Its dimension is given by the number of free parameters, that is $(k_1 - 1) + \dots + (k_r - 1) = m - r$. All its permutations $\mathcal{P}_\pi \bar{\Gamma}_s(\boldsymbol{\theta}_*^r)$ are also affine subspaces with the same dimension. For two of these subspaces to intersect, there should be a point contained in both subspaces. However, observe that the incoming weight vectors for two distinct subspaces are never the same thus an intersection point is not possible.

For the number of these subspaces, it is enough to count the distinct configurations of the incoming weight vectors, since the outgoing weight vectors follow the incoming ones. The formula for the number of distinct permutations of

$$(\underbrace{w_1^*,\dots,w_1^*}_{k_1},\underbrace{w_2^*,\dots,w_2^*}_{k_2},\dots,\underbrace{w_r^*,\dots,w_r^*}_{k_r}).$$

is $\frac{m!}{k_1!\cdots k_r!}$ for a given tuple $(k_1,\dots,k_r)$ with $k_i \geq 1$ and $k_1+\cdots+k_r=m$. Summing over all such tuples, we find the formula for $G(r,m)$.

Second, we will show that all points in $\bar{\Theta}_{r\to m}(\boldsymbol{\theta}_*^r)$ are critical. To do so we show that all points $\boldsymbol{\theta}_*^m \in \bar{\Gamma}_s(\boldsymbol{\theta}_*^r)$ are critical, then since $\nabla L^m(\mathcal{P}_\pi \boldsymbol{\theta}_*^m) = \mathcal{P}_\pi \nabla L^m(\boldsymbol{\theta}_*^m) = 0$, we obtain the result for all points in $\bar{\Theta}_{r\to m}(\boldsymbol{\theta}_*^r)$. For $i \in [r]$, we denote the gradient components with respect to the i-th incoming weight vector and the i-th outgoing weight vector as follows

$$\nabla_i^w L^r(\boldsymbol{\theta}_*^r) = \frac{(a_i^*)^T}{N} \sum_{(x,y)\in \text{Trn}} c'(f^{(2)}(x|\boldsymbol{\theta}_*^r),y)\sigma'(w_i^*\cdot x)x,$$
$$\nabla_i^a L^r(\boldsymbol{\theta}_*^r) = \frac{1}{N} \sum_{(x,y)\in \text{Trn}} c'(f^{(2)}(x|\boldsymbol{\theta}_*^r),y)\sigma(w_i^*\cdot x).$$

By introducing the $d_{\text{out}} \times d_{\text{in}}$ matrix U and the d_{out}-dimensional vector V as

$$U(w) := \frac{1}{N} \sum_{(x,y)\in \text{Trn}} c'(f^{(2)}(x|\boldsymbol{\theta}_*^r),y)\sigma'(w\cdot x)x,$$
$$V(w) := \frac{1}{N} \sum_{(x,y)\in \text{Trn}} c'(f^{(2)}(x|\boldsymbol{\theta}_*^r),y)\sigma(w\cdot x).$$

we have $\nabla_i^w L^r(\boldsymbol{\theta}_*^r) = (a_i^*)^T U(w_i^*)$ and $\nabla_i^a L^r(\boldsymbol{\theta}_*^r) = V(w_i^*)$. Since $\boldsymbol{\theta}_*^r$ is a critical point, we have $(a_i^*)^T U(w_i^*) = 0$ and $V(w_i^*) = 0$ for all $i \in [r]$. For $\boldsymbol{\theta}_*^m \in \bar{\Gamma}_s(\boldsymbol{\theta}_*^r)$, we have $f^{(2)}(x|\boldsymbol{\theta}_*^m) = f^{(2)}(x|\boldsymbol{\theta}_*^r)$ and we write down the gradient components for L^m at $\boldsymbol{\theta}_*^m$

$$\nabla_{K_i+j}^w L^m(\boldsymbol{\theta}_*^m) = \frac{\beta_i^j (a_i^*)^T}{n} \sum_{(x,y)\in \text{Trn}} c'(f^{(2)}(x|\boldsymbol{\theta}_*^m),y)\sigma'(w_i^*\cdot x)x = \beta_i^j (a_i^*)^T U(w_i^*)$$
$$\nabla_{K_i+j}^a L^m(\boldsymbol{\theta}_*^m) = \frac{1}{n} \sum_{(x,y)\in \text{Trn}} c'(f^{(2)}(x|\boldsymbol{\theta}_*^m),y)\sigma(w_i^*\cdot x) = V(w_i^*)$$

where $K_i = k_1+\cdots+k_{i-1}$ and $j \in [k_i]$ for all $i \in [r]$. Since all gradient components are zero, thus $\boldsymbol{\theta}_*^m$ is a critical point of L^m.

□

Proposition B.9 (Proposition 4.4 in the main)**.** *For twice-differentiable c and σ, for all $\boldsymbol{\theta}_*^m \in \bar{\Theta}_{r\to m}(\boldsymbol{\theta}_*^r)$, the spectrum of the Hessian $\nabla^2 L^m(\boldsymbol{\theta}_*^m)$ has $(m-r)$ zero eigenvalues. Moreover, if $\boldsymbol{\theta}_*^r$ is a strict saddle, then all points in $\bar{\Theta}_{r\to m}(\boldsymbol{\theta}_*^r)$ are also strict saddles.*

Proof. Because any $\boldsymbol{\theta}_*^m \in \bar{\Theta}_{r\to m}(\boldsymbol{\theta}_*^r)$ lies in an equal-loss affine subspace of dimension $(m-r)$, it has at least $(m-r)$ zero eigenvalues in the Hessian.

For $\boldsymbol{\theta}_*^r$ that is a strict saddle of L^r, we have an eigenvector β such that $\beta^T \nabla^2 L^r(\boldsymbol{\theta}_*^r)\beta < 0$. Since $\bar{\Theta}_{r\to m}(\boldsymbol{\theta}_*^r)$ is an equal-loss manifold where all the points have the same loss as $\boldsymbol{\theta}_*^r$, we have $L^r(\boldsymbol{\theta}_*^r) = L^m(\boldsymbol{\theta}_*^m) = L^m(U\boldsymbol{\theta}_*^r)$ where U is a linear map. Finally, we have $(U\beta)^T \nabla^2 L^m(\boldsymbol{\theta}_*^m) U\beta = \beta^T \nabla^2 L^r(\boldsymbol{\theta}_*^r)\beta < 0$ by the chain rule, and therefore $\nabla^2 L^m(\boldsymbol{\theta}_*^m)$ cannot be a positive semidefinite matrix, i.e. it has a negative eigenvalue, which completes the proof. □

B.5. Combinatorial Analysis

For proving the exact combinatorial results presented in the main (Proposition 4.5 and Lemma 4.6) it will be convenient to use Newton's series for finite differences (Milne-Thomson, 2000):

Definition B.2. *Let p be a polynomial of degree d, we define the k-th forward difference of the polynomial $p(x)$ at 0 as*

$$\Delta^k[p](0) = \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} p(i).$$

Hence, we can write $p(x)$ as

$$p(x) = \sum_{k=0}^{d} \binom{x}{k} \Delta^k[p](0). \tag{2}$$

Rearranging the summands in Equation 2, one observes that Newton's series for finite differences is a discrete analog of Taylor's series

$$p(x) = \sum_{k=0}^{d} \frac{\Delta^k[p](0)}{k!} [x]_k$$

where $(x)_k = x(x-1)\dots(x-k+1)$ is the falling factorial.

We now proceed with proving Proposition 4.5 in the main.

Proposition B.10 (Proposition 4.5 in the main)**.** *For $r \leq m$, we have*

$$G(r,m) = \sum_{i=1}^{r} \binom{r}{i} (-1)^{r-i} i^m, \tag{3}$$

$$T(r,m) = G(r,m) + \sum_{u=1}^{m-r} \binom{m}{u} G(r, m-u) g(u). \tag{4}$$

where $g(u) = \sum_{j=1}^{u} \frac{1}{j!} G(j,u)$.

Proof. The proof of the theorem is divided in the two next Propositon. In Proposition B.11 we prove Equation (3), while in Proposition B.13, using a counting argument (Lemma B.12), we prove Equation (4). □

Proposition B.11. *For $r \leq m$, we have*

$$G(r,m) = \sum_{i=1}^{r} \binom{r}{i} (-1)^{r-i} i^m.$$

Proof. First, recall that, by Proposition B.8, we have that

$$G(r,m) := \sum_{\substack{k_1+\dots+k_r=m \\ k_i \geq 1}} \binom{m}{k_1, \dots, k_r}.$$

The above can be restated by using the identity

$$\sum_{\substack{k_1+\dots+k_r=m \\ k_i \geq 0}} \binom{m}{k_1, \dots, k_r} = \sum_{\ell=0}^{r} \binom{r}{\ell} \sum_{\substack{k_1+\dots+k_r=m \\ k_i \geq 0}} \binom{m}{k_1, \dots, k_r} \mathbb{1}_{I_\ell}(k_1, \dots, k_r) \tag{5}$$

where $I_\ell := \{(0, \dots, 0, k_{\ell+1}, \dots, k_r) : k_i \geq 1 \text{ for } \ell+1 \leq i \leq r\}$. Equation (5) is equivalent to

$$r^m = \sum_{\ell=0}^{r} \binom{r}{\ell} G(r-\ell, m) \tag{6}$$

$$= \sum_{\ell=0}^{r} \binom{r}{\ell} G(\ell, m), \tag{7}$$

with the convention that $G(0,m) = 0$. Newton's series for finite differences (Equation (2)), applied to the polynomial $p(x) = x^m$ at $x = r$, yields

$$r^m = \sum_{\ell=0}^{r} \binom{r}{\ell} \sum_{i=0}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} i^m. \tag{8}$$

Note that the outer summation goes up to r instead of m since the terms with a factor $\binom{r}{k}$ for $k \geq r+1$ are zero. Hence we have

$$\sum_{\ell=0}^{r} \binom{r}{\ell} \left[\sum_{i=0}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} i^m - G(\ell, m) \right] = 0. \tag{9}$$

Indeed, with m fixed, the solution

$$G(\ell, m) = \sum_{i=0}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} i^m \tag{10}$$

is the unique solution for the Equation (9) with initial value given by the condition $1^m = 1$. The uniqueness follows from an immediate induction argument: since

$$G(1,m) = \sum_{k_1=m} \binom{m}{k_1} = 1 = \sum_{i=0}^{1} \binom{1}{i} (-1)^{1-i} i^m,$$

the initial step of induction is verified. Then, for the induction hypothesis, for $k = 1, \ldots, r-1$, the first $r-1$ term in the summation in Equation (9) are null, leaving us with the condition

$$G(r,m) = \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} i^m.$$

□

The Proposition above, which holds for $r < m$ shows that $G(r,m)$ are the forward finite difference at 0 for $p(x) = x^m$, i.e. $G(r,m) = \Delta^r[p](0)$. We now comment on the meaning of the formula for $r \geq m$. For a given polynomial $p(x)$ define the *rescaled* Newton's finite differences $\Delta_h^r[p](0)$ as Newton's finite differences (at 0) for the polynomial $p(hx)$; hence, we can write the r-th derivative of the polynomial p as the $h \to 0$ limit of the h-the r-th Newton's finite difference:

$$p^{(r)}(0) = \lim_{h \to 0^+} \frac{\Delta_h^r[p](0)}{h^r} = \lim_{h \to 0^+} \frac{1}{h^r} \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} (hi)^m = \lim_{h \to 0^+} \frac{1}{h^{r-m}} G(r,m).$$

Hence for $r = m$ we obtain $G(m,m) = m!$, whereas for $r > m$ we find $G(r,m) = 0$.

In order to prove Equation (4), we introduce the following Lemma B.12, which is in fact a counting of the same number in two ways.

Lemma B.12. *For $j \leq n$, we have*

$$\frac{1}{j!} G(j,n) = \sum_{\substack{c_1 + 2c_2 + \cdots + nc_n = n \\ c_1 + c_2 + \cdots + c_n = j \\ c_i \geq 0}} \frac{n!}{1!^{c_1} 2!^{c_2} \cdots n!^{c_n}} \frac{1}{c_1! \cdots c_n!}.$$

Proof. By definition, we have

$$G(j,n) = \sum_{\substack{b_1 + \ldots + b_j = n \\ b_i \geq 1}} \binom{n}{b_1, \ldots, b_j}.$$

Starting from a tuple $(b_1, \dots, b_j)$, consider the tuple $(c_1, \dots, c_n)$ where c_i is the number of occurence of i in $(b_1, \dots, b_j)$. Therefore we have

$$\binom{n}{b_1, \dots, b_j} = \binom{n}{\underbrace{1, \dots, 1}_{c_1}, \underbrace{2, \dots, 2}_{c_2}, \dots, \underbrace{n}_{c_n}} = \frac{n!}{1!^{c_1} \cdots n!^{c_n}}. \tag{11}$$

Moreover, any c-tuple $(c_1, \dots, c_n)$ appears in

$$\binom{j}{c_1, \dots, c_n} = \frac{j!}{c_1! \cdots c_n!} \tag{12}$$

b-tuples that are exactly $(b_1, \dots, b_j)$. From Equation (11) and Equation (12) and summing over all tuples $(c_1, \dots, c_n)$ we conclude. □

We are now in position to prove the closed-form formula for T, Equation (4).

Proposition B.13. *For $r \leq m$, we have*

$$T(r, m) = G(r, m) + \sum_{u=1}^{m-r} \binom{m}{u} G(r, m-u) g(u) \tag{13}$$

where $g(u) = \sum_{j=1}^{u} \frac{1}{j!} G(j, u)$.

Proof. Let $u = b_1 + \cdots + b_j$ and let c_i be, as in Lemma B.12, the number of occurrences of i among $(b_1, \dots, b_j)$. Recall that for T we have the identity

$$T(r, m) := \sum_{j=0}^{m-r} \sum_{\substack{sum(s)=m \\ k_i \geq 1, b_i \geq 1}} \binom{m}{k_1, \dots, k_r, b_1, \dots, b_j} \frac{1}{c_b}.$$

We rewrite the outer summation in T from the number of b_i's to the summation of b_i's and we obtain

$$T(r, m) = \sum_{u=0}^{m-r} \sum_{j=0}^{u} \binom{m}{u} \sum_{\substack{k_1 + \cdots + k_r = m-u \\ b_1 + \cdots + b_j = u \\ k_i \geq 1, b_i \geq 1}} \binom{m-u}{k_1, \dots, k_r} \binom{u}{b_1, \dots, b_j} \frac{1}{c_1! c_2! \cdots c_{m-r}!}$$

where we split the inner summation and the multinomial coefficient into two parts: one that comes from the incoming weight vectors and the others come from the zero-type neurons $(w'_1, \dots, w'_j)$. Using the formula for G on $(k_1, \dots, k_r)$, we simplify as follows

$$T(r, m) = \sum_{u=0}^{m-r} \binom{m}{u} G(r, m-u) \sum_{j=0}^{u} \sum_{\substack{b_1 + \cdots + b_j = u \\ b_i \geq 1}} \binom{u}{b_1, \dots, b_j} \frac{1}{c_1! c_2! \cdots c_{m-r}!}.$$

Finally using Lemma B.12, we find

$$T(r, m) = \sum_{u=0}^{m-r} \binom{m}{u} G(r, m-u) \sum_{j=0}^{u} \frac{1}{j!} G(j, u)$$

where $G(0, 0) = 1$. Splitting the case $u = 0$, we derive the closed form formula

$$T(r, m) = G(r, m) + \sum_{u=1}^{m-r} \binom{m}{u} G(r, m-u) \sum_{j=1}^{u} \frac{1}{j!} G(j, u).$$

□

Lemma B.14 (Lemma 4.6 in the main)**.** *For any $k \geq 0$ fixed, we have,*

$$G(m-k,m) \sim T(m-k,m) \sim \frac{m^k}{2^k k!} m!, \text{ as } m \to \infty.$$

For any fixed $r \geq 0$, we have $G(r,m) \sim r^m$ as $m \to \infty$.

Proof. We begin to show that

$$\lim_{r\to\infty} \frac{1}{(r+k)!r^k} G(r, r+k) = \frac{1}{2^k k!}. \tag{14}$$

In particular, we observe that for $k=1$ we have that

$$G(r,r+1) = \sum_{\substack{k_1+\ldots+k_r=r+1 \\ k_i\geq 1}} \binom{r+1}{k_1,\ldots,k_r} = \binom{r}{1}\binom{r+1}{2,1,\ldots,1} = r\frac{(r+1)!}{2!}.$$

We find that the asymptotic in Equation (14) is in fact an exact equality for any $r > 0$. For a generic $k \geq 0$, we divide the summation in G according to the number of 1's in $(k_1,\ldots,k_r)$

$$\begin{aligned} G(r,r+k) &= \sum_{\substack{k_1+\cdots+k_r=r+k \\ k_i\geq 1}} \binom{r+k}{k_1,\ldots,k_r} \\ &= \binom{r}{k}\binom{r+k}{\underbrace{2,\ldots,2}_{k},\underbrace{1,\ldots,1}_{r-k}} + \sum_{n=1}^{k-1}\binom{r}{n} \sum_{\substack{k_1+\cdots+k_n=n+k \\ k_i\geq 2}} \binom{r+k}{k_1,\ldots,k_n,\underbrace{1,\ldots,1}_{r-n}}. \end{aligned} \tag{15}$$

For a given tuple $(k_1,\ldots,k_n)$, let $c=(c_2,\ldots,c_n)$, with $\sum_{i=2}^n c_i = n$ and c_i is the number of occurrences of i among $(k_1,\ldots,k_n)$, hence we have

$$\binom{r+k}{k_1,\ldots,k_n,1,\ldots,1} = \frac{(r+k)!}{2!^{c_2}\cdots n!^{c_n}}.$$

Since for a given $c=(c_2,\ldots,c_n)$ there are $\binom{n}{c_2,\ldots,c_n}$ n-tuples $(k_1,\ldots,k_n)$ with such occurrences, we rewrite Equation (15) as

$$G(r,r+k) = \binom{r}{k}\frac{(r+k)!}{2^k} + \sum_{n=1}^{k-1}\binom{r}{n} \sum_{\substack{2c_2+\cdots+nc_n=n+k \\ c_2+\cdots+c_n=n}} \binom{n}{c_2,\ldots,c_n}\frac{(r+k)!}{2!^{c_2}\cdots n!^{c_n}}.$$

Dividing both sides by $(r+k)!r^k$, we find

$$\frac{G(r,r+k)}{(r+k)!r^k} = \frac{1}{2^k k!}\frac{r(r-1)\ldots(r-k+1)}{r^k} + \sum_{n=1}^{k-1}\sum_{\substack{2c_2+\cdots+nc_n=n+k \\ c_2+\cdots+c_n=n}} \frac{r(r-1)\ldots(r-n+1)}{r^k} C_c, \tag{16}$$

where $C_c := 1/(c_2!\cdots c_n!\cdot 2!^{c_2}\cdots n!^{c_n})$. For $n \leq k$, we have the following immediate double inequality:

$$r^{n-k}\left(\frac{r-n+1}{r}\right)^n \leq \frac{r(r-1)\ldots(r-n+1)}{r^k} \leq r^{n-k}.$$

Together with Equation (16), the above double inequality leads to

$$\begin{aligned} &\frac{1}{2^k k!}\left(\frac{r-k+1}{r}\right)^k + \sum_{n=1}^{k-1}\sum_{\substack{2c_2+\cdots+nc_n=n+k \\ c_2+\cdots+c_n=n}} r^{n-k}\left(\frac{r-n+1}{r}\right)^n C_c \\ &\qquad \leq \frac{1}{(r+k)!r^k}G(r,r+k) \leq \frac{1}{2^k k!} + \sum_{n=1}^{k-1}\sum_{\substack{2c_2+\cdots+nc_n=n+k \\ c_2+\ldots+c_n=n}} r^{n-k}C_c. \end{aligned}$$

In the limit $r \to \infty$, both the lower and the upper bound converge to $\frac{1}{2^k k!}$, hence giving

$$G(r, r+k) \sim \frac{r^k (r+k)!}{2^k k!} \sim \frac{(r+k)^k (r+k)!}{2^k k!};$$

finally, by choosing $r = m - k$, we recover the first asymptotic of the Lemma. In order to prove the asymptotic for $T(m-k, m)$ we divide both sides in Equation (13) (with $r = m - k$) by $G(m-k, m)$:

$$\frac{T(m-k,m)}{G(m-k,m)} = 1 + \sum_{u=1}^{k} \binom{m}{u} \frac{G(m-k, m-u)}{G(m-k,m)} g(u).$$

The limit of $T(m-k, m)$ as $m \to \infty$, is then obtained from the asymptotic of $G(m-k, m)$ above:

$$1 + \sum_{u=1}^{k} \binom{m}{u} \frac{G(m-k,m-u)}{G(m-k,m)} g(u) \sim 1 + \sum_{u=1}^{k} \frac{m^u}{u!} c_u \frac{m^{k-u}(m-u)!}{m^k m!} g(u) \sim 1 + \sum_{u=1}^{k} \frac{g(u)}{u!} \frac{c_u}{m^u} \sim 1$$

hence, for large m, $T(m-k, m)$ and $G(m-k, m)$ grows at the same rate.

Finally, with an induction argument, we show that $G(r, m) \sim r^m$ for fixed r and $m \gg r$. For $r = 1$, we have $G(1, m) = 1$. For $r = 2$, we have $G(2, m) = 2^m - 2 \sim 2^m$. We assume that for all $\ell = 1, \dots, r-1$, we have $G(\ell, m) \sim \ell^m$. Normalizing Equation (6) by $1/r^m$, as $m \to \infty$ we have

$$1 = \frac{1}{r^m} G(r,m) + \frac{1}{r^m} \sum_{\ell=1}^{r-1} \binom{r}{\ell} G(\ell, m) \sim \frac{1}{r^m} G(r,m) + \sum_{\ell=1}^{r-1} \frac{r^\ell}{\ell!} \left(\frac{\ell}{r}\right)^m \sim \frac{1}{r^m} G(r,m).$$

which completes the induction step, thus the Lemma. □

Thanks to the Propositions and Lemmas demonstrated in this section, we are now in position of proving the asymptotic behaviours presented in Equations (6) and (7) of the main, for *mildly* and *vastly parameterized* regimes, respectively. We assume an overparameterized network of width $m = r^* + n$ where the minimal width r^* is large.

Mildly Overparameterized (small h). For fixed k and h, in the limit $r^* \to \infty$, Lemma B.14 (Lemma 4.6 in the main) gives the following asymptotic for $G(r^* - k, m)$ and for $T(r^*, m)$:

$$G(r^*-k, m) \sim \frac{(m)^{k+h}}{2^{k+h}(k+h)!} m! \sim \frac{(r^*)^{k+h}}{2^{k+h}(k+h)!} m!\,,$$
$$T(r^*, m) \sim \frac{m^h}{2^h h!} m! \sim \frac{(r^*)^h}{2^h h!} m!\,.$$

Taking the ratio of the two quantities above, we find

$$\frac{G(r^*-k,m)}{T(r^*,m)} \sim \frac{(r^*)^{k+h}}{2^{k+h}(k+h)!} \frac{2^h h!}{(r^*)^h} = \frac{(r^*)^k}{2^k (k+h)\cdots(h+1)}.$$

Vastly Overparameterized ($h \gg r^*$). We consider the case where h is much bigger than r^*.

Using Equation (6) at $r = r^* - 1$, we find

$$\sum_{\ell=1}^{r^*-1} \binom{r^*-1}{\ell} G(\ell, m) = (r^*-1)^m.$$

We also have that $T(r^*, m) \geq G(r^*, m)$. Thus if the numbers a_k of critical points in a network of width $k \in [r^* - 1]$ are bounded by $\binom{r^*-1}{r^*-k}$, we have

$$\frac{\sum_{k=1}^{r^*-1} a_k G(r^*-k, m)}{T(r^*,m)} \leq \frac{\sum_{r=1}^{r^*-1} \binom{r^*-1}{r} G(r,m)}{G(r^*,m)} = \frac{(r^*-1)^m}{G(r^*,m)}.$$

On the other hand, since $r^* \ll m$, we have that (Lemma B.14 , i.e. Lemma 4.6 in the main)

$$\frac{(r^*-1)^m}{G(r^*,m)} \sim \left(\frac{r^*-1}{r^*}\right)^m$$

as the limit $m \to \infty$. Thus the inequality (7) in the main holds for large m. Although beyond the scope of the paper, it is worth to point out that a more refined asymptotic analysis for G can be carried on by means of the Nørlund-Rice integral and saddle point techniques.

B.6. Multi-Layer ANNs

In the case of multi-layers, the equivalence of two incoming weight vectors in the intermediate layers should be understood in the general sense, i.e. all incoming weight vectors of layer ℓ are the outgoing weight vectors of layer $\ell-1$ that can be written as

$$\{(\underbrace{(a_1^1)_d,\dots,(a_1^{k_1})_d)}_{k_1},\dots,\underbrace{(a_r^1)_d,\dots,(a_r^{k_r})_d}_{k_r},\underbrace{(\alpha_1^1)_d,\dots,(\alpha_1^{b_1})_d}_{b_1},\dots,\underbrace{(\alpha_1^1)_d,\dots,(\alpha_r^{b_j})_d)}_{b_j} : \sum_{i=1}^{k_t}(a_t^i)_d = (a_t)_d \text{ and } \sum_{i=1}^{b_t}(\alpha_t^i)_d = 0\}$$

where $d \in [r_\ell]$. All weight vectors in this set are equivalent in the sense that they produce the same neuron in layer ℓ.

For the general shape of the multi-layer expansion manifold, let us consider first a three-layer network. If we add one neuron to the first hidden layer, we have that $\Theta^{(1)}_{\boldsymbol{r}\to\boldsymbol{m}}(\boldsymbol{\theta^r})$ is connected. If we do not add a new neuron in the second hidden layer, the permutations of the neurons in the second hidden layer would bring $r_2!$ disconnected components where each one of the disconnected components have $T(r_1, r_1+1)$ affine subspaces that are connected to each other. Note that in this case the overall manifold $\Theta_{\boldsymbol{r}\to\boldsymbol{m}}(\boldsymbol{\theta^r})$ is disconnected. However, adding one neuron to the second hidden layer, every $r_2!$ disconnected components get connected through the parameters of the neurons in the second hidden layer, which yields a connected multi-layer expansion manifold $\Theta_{\boldsymbol{r}\to\boldsymbol{m}}(\boldsymbol{\theta^r})$.

In general, adding n_1 neurons to the first hidden layer results in $T(r_1, r_1+n_1)$ connected affine subspaces instead of the usual $r_1!$ discrete (i.e. disconnected) points. Adding n_2 neurons to the second hidden layer brings $T(r_2, r_2+n_2)$ affine subspaces instead of the usual $r_2!$ points, for each one of the $T(r_1, r_1+n_1)$ affine subspaces. Note that this is multiplicative because every combination of the parameters in the first hidden layer can be paired with every combination of the parameters in the second hidden layer which results in a distinct affine subspace. Similarly, via induction, if $n_\ell \geq 1$ for all $\ell \in [L-1]$, adding $(n_1,\dots,n_{L-1})$ neurons to each one of the hidden layers make a connected manifold of $\prod_{\ell=1}^{L-1} T(r_1, r_1+n_1)$ affine subspaces.

References

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v9/glorot10a.html.

Johnson, S. G. The nlopt nonlinear-optimization package. URL http://github.com/stevengj/nlopt.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv e-prints*, December 2014. URL https://arxiv.org/abs/1412.6980.

Milne-Thomson, L. M. *The calculus of finite differences*. American Mathematical Soc., 2000.