A. Proof of Theorem 1

In (Friedland, 2006) the theorem is proven for square matrices. In fact Theorem 1 can be deduced from this case by the following argument:

For a sequence of non-square matrices $(A_k)_{k\in\mathbb{N}} \in \mathbb{R}^{m_k\times l_k}$ of finite size $m_k, l_k \leq L$ we can always find a finite set of subsequent matrices that when multiplied together are a square matrix.

$$\underbrace{A_1 \cdot ... \cdot A_{n_1}}_{=:\bar{A}_1\in\mathbb{R}^{m\times m}} \cdot \underbrace{A_{n_1+1} \cdot ... \cdot A_{n_2}}_{=:\bar{A}_2\in\mathbb{R}^{m\times m}} \cdot \underbrace{A_{n_2+1} \cdot ... \cdot A_{n_3}}_{=:\bar{A}_3\in\mathbb{R}^{m\times m}} \cdot ... \tag{17}$$

The matrices $\bar{A}_1, \bar{A}_2, \bar{A}_3, ...$ define a sequence of non-negative square matrices that fulfill the conditions in (Friedland, 2006) and therefore converge to a rank-1 matrix.

Our proof requires no knowledge of algebraic geometry and uses the cosine similarity to show convergence. First, we outline the conditions on the matrix sequence A_n. Then, we state the theorem again and sketch our proof to give the reader a better overview. Finally, we prove the theorem in 5 steps.

Conditions on A_n The first obvious condition is that the $(A_n)_{n\in\mathbb{N}}$ is a sequence of non-negative matrices such that A_i, A_{i+1} have the correct size to be multiplied together. Secondly, as we calculate angles between column vector in our proof, no column of A_n should be zero. The angle between a zero vector and any other vector is undefined. Finally, the size of A_n should not increase infinitely, i.e. an upper bound on the size of the A_i's exists such that $A_i \in \mathbb{R}^{m\times l}$ where $m, l \leq L$ for some $L \in \mathbb{N}$.

Definition 1. We say $\lim_{n\to\infty} A_n$ **exists**, if for all $m, l \in \mathbb{N}$ the subsequences $A_{n_{k(m,l)}} \in \mathbb{R}^{m\times l}$ that consist of all terms that have size $m \times l$ converge elementwise. Note that even if there only finitely many, say A_N is the last term with $A_n \in \mathbb{R}^{m\times l}$, we say $\lim_{k\to\infty} A_{n_{k(l,m)}} = A_N$.

Theorem 1. *Let $A_1, A_2, A_3 \ldots$ be a sequence of non-negative matrices as described above such that $\lim_{n\to\infty} A_n$ exists. We exclude the cases where one column of $\lim_{n\to\infty} A_n$ is the zero vector or two columns are orthogonal to each other. Then the product of all terms of the sequence converges to a rank-1 matrix $\bar{C}$:*

$$\bar{C} := \prod_{i=1}^{\infty} A_i = \bar{\boldsymbol{c}}\boldsymbol{\gamma}^T \,. \tag{18}$$

Matrices of this form are *excluded* for $\lim_{n\to\infty} A_n$:

$$\begin{array}{c|c}
\left(\underbrace{\boldsymbol{v}_1 \;\; ... \;\; \boldsymbol{v}_l}_{\text{arbitrary}} \;\; \underbrace{\boldsymbol{v}_{l+1} \;\; ... \;\; \boldsymbol{v}_m}_{\text{orthogonal}} \;\; \underbrace{\boldsymbol{v}_{m+1} \;\; ... \;\; \boldsymbol{v}_n}_{\boldsymbol{v}_i = 0}\right) & \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \\
\text{Theory} & \text{Example}
\end{array}$$

up to ordering of the columns.

Proof sketch To show that $\prod_i^\infty A_i$ converges to a rank-1 matrix, we do the following steps:

(1) We define a sequence s_n as the cosine of the maximum angle between the column vectors of $M_n := \prod_{i=1}^n A_i$.

(2) We show that the sequence s_n is monotonic and bounded and therefore converging.

(3) We assume $\lim_{n\to\infty} s_n \neq 1$ and analyze two cases where we do not get a contradiction. Each case yields an equation on $\lim_{n\to\infty} A_n$.

(4) In both cases, we find lower bounds on s_n: $\alpha_n s_{n-1}$ and $\alpha'_n s_{n-1}$ that are becoming infinitely large, unless we have $\lim_{n\to\infty} \alpha_n = 1$ (case 1) or $\lim_{n\to\infty} \alpha'_n = 1$ (case 2).

(5) The lower bounds lead to equations on $\lim_{n\to\infty} A_n$ for non-convergence. The only solutions, we obtain for $\lim_{n\to\infty} A_n$, are those explicitly excluded in the theorem. We still get a contradiction and $\lim_{n\to\infty} s_n = 1 \Rightarrow \prod_i^\infty A_i = \bar{c}\gamma^T$.

Proof **(1)** Let $M_n := \prod_{i=1}^n A_i$ be the product of the matrices $A_1 \cdot \ldots \cdot A_n$. We define a sequence on the angles of column vectors of M_n using the cosine similarity. Let $\boldsymbol{v}_1(n), ..., \boldsymbol{v}_{k(n)}(n)$ be the column vectors of M_n. Note, the angles are well defined between the columns of M_n. The columns of M_n cannot be a zero vector as we required A_n to have no zero columns. Let s_n be the cosine of the maximal angle between the columns of M_n:

$$s_n = \min_{i\neq j} \mathrm{s}_{\cos}(\boldsymbol{v}_i(n), \boldsymbol{v}_j(n)) := \min_{i,j} \frac{\langle \boldsymbol{v}_i(n), \boldsymbol{v}_j(n)\rangle}{\|\boldsymbol{v}_i(n)\|\|\boldsymbol{v}_j(n)\|}, \tag{19}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. We show that the maximal angle converges to 0 as $\lim_{n\to\infty} s_n = 1$, which is equivalent to M_n converging to a rank-1 matrix. In the following, we take a look at two consecutive elements of the sequence s_n and check by how much the sequence increases.

(2) We show that the sequence s_n is monotonic and bounded and therefore converging. Assume $\boldsymbol{a}_{n+1}$ and $\boldsymbol{b}_{n+1}$ are the two columns of A_{n+1} which produce the columns $\boldsymbol{v}_m(n+1)$ and $\boldsymbol{v}_{m'}(n+1)$ of M_{n+1} with the maximum angle:

$$s_{n+1} = \mathrm{s}_{\cos}(\boldsymbol{v}_m(n+1), \boldsymbol{v}_{m'}(n+1)) = \mathrm{s}_{\cos}(M_n \boldsymbol{a}_{n+1}, M_n \boldsymbol{b}_{n+1}). \tag{20}$$

We also assume that $\|\boldsymbol{v}_i(n)\| = 1$ for all i, since the angle is independent of length. To declutter notation, we write $\boldsymbol{v}_i(n) =: \boldsymbol{v}_i$, $\boldsymbol{a}_n = \boldsymbol{a} = (a_1, ..., a_k)^T$ and $\boldsymbol{b}_n = \boldsymbol{b} = (b_1, ..., b_k)^T$. We now show that s_n is monotonic and use the definition of the cosine similarity:

$$s_{n+1} = \frac{\sum_{ij} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{\|\sum_i a_i \boldsymbol{v}_i\| \|\sum_i b_i \boldsymbol{v}_i\|} \tag{21}$$

Using the triangle inequality $\|\sum_i a_i \boldsymbol{v}_i\| \leq \sum_i a_i \|\boldsymbol{v}_i\|$ we get:

$$s_{n+1} \geq \frac{\sum_{ij} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{(\sum_i a_i \|\boldsymbol{v}_i\|)(\sum_i b_i \|\boldsymbol{v}_i\|)} \tag{22}$$

As we assumed that the $\|\boldsymbol{v}_i\| = 1$, we know that $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = s_{\cos}(\boldsymbol{v}_i, \boldsymbol{v}_j)$ which must be greater than the smallest cosine similarity s_n:

$$s_{n+1} \geq \frac{\sum_{ij} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{(\sum_i a_i)(\sum_i b_i)} \geq \frac{\sum_{ij} a_i b_j}{(\sum_i a_i)(\sum_i b_i)} s_n = s_n \tag{23}$$

Therefore s_n is monotonically increasing and upper-bounded by 1 as the cosine. Due to the monotone convergence theorem, it will converge. The rest of the proof investigates if the sequence s_n converges to 1 and if so, under which conditions.

(3) We look at two consecutive sequence elements and measure the factor α by which they increase: $s_{n+1} \geq \alpha s_n$. We are using proof by contradiction and assume that s_n does not converge to 1. We get two cases, each with a lower bound on the factor of increase. For both cases, we find a lower bound for $\alpha > 1$ for all n which would mean that s_n is diverging to ∞ – a contradiction. Under certain conditions on $\lim_{n\to\infty} A_n$, we do not find a lower bound $\alpha > 1$. We find that these conditions correspond to the conditions explicitly excluded in the theorem and therefore M_n converges to a rank-1 matrix.

Case 1: Let $t_n := \langle \boldsymbol{v}_l(n), \boldsymbol{v}_m(n) \rangle$ $(l \neq m)$ and assume that there exists a subsequence t_{n_k} of t_n that does not converge to $\lim_{n\to\infty} s_n$. So there is an $\varepsilon > 0$ such that $\langle \boldsymbol{v}_l(n_k), \boldsymbol{v}_m(n_k) \rangle \geq (1+\varepsilon) s_{n_k}$ for all $k \geq K$ for some $K \in \mathbb{N}$.

We multiply the first lower bound of equation 23 by $1 = \frac{s_n}{s_n}$ and get:

$$s_{n+1} \geq \frac{\sum_{ij} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{(\sum_i a_i)(\sum_i b_i)} = \frac{\sum_{ij} a_i b_j \frac{\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{s_n}}{(\sum_i a_i)(\sum_i b_i)} s_n, \tag{24}$$

We will now pull terms corresponding to the pair (l, m) out of the sum and for all terms in the sum, we lower bound $\frac{\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{s_n} \geq 1$ by one. Let the set $I := \{(i,j) \mid (i,j) \neq (l,m), (m,l)\}$ index all other terms:

$$s_{n+1} \geq \frac{\sum_I a_i b_j + (a_l b_m + a_m b_l) \frac{\langle \boldsymbol{v}_l, \boldsymbol{v}_m \rangle}{s_n}}{\left(\sum_i a_i\right)\left(\sum_i b_i\right)} s_n \tag{25}$$

We know that $\langle \boldsymbol{v}_l(n_k), \boldsymbol{v}_m(n_k) \rangle \geq (1+\varepsilon) s_{n_k}$:

$$s_{n_{k+1}} \geq s_{n_k+1} \geq \frac{\sum_I a_i b_j + (a_l b_m + a_m b_l)(1+\varepsilon)}{\left(\sum_i a_i\right)\left(\sum_i b_i\right)} s_{n_k} \tag{26}$$

We absorb the m, l factors back into the sum:

$$s_{n_{k+1}} \geq \frac{\sum_{ij} a_i b_j + \overbrace{(a_l b_m + a_m b_l)}^{=: r_{n_k}} \varepsilon}{\left(\sum_i a_i\right)\left(\sum_i b_i\right)} s_{n_k} = \left(1 + \frac{r_{n_k} \varepsilon}{\left(\sum_i a_i\right)\left(\sum_i b_i\right)}\right) s_{n_k} \geq \underbrace{\left(1 + \frac{r_{n_k}\varepsilon}{\bar{q}}\right)}_{=: \alpha_{n_k}} s_{n_k} \tag{27}$$

where $\bar{q}$ is an upper bound on $\sum_{ij} a_i b_j$ which exists since $\lim_{n\to\infty} A_n$ exists, which is also why $\lim_{k\to\infty} r_{n_k}$ exists.

(4) Case 1: Define $r_n = a_l(n) b_m(n) + a_m(n) b_l(n)$ analogous to r_{n_k}. So if $\lim_{n\to\infty} r_n = \lim_{k\to\infty} r_{n_k} \neq 0$, the factor by which s_n increases would be greater that one by a constant – a contradiction:

$$s_{n_k} \geq (1+c)^{n'} s_{n_1} > 0 \tag{28}$$

where $n_k - n'$ is the number of cases where $r_{n_k} = 0$ and $c > 0$ is a lower bound on the set $\{\frac{\varepsilon r_{n_k}}{\bar{q}} \neq 0\}$. As we assumed $\lim_{n\to\infty} r_n \neq 0$, $c > 0$ for an infinite number of cases and therefore $n' \to \infty$ when $k \to \infty$.

To end case 1, we have to ensure that the first sequence element is greater than zero: $s_{n_1} \geq s_1 > 0$. This is not the case if the first N matrices have two orthogonal columns, $s_1 = \ldots = s_N = 0$. We can then skip the first N matrices and define s_1 on A_{N+1} (set $A_i = A_{N+i}$). We know N has to be finite, as $\lim_{n\to\infty} A_n$ has no two columns that are orthogonal.

(3) Case 2: No subsequence of t_n, as defined in case 1, exists , i.e. for all $\varepsilon > 0$ no subsequence $t_{n_k} = \langle \boldsymbol{v}_l(n), \boldsymbol{v}_m(n) \rangle \geq (1+\varepsilon) s_{n_k}$ exists. Then t_n and s_n converge to the same value: $\lim_{n\to\infty} t_n - s_n = 0$. Since we assumed that s_n does not converge to one, it must converge to a value smaller than 1 by a constant ε'. An $N \in \mathbb{N}$ exists such that for all $n \geq N$ there is an $\varepsilon' > 0$ with $\langle \boldsymbol{v}_l(n), \boldsymbol{v}_m(n) \rangle \leq 1 - \varepsilon'$. We derive a second lower bound:

$$s_{n+1} = \frac{\sum_{ij} a_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle}{\| \sum_i a_i \boldsymbol{v}_i \| \| \sum_i b_i \boldsymbol{v}_i \|} \geq \frac{\sum_{ij} a_i b_j}{\| \sum_i a_i \boldsymbol{v}_i \| \| \sum_i b_i \boldsymbol{v}_i \|} s_n, \tag{29}$$

where we used $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle \geq s_n$. We now find a lower bound for the square of this factor. The steps are similar to case 1:

$$\frac{(\sum_{ij} a_i b_j)^2}{\| \sum_i a_i \boldsymbol{v}_i \|^2 \| \sum_i b_i \boldsymbol{v}_i \|^2} = \frac{(\sum_{ij} a_i b_j)^2}{(\sum_{ij} a_i a_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle)(\sum_{ij} b_i b_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle)} \tag{30}$$

$$= \frac{(\sum_{ij} a_i b_j)^2}{(\sum_I a_i a_j + 2 a_l a_m \langle \boldsymbol{v}_l, \boldsymbol{v}_m \rangle)(\sum_I b_i b_j + 2 b_l b_m \langle \boldsymbol{v}_l, \boldsymbol{v}_m \rangle)} \tag{31}$$

$$\geq \frac{(\sum_{ij} a_i b_j)^2}{(\sum_I a_i a_j + 2 a_l a_m (1-\varepsilon'))(\sum_I b_i b_j + 2 b_l b_m (1-\varepsilon'))} \tag{32}$$

$$= \frac{q^2}{q^2 - \varepsilon' \left(\sum_{ij} 2 a_l a_m b_i b_j + \sum_{ij} 2 b_l b_m a_i a_j - 4\varepsilon' a_l a_m b_l b_m \right)} \tag{33}$$

$$\geq \frac{q^2}{q^2 - \varepsilon'^2 \underbrace{\left(\sum_{ij} 2 a_l a_m b_i b_j + \sum_{ij} 2 b_l b_m a_i a_j - 4 a_l a_m b_l b_m \right)}_{=: r'_n}} \tag{34}$$

$$= \frac{q^2}{q^2 - \varepsilon'^2 r'_n} = 1 + \frac{\varepsilon'^2 r'_n}{q^2 - \varepsilon'^2 r'_n} \geq 1 + \frac{\varepsilon'^2 r'_n}{\bar{q}^2 - \varepsilon'^2 r'_n} =: \alpha'^2_n \tag{35}$$

where $q^2 = (\sum_{ij} a_i b_j)^2$ and $\bar{q}^2$ is an upper bound on q^2 for all n. Note that $q^2 - \varepsilon'^2 r'_n > 0$, since q^2 has all terms that r'_n has but more.

(4) Case 2: So r'_n is a sequence that converges to zero. Otherwise, the factor by which s_n increases would be greater than one by at least a constant for infinitely many n. As in the previous case 1, this would lead to a contradiction.

(5) Case 1 and case 2 are complements from which we obtain two equations for $\lim_{n\to\infty} A_n$. Let $\boldsymbol{a}_k = (a_1, ..., a_k)^T$ and $\boldsymbol{b}_k = (b_1, ..., b_k)^T$ be columns of $\lim_{n\to\infty} A_n$. We get one equation per case. For all (i, j) with $i < j$ we have:

$$\textbf{Case 1: } \lim_{n\to\infty} r_n = 0 \Rightarrow a_i b_j = a_j b_i = 0 \quad \text{or} \quad \textbf{Case 2: } \lim_{n\to\infty} r'_n = 0 \Rightarrow a_i a_j = b_i b_j = 0, \tag{36}$$

where the first equation comes from $\lim_{n\to\infty} r_n = 0$ and the second from $\lim_{n\to\infty} r'_n = 0$.

For equation 36 to be true, the following set of equations have to hold:

$$S(k) := \{\forall i = 1, ..., k : a_i = 0 \neq b_i, a_i \neq 0 = b_i \text{ or } a_i = b_i = 0\} \tag{37}$$

$$\cup \quad \{\exists l : a_l \neq 0 \neq b_l \text{ and } \forall i \neq l : a_i = b_i = 0\}. \tag{38}$$

This is equivalent to the matrix being rank one already or one of the columns is the zero vector or two are orthogonal to each other. To show why this statement holds, we are using induction on k. For $k = 2$, we have the following set of solutions:

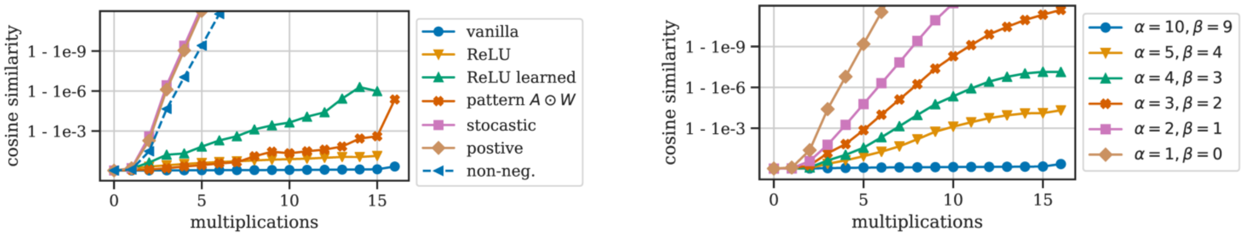
$$\begin{array}{c|c} \begin{pmatrix}0 & b_1\\ 0 & b_2\end{pmatrix} \quad \begin{pmatrix}a_1 & 0\\ a_2 & 0\end{pmatrix} \quad \begin{pmatrix}0 & 0\\ a_2 & b_2\end{pmatrix} \quad \begin{pmatrix}a_1 & b_1\\ 0 & 0\end{pmatrix} & \begin{pmatrix}a_1 & 0\\ 0 & b_2\end{pmatrix} \quad \begin{pmatrix}0 & b_1\\ a_2 & 0\end{pmatrix} \\ \\ \text{Case 1} & \text{Case 2} \end{array}$$

Case one provides rank 1 matrices only and case two gives orthogonal columns. So, the statement holds for $k = 2$.

Next assume we solved the problem for columns with k entries and want to deduce the case where we have $k + 1$ entries (i.e. they satisfy the equations in $S(k + 1)$). The pair a_{k+1}, b_{k+1} satisfies either one of the three equations in line equation 37: $a_{k+1} = b_{k+1} = 0$, $a_{k+1} = 0 \neq b_{k+1}$ or $a_{k+1} \neq 0 = b_{k+1}$. If $a_{k+1} = b_{k+1} = 0$, the rest of the non-trivial equations will be the same set of equations that one will get in the case of k entries. If $a_{k+1} = 0 \neq b_{k+1}$, we would be left with equation $a_i b_{k+1} = 0$ (Case 1) or $b_i b_{k+1} = 0$ (Case 2) which would mean that for all $i \leq k$ either $a_i = 0$ or $b_i = 0$, which will satisfy the equations in $S(k + 1)$. We get an analogous argument in the case of $a_{k+1} \neq 0 = b_{k+1}$.

The other possibility is that $a_{k+1} \neq 0 \neq b_{k+1}$. But in this case both equations from case one and two $a_{k+1} b_i = a_i b_{k+1} = 0$ and $a_{k+1} a_i = b_{k+1} b_i = 0$ lead to $a_i = b_i = 0$ for all $i \leq k$ and this satisfies $S(k + 1)$ in line equation 38, concluding the induction.

This completes the proof. Since $\lim_{n\to\infty} s_n \neq 1$ only if $\lim_{n\to\infty} A_n$ has a column that is the zero vector, a multiple of a standard basis vector or it has two columns that are orthogonal to each. Exactly, the conditions excluded in the theorem. For all other cases, we get a contradiction: therefore $\lim_{n\to\infty} s_n = 1$ and M_n converges to a rank-1 matrix. $\square$



(a) different matrix properties (see text)

(b) $\alpha W^+ + \beta W^-$ where $W_{[ij]} \sim \mathcal{N}(0, 1)$

Figure 7: Simulated convergence for a matrix chain.

B. Convergence Speed & Simulation of Matrix Convergences

We proved that $M_n = \prod_i^n A_i$ converges to a rank-1 matrix for $n \to \infty$, but which practical implications has this for a 16 weight-layered network? How quickly is the convergence for matrices considered in neural networks?

We know that s_n increases by a factor $(1 + c)$ greater than 1 ($c > 0$):

$$s_n \geq (1 + c) s_{n-1} \tag{39}$$

Each iteration yields such a factor and we get a chain of factors:

$$s_n \geq (1 + c_n)(1 + c_{n-1})...(1 + c_2) s_1 \tag{40}$$

Although the multiplication chain of c_n has some similarities to an exponential form γ^n, s_n does not have to converge exponentially as the individual c_n have to decrease (s_n bounded by 1). We investigated the convergence speed using a simulation of random matrices and find that non-negative matrices decay exponentially fast towards 1.

We report the converging behavior for matrix chains which resembles a VGG-16. As in the backward pass, we start from the last layer. The convolutional kernels are considered to be 1x1, e.g. for a kernel of size (3, 3, 256, 128), we use a matrix of size (256, 128).

We test out the effect of different matrix properties. For *vanilla*, we sample the matrix entries from a normal distribution. Next, we apply a *ReLU* operation after each multiplication. For *ReLU learned*, we used the corresponding learned VGG parameters. We generate *non-negative* matrices containing 50% zeros by clipping random matrices to $[0, \infty]$. And *positive* matrices by taking the absolute value. We report the median cosine similarity between the column vectors of the matrix.

The y-axis of Figure 7a has a logarithmic scale. We observe that the positive, stochastic, and non-negative matrices yield a linear path, indicating an exponential decay of the form: $1 - \exp(-\lambda n)$. The 50% zeros in the non-negative matrices only result in a bit lower convergence slope. After 7 iterations, they converged to a single vector up to floating point imprecision.

We also investigated how a slightly negative matrix influences the convergence. In Figure 7b, we show the converges of matrices: $\alpha W^+ + \beta W^-$ where $W^+ = \max(0, W), W^- = \min(0, W)$ and $W \sim \mathcal{N}(0, I)$. We find that for small enough $\beta < 4$ values the matrix chains still converge. This simulation motivated us to include LRP$_{\alpha 5 \beta 4}$ in our evaluation which show less convergence on VGG-16, but its saliency maps also contain more noise.

C. Pattern Attribution

We derive equation 9 from the original equation given in (Kindermans et al., 2018). We will use the notation from the original paper and denote a weight vector with $\boldsymbol{w} = W_{l_{[i,:]}}$ and the corresponding pattern with $\boldsymbol{a} = A_{l_{[i,:]}}$. The output is $y = \boldsymbol{w}^T \boldsymbol{x}$.

Derivation of Pattern Computation For the positive patterns of the two-component estimator $S_{\boldsymbol{a}+-}$, the expectation is taken only over $\{\boldsymbol{x} | \boldsymbol{w}^T \boldsymbol{x} > 0\}$. We only show it for the positive patterns $\boldsymbol{a}_+$. As our derivation is independent of the subset of $\boldsymbol{x}$ considered, it would work analogously for negative patterns or the linear estimator $S_{\boldsymbol{a}}$.

The formula to compute the pattern $\boldsymbol{a}_+$ is given by:

$$\begin{aligned}\boldsymbol{a}_+ &= \frac{\mathbb{E}_+\left[\boldsymbol{x}y\right] - \mathbb{E}_+\left[\boldsymbol{x}\right]\mathbb{E}_+\left[y\right]}{\boldsymbol{w}^T\mathbb{E}_+\left[\boldsymbol{x}y\right] - \boldsymbol{w}^T\mathbb{E}_+\left[\boldsymbol{x}\right]\mathbb{E}_+\left[y\right]} \\ &= \frac{\operatorname{cov}[\boldsymbol{x}, \boldsymbol{w}^T\boldsymbol{x}]}{\boldsymbol{w}^T\operatorname{cov}[\boldsymbol{x}, \boldsymbol{w}^T\boldsymbol{x}]},\end{aligned} \tag{41}$$

where $\operatorname{cov}[\boldsymbol{x}, \boldsymbol{w}^T\boldsymbol{x}] = \mathbb{E}_+[\boldsymbol{x}y] - \mathbb{E}_+[\boldsymbol{x}]\mathbb{E}_+[y]$. Using the bilinearity of the covariance matrix ($\operatorname{cov}[\boldsymbol{b}, \boldsymbol{c}^T\boldsymbol{d}] = \operatorname{cov}[\boldsymbol{b}, \boldsymbol{d}]\boldsymbol{c}$), gives:

$$\boldsymbol{a}_+ = \frac{\operatorname{cov}[\boldsymbol{x}, \boldsymbol{x}]\boldsymbol{w}}{\boldsymbol{w}^T\operatorname{cov}[\boldsymbol{x}, \boldsymbol{x}]\boldsymbol{w}}. \tag{42}$$

Using the notation $\operatorname{cov}[\boldsymbol{h}] = \operatorname{cov}[\boldsymbol{x}, \boldsymbol{x}]$ gives equation 9.

Connection to power iteration A step of the power iteration is given by:

$$\boldsymbol{v}_{k+1} = \frac{M\boldsymbol{v}_k}{\|M\boldsymbol{v}_k\|} \tag{43}$$

The denominator in equation 9 is $\boldsymbol{w}^T\operatorname{cov}[\boldsymbol{h}]\boldsymbol{w}$. Using the symmetry of $\operatorname{cov}[\boldsymbol{h}]$, we have:

$$\left\|\operatorname{cov}[\boldsymbol{h}]^{1/2}\boldsymbol{w}\right\| = (\boldsymbol{w}^T\operatorname{cov}[\boldsymbol{h}]^{1/2}\operatorname{cov}[\boldsymbol{h}]^{1/2}\boldsymbol{w})^{1/2} = (\boldsymbol{w}^T cov[\boldsymbol{h}]\boldsymbol{w})^{1/2} \tag{44}$$

This should be similar to the norm $\|\operatorname{cov}[h]\boldsymbol{w}\|$. As only a single step of the power iteration is performed, the scaling should not matter that much. The purpose of the scaling in the power-iteration algorithm is to keep the vector v_k from exploding or converging to zero.

D. CIFAR-10 Network Architecture

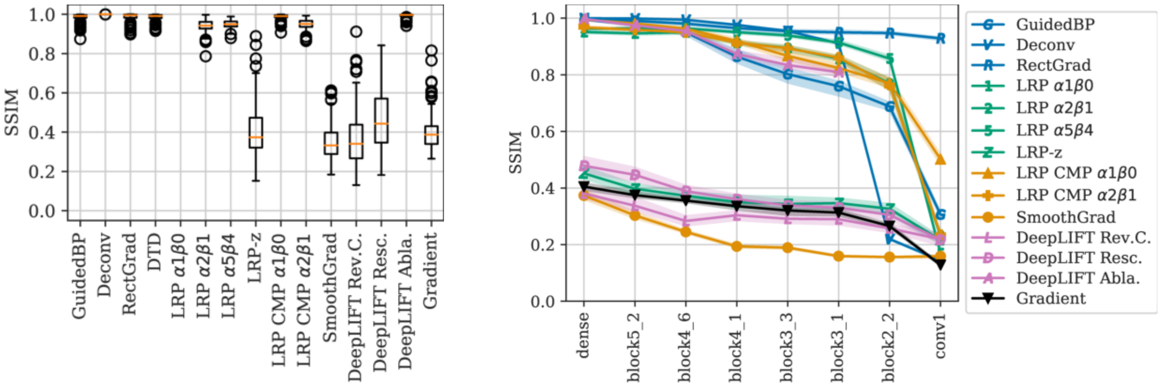
```
# network architecture as a keras model
model = Sequential()

model.add(InputLayer(input_shape=(32, 32, 3), name='input'))
model.add(Conv2D(32, (3, 3), padding='same', name='conv1'))
model.add(Activation('relu', name='relu1'))
model.add(Conv2D(64, (3, 3), padding='same', name='conv2'))
model.add(Activation('relu', name='relu2'))
model.add(MaxPooling2D(pool_size=(2, 2), name='pool2'))

model.add(Conv2D(128, (3, 3), padding='same', name='conv3'))
model.add(Activation('relu', name='relu3'))
model.add(Conv2D(128, (3, 3), padding='same', name='conv4'))
model.add(Activation('relu', name='relu4'))
model.add(MaxPooling2D(pool_size=(2, 2), name='pool4'))

model.add(Flatten(name='flatten'))
model.add(Dropout(0.5, name='dropout5'))
model.add(Dense(1024, name='fc5'))
model.add(Activation('relu', name='relu5'))
model.add(Dropout(0.5, name='dropout6'))
model.add(Dense(10, name='fc6'))
model.add(Activation('softmax', name='softmax'))
```

E. Results on ResNet-50

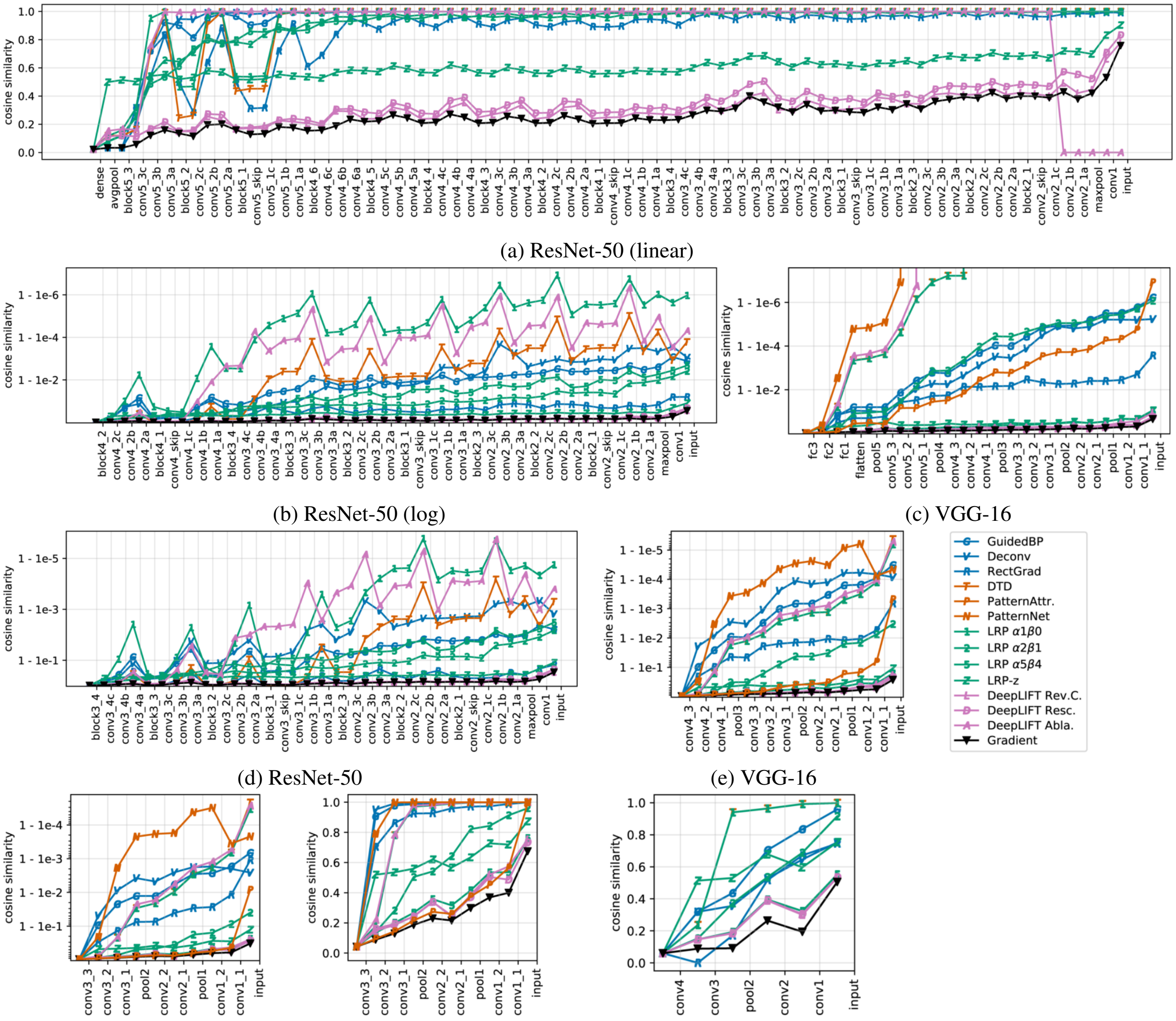


(a) Random Logits

(b) Cascading Parameter Randomization

Figure 8: Effect of (a) randomizing the logits or (b) the parameters on a ResNet-50.

F. Additional Cosine Similarity Figures



(a) ResNet-50 (linear)

(b) ResNet-50 (log)

(c) VGG-16

(d) ResNet-50

(e) VGG-16

(f) VGG-16

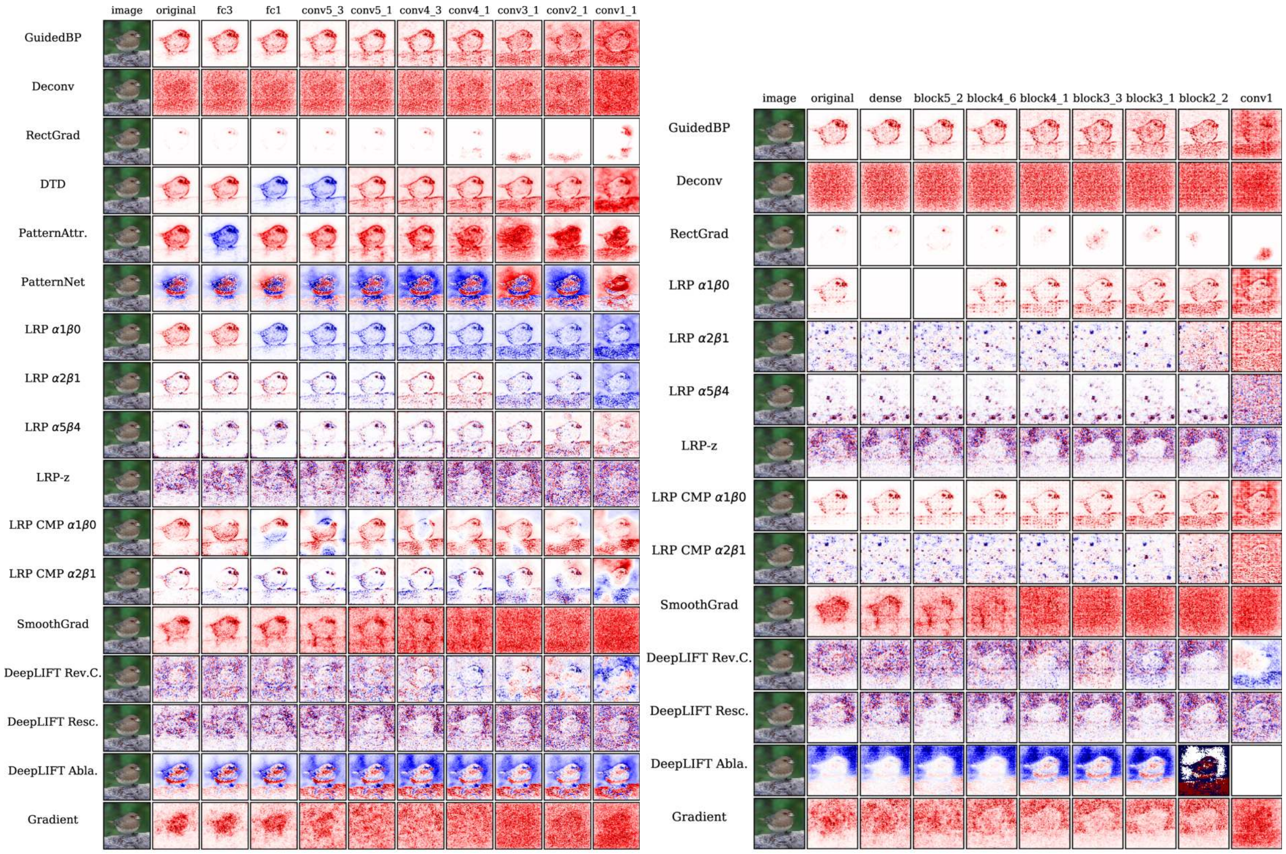
(g) VGG-16 (linear)

(h) CIFAR-10 (linear)

Figure 9: Convergence measured using the CSC for different starting layers.

G. Saliency maps for Sanity Checks

For visualization, we normalized the saliency maps to be in $[0, 1]$ if the method produce only positive relevance. If the method also estimates negative relevance, than it is normalized to $[-1, 1]$. The negative and positive values are scaled equally by the absolute maximum. For the sanity checks, we scale all saliency maps to be in $[0, 1]$.



(a) VGG-16.

(b) ResNet-50.

Figure 10: Saliency maps for sanity checks. Parameters are randomized starting from last to first layer.