Appendix: Feature Learning in L_2-regularized DNNs

The appendix is structured as follows:

1. In Section A, we describe the Experimental setup.
2. In Section B, we prove Proposition 1 of the main underlying the first reformulation.
3. In Section C, we prove Proposition 3 for the second reformulation. We also give an example of a local minimum of the original loss which is not a local minimum in the second reformulation.
4. In Section D, we prove Proposition 8, 13, 14 and 15 of the main.

A Experimental Setup

The experiments were done on fully-connected DNNs of depth $L = 3$ with varying widths.

We used the MNIST dataset [2] under the 'Creative Commons Attribution-Share Alike 3.0' license. For the MNIST examples we trained the networks on the multiclass cross-entropy loss with L_2-regularization.

We also used synthetic data sampled from a teacher network. The network has depth $L = 3$, widths $\mathbf{n} = (50, 10, 10, 10)$ with random Gaussian weights. The cost used was the Mean Squared Error (MSE).

For the experiments of Figure 1 of the main, the DNN was trained with full batch GD. For the experiments of Figure 2 we first trained with Adam [1] and finished with full batch GD (GD seems to be better suited to consistently reach the bottom of the local minima, though Adam trains faster overall). For the right plot of Figure 2, three independent networks were trained for every width and the one with the smallest loss at the end of training was chosen (the plotted test error is that of the chosen network).

The goal of Figure 2 is to identify the start of the plateau, note however that we cannot guarantee that our training procedures actually approaches a global minimum. Interestingly it was easier to observe a plateau on MNIST rather than on the teacher network data, which is why we had to take the minimum over 3 trials in the teacher setting. This could be due to the change of loss (from cross entropy to the MSE) or due to the change of the data. Note that in Figure 2 (right), it is unclear whether the 'failed' trials , i.e. the small blue dots with a loss above the plateau even for large widths, are stuck at local minima of the loss or if they could have reached the plateau if we had trained them longer.

The experiments each took between 1 and 4 hour on a single NVIDIA GeForce GTX 1080.

B Equivalence for the first reformulation

Proposition 1 (Proposition 1 of the main)**.** *The infimum of* $\mathcal{L}_\lambda(\mathbf{W}) = C(Z_L(X; \mathbf{W})) + \lambda \|\mathbf{W}\|^2$, *over the parameters* $\mathrm{W} \in \mathbb{R}^P$ *is equal to the infimum of*

$$\mathcal{L}^r_\lambda(Z_1, \ldots, Z_L) = C(Z_L) + \lambda \sum_{\ell=1}^{L} \left\| Z_\ell \left(Z^\sigma_{\ell-1}\right)^+ \right\|_F^2$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

over the set $\mathcal{Z}$ of hidden representations $\mathbf{Z} = (Z_\ell)_{\ell=1,\dots,L}$ such that $Z_\ell \in \mathbb{R}^{n_\ell \times N}$, $\mathrm{Im} Z_{\ell+1}^T \subset \mathrm{Im}\,(Z_\ell^\sigma)^T$, with the notations $Z_0^\sigma = \begin{pmatrix} X \\ \beta \mathbf{1}_N^T \end{pmatrix}$ and $Z_\ell^\sigma = \begin{pmatrix} \sigma(Z_\ell) \\ \beta \mathbf{1}_N^T \end{pmatrix}$.

Furthermore, if $\mathbf{W}$ is a local minimizer of $\mathcal{L}_\lambda$ then $(Z_1(X;\mathbf{W}), \dots, Z_L(X;\mathbf{W}))$ is a local minimizer of $\mathcal{L}_\lambda^r$. Conversely, keeping the same notations, if $(Z_\ell)_{\ell=1,\dots,L}$ is a local minimizer of $\mathcal{L}_\lambda^r$, then $\mathrm{W} = (Z_\ell (Z_{\ell-1}^\sigma)^+)_{\ell=1,\dots,L}$ is a local minimizer of $\mathcal{L}_\lambda$.

Proof. We write Φ for the map which sends some weights $\mathbf{W}$ to the hidden representations $(Z_1(X;\mathbf{W}), \dots, Z_L(X;\mathbf{W}))$ and Ψ for the map which sends some hidden representations $\mathbf{Z} \in \mathcal{Z}$ to $\mathbf{W}$ with weight matrices $W_\ell = Z_\ell \left(Z_{\ell-1}^\sigma\right)^+$.

We clearly have $\Phi(\Psi(\mathbf{Z})) = \mathbf{Z}$ for any $\mathbf{Z} \in \mathcal{Z}$, however it is not true in general that $\Psi(\Phi(\mathbf{W}))$ for all $\mathbf{W}$ (actually this is true iff $\mathbf{W}$ lies in the image of Ψ).

Let $\mathcal{L}_\lambda(\mathbf{W}) = C(Y_\mathbf{W}) + \lambda \|\mathbf{W}\|^2$ and $\mathcal{L}_\lambda^r(\mathbf{Z}) = C(Z_L) + \lambda \sum_{\ell=1}^L \left\| Z_\ell \left(Z_{\ell-1}^\sigma\right)^+ \right\|_F^2$. One can show that $\mathcal{L}_\lambda(\Psi(\mathbf{Z})) = \mathcal{L}_\lambda^r(\mathbf{Z})$ for all $\mathbf{Z} \in \mathcal{Z}$ and $\mathcal{L}_\lambda^r(\Phi(\mathbf{W})) \leq \mathcal{L}_\lambda(\mathbf{W})$ for all $\mathbf{W}$ (actually $\mathcal{L}_\lambda^r(\Phi(\mathbf{W})) = \mathcal{L}_\lambda(\mathbf{W})$ if $\mathbf{W} \in \mathrm{Im}\Psi$ and $\mathcal{L}_\lambda^r(\Phi(\mathbf{W})) < \mathcal{L}_\lambda(\mathbf{W})$ otherwise). The first fact implies that $\inf_\mathbf{W} \mathcal{L}_\lambda(\mathbf{W}) \leq \inf_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}_\lambda^r(\mathbf{Z})$ while the second implies $\inf_\mathbf{W} \mathcal{L}_\lambda(\mathbf{W}) \geq \inf_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}_\lambda^r(\mathbf{Z})$, furthermore the maps Φ and Ψ must map global minimizers to global minimizers.

Local Minima: We now extend the correspondence to local minima and saddles:

We prove that if $\mathbf{Z}$ is a local minimum of $\mathbf{Z} \mapsto \mathcal{L}_\lambda^r(\mathbf{Z})$ then $\mathbf{W} = \Psi(\mathbf{Z})$ is a local minimum of $\mathbf{W} \mapsto \mathcal{L}_\lambda(\mathbf{W})$ through the contrapositive: if $\mathbf{W} = \Psi(Z)$ is not a local minimum of the loss $\mathbf{W} \mapsto \mathcal{L}_\lambda(\mathbf{W})$ (i.e. there is a sequence of weights $\mathbf{W}_1, \mathbf{W}_2, \dots$ which converges to $\mathbf{W}$ with $\mathcal{L}_\lambda(\mathbf{W}_i) < \mathcal{L}_\lambda(\mathbf{W})$ for all i) then $\mathbf{Z}$ is not a local minimum. We simply consider the sequence $\mathbf{Z}_i = \Phi(\mathbf{W}_i)$ which converges to $\mathbf{Z} = \Phi(\Psi(\mathbf{Z}))$ by the continuity of Φ. This sequence satisfies $\mathcal{L}_\lambda^r(\mathbf{Z}_i) \leq \mathcal{L}_\lambda(\mathbf{W}_i) < \mathcal{L}_\lambda(\mathbf{W}) = \mathcal{L}_\lambda^r(\mathbf{Z})$, proving that $\mathbf{Z}$ is not a local minimum.

Let us now prove if $\mathbf{W}$ is a local minimum of $\mathbf{W} \mapsto \mathcal{L}_\lambda(\mathbf{W})$ then $\mathbf{Z} = \Phi(\mathbf{W})$ is a local minimum of $\mathbf{Z} \mapsto \mathcal{L}_\lambda^r(\mathbf{Z})$, again using the contrapositive. Assume that there is a sequence $\mathbf{Z}_1, \mathbf{Z}_2, \dots$ which converges to $\mathbf{Z} = \Phi(\mathbf{W})$ with $\mathcal{L}_\lambda^r(\mathbf{Z}_i) < \mathcal{L}_\lambda^r(\mathbf{Z})$ for all i. We consider the sequence $\mathbf{W}_i = \Psi(\mathbf{Z}_i)$, however this sequence might not be convergent since Ψ is not continuous, however we know the sequence is bounded, since $\|\mathbf{W}_i\|^2 = \sum_{\ell=1}^L \left\| Z_{i,\ell} \left(Z_{i,\ell-1}^\sigma\right)^+ \right\|_F^2 \leq \frac{1}{\lambda} \mathcal{L}_\lambda^r(\mathbf{Z}_i) < \frac{1}{\lambda} \mathcal{L}_\lambda^r(\mathbf{Z})$, this implies that there exists a subsequence $\mathbf{Z}_{k_i}$ such that $\mathbf{W}_{k_i} = \Psi\left(\mathbf{Z}_{k_i}\right)$ converges to some weights $\mathbf{W}'$. Note that since $\Phi(\mathbf{W}) = \Phi(\mathbf{W}')$ the weight matrices must agree up to 'useless weights', i.e. for all ℓ

$$W_\ell = Z_\ell \left(Z_{\ell-1}^\sigma\right)^+ + \tilde{W}_\ell$$
$$W'_\ell = Z_\ell \left(Z_{\ell-1}^\sigma\right)^+ + \tilde{W}'_\ell.$$

If $\tilde{W}_\ell \neq 0$ then $\mathbf{W}$ is not a local minimum (since we could choose the weights $W_\ell^\epsilon = Z_\ell \left(Z_{\ell-1}^\sigma\right)^+ + (1-\epsilon)\tilde{W}_\ell$ for any $0 < \epsilon < 1$ to get a lower loss). We may therefore assume $\tilde{W}_\ell = 0$, but this implies that $\tilde{W}'_\ell = 0$ too since $\|\mathbf{W}\| = \|\mathbf{W}'\|$ and therefore $\mathbf{W}' = \mathbf{W}$ and therefore $\mathbf{W}$ is not a local minimum since the sequence $\mathbf{W}_{k_i}$ approaches $\mathbf{W}$ with a strictly lower loss. □

B.1 Optimization

It is possible to optimize the first reformulation directly, using projected gradient descent to guarantee that the constraints $\mathrm{Im} Z_{\ell+1}^T \subseteq \mathrm{Im}\,(Z_\ell^\sigma)^T$ remain satisfied. As we show now, this projection is unnecessary in the continuous case, which suggests that it might also be unnecessary in gradient descent with a small enough learning rate.

Assume there is a ℓ s.t. $\mathrm{Im} Z_{\ell+1}^T \nsubseteq \mathrm{Im}\,(Z_\ell^\sigma)^T$, i.e. there is a vector $v \in \mathbb{R}^N$ (with $\|v\| = 1$) such that $v \in \ker Z_\ell^\sigma$ but $\|Z_{\ell+1} v\| > 0$. Consider any $\tilde{\mathcal{Z}}$ such that $\left\| \tilde{\mathcal{Z}} - \mathcal{Z} \right\| \leq \epsilon$, then

$$\left\| \tilde{Z}_{\ell+1} \left(\tilde{Z}_\ell^\sigma\right)^+ \right\|_F^2 \geq \left\| \tilde{Z}_{\ell+1} v v^T \left(\tilde{Z}_\ell^\sigma\right)^+ \right\|_F^2 = \left\| \tilde{Z}_{\ell+1} v \right\|^2 \left\| v^T \left(\tilde{Z}_\ell^\sigma\right)^+ \right\|^2 \geq \frac{\|Z_{\ell+1} v\|^2 - \epsilon}{\epsilon}.$$

This implies that the loss explodes in the vicinity of any point where the constraints are not satisfied. As a result, gradient flow on the cost $\mathcal{L}^r_\lambda$ starting from a value with a non-zero loss will never approach a non-acceptable point (where $\mathrm{Im} Z^T_{\ell+1} \not\subseteq \mathrm{Im}\left(Z^\sigma_\ell\right)^T$) since the loss is decreasing during gradient flow.

C Equivalence for the second reformulation

Proposition 2 (Proposition 3 of the main)**.** *For positively homogeneous non-linearities σ, the infimum of $\mathcal{L}_\lambda(\mathbf{W}) = C(Z_L(X;\mathbf{W})) + \lambda \|\mathbf{W}\|^2$, over the parameters $\mathrm{W} \in \mathbb{R}^P$ is equal to the infimum over $\mathcal{K}_\mathbf{n}(X)$ of*

$$\mathcal{L}^{\mathrm{k}}_\lambda(\mathbf{K}, Z_L) = C(Z_L) + \lambda \sum_{\ell=1}^{L} \mathrm{Tr}\left[K_\ell \left(K^\sigma_{\ell-1}\right)^+\right].$$

The set $\mathcal{K}_\mathbf{n}(X)$ is the set of covariances $\mathbf{K} = ((K_1, K^\sigma_1), \dots, (K_{L-1}, K^\sigma_{L-1}))$ and outputs Z_L such that for all hidden layer $\ell = 1, \dots, L-1$:

- *the pair (K_ℓ, K^σ_ℓ) belongs to the (translated) n_ℓ-conical hull*

$$S_{n_\ell,\beta} = \mathrm{cone}_{n_\ell}\left(\left\{\left(xx^T, \sigma(x)\sigma(x)^T\right) : x \in \mathbb{R}^N\right\}\right) + (0, \beta^2 \mathbf{1}_{N\times N}),$$

- $\mathrm{Im} K_\ell \subset \mathrm{Im} K^\sigma_{\ell-1}$, *with the notation $K^\sigma_0 = X^T X + \beta^2 \mathbf{1}_{N\times N}$ and for the outputs,* $\mathrm{Im} Z_L \subset \mathrm{Im} K^\sigma_{L-1}$.

Proof. Consider the map Ψ that maps parameters $\mathbf{W}$ to the the tuple $(\mathbf{K}, Z_L)$. We simply need to show that the image of Ψ is the set $\mathcal{K}_\mathbf{n}(X)$. The fact that $\mathrm{Im}\Psi \subset \mathcal{K}_\mathbf{n}(X)$ can easily be checked.

To prove $\mathrm{Im}\Psi \supset \mathcal{K}_\mathbf{n}(X)$ we need to construct a pre-image $\mathbf{W} \in \Psi^{-1}(\mathbf{K}, Z_L)$ from any tuple $(\mathbf{K}, Z_L)$ in $\mathcal{K}_\mathbf{n}(X)$. For every hidden layer ℓ, we have $(K_\ell, K^\sigma_\ell) \in S_{n_\ell,\beta}$. There are hence representations $Z_\ell \in \mathbb{R}^{n_\ell \times N}$ such that $K_\ell = Z^T_\ell Z_\ell$ and $K^\sigma_\ell = \sigma\left(Z_\ell\right)^T \sigma\left(Z_\ell\right) + \beta^2 \mathbf{1}_{N\times N}$, furthermore for all ℓ, we have $\mathrm{Im} Z^T_\ell = \mathrm{Im} K_\ell$ and $\mathrm{Im}\begin{pmatrix} \sigma\left(Z_\ell\right) \\ \beta \mathbf{1}^T_N \end{pmatrix} = \mathrm{Im} K^\sigma_\ell$, which implies that $\mathrm{Im} Z^T_\ell \subset \mathrm{Im}\left(Z^\sigma_{\ell-1}\right)^T$ and therefore that the tuple $(Z_1, \dots Z_L)$ is in the set $\mathcal{Z}_\mathbf{n}$ and we can choose the weight matrices $W_\ell = Z_\ell \left(Z^\sigma_{\ell-1}\right)^+$ to obtain a preimage $\mathbf{W} \in \Psi^{-1}\left(\mathbf{K}, Z_L\right)$. □

C.1 Non-correspondence of the local minima

Let us consider the map $\Gamma : \mathbf{Z} \mapsto (\mathbf{K}, Z_L)$ which maps each hidden representation Z_ℓ to the kernel pair $(Z^T_\ell Z_\ell, \left(Z^\sigma_\ell\right)^T Z^\sigma_\ell)$. The continuity of Γ implies that if $\Gamma(\mathbf{Z})$ is a local minimum then so is $\mathbf{Z}$. The converse is not true, instead we have:

Proposition 3. *A kernel and outputs pair $(\mathbf{K}, Z_L)$ is a local minimum if all $\mathbf{Z} \in \Gamma^{-1}(\mathbf{K})$ are local minima.*

Proof. We will prove the contrapositive of this statement: if $(\mathbf{K}, Z_L)$ is a saddle (i.e. there is a sequence $(\mathbf{K}_1, Z_{L,1}), (\mathbf{K}_2, Z_{L,2}), \dots$ which converges to $(\mathbf{K}, Z_L)$ such that $\mathcal{L}_\lambda(\mathbf{K}_i, Z_{L,i}) < \mathcal{L}_\lambda(\mathbf{K}, Z_L)$), then there is a $\mathbf{Z} \in \Gamma^{-1}(\mathbf{K}, Z_L)$ which is a saddle.

First note that for any i, $\Gamma^{-1}(\mathbf{K}_i, Z_{L,i})$ is compact (it is closed and bounded since $\|Z_\ell\|^2_F = \mathrm{Tr}\left[K_\ell\right] < \infty$). There is hence a sequence $\mathbf{Z}_1, \mathbf{Z}_2 \dots$ with $\mathbf{Z}_i \in \Gamma^{-1}\left(\mathbf{K}_i, Z_{L,i}\right)$ which converges to some $\mathbf{Z}$. By the continuity of Γ, we have $\Gamma(\mathbf{Z}) = (\mathbf{K}, Z_L)$ and we have $\mathcal{L}_\lambda\left(\mathbf{Z}_i\right) = \mathcal{L}_\lambda\left(\mathbf{K}_i, Z_{L,i}\right) < \mathcal{L}_\lambda\left(\mathbf{K}, Z_L\right) = \mathcal{L}_\lambda\left(\mathbf{Z}\right)$, hence proving that $\mathbf{Z}$ is a saddle as needed. □

Let us now give an example of a set of weights $\mathbf{W}$ of a depth $L = 2$ network which is a local minimum of $\mathcal{L}_\lambda$ but such that the corresponding covariances $(\mathbf{K}, Z_L)$ are not a local minimum of $\mathcal{L}^k_\lambda$:

Proposition 4. *Consider a shallow ReLU network ($L = 2$) of widths $n_0 = 1, n_1 = 2, n_2 = 1$ with no bias $\beta = 0$. Consider the MSE error $\mathcal{L}_\lambda(\mathbf{W}) = \frac{1}{N}\|Y(X;\mathbf{W}) - Y\|^2_F$ for the size $N = 2$ dataset with inputs $X = \begin{pmatrix} 1 & -1 \end{pmatrix}$ and outputs $Y = \begin{pmatrix} 1 & 1 \end{pmatrix}$.*

For any $\lambda < 1$ and any choices of $a_1, a_2 > 0$ s.t. $a_1^2 + a_2^2 = 1 - \lambda$ the parameters

$$W_1 = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, W_2 = \begin{pmatrix} a_1 & a_2 \end{pmatrix}$$

are a local minimum of the loss $\mathcal{L}_\lambda(\mathbf{W})$ however, the corresponding covariances and outputs $(K_1, K_1^\sigma), Z_2$ are not a local minimum of the second reformulation $\mathcal{L}_\lambda^c((K_1, K_1^\sigma), Z_2)$.

Proof. Consider a depth $L = 2$ network with no bias ($\beta = 0$) and widths $\mathbf{n} = (1, 2, 1)$ with a training set of size $N = 2$, with inputs $X = (1, -1)$ and outputs $Y = (1, 1)$. Let us consider this loss in the region where all four weights are positive:

$$W_1 = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, W_2 = \begin{pmatrix} b_1 & b_2 \end{pmatrix}$$

with $a_1, a_2, b_1, b_2 \geq 0$. We then have the following activations

$$\begin{aligned} Z_1 &= \begin{pmatrix} a_1 & -a_1 \\ a_2 & -a_2 \end{pmatrix} \\ \sigma(Z_1) &= \begin{pmatrix} a_1 & 0 \\ a_2 & 0 \end{pmatrix} \\ Z_2 &= \begin{pmatrix} a_1 b_1 + a_2 b_2 & 0 \end{pmatrix}. \end{aligned}$$

The cost therefore takes the form

$$\mathcal{L}_\lambda(\mathbf{W}) = (1 - a_1 b_1 - a_2 b_2)^2 + 1 + \lambda\left(a_1^2 + a_2^2 + b_1^2 + b_2^2\right).$$

Let us now reformulate the loss in terms of the two positive values

$$\begin{aligned} c &= \left(\frac{a_1 + b_1}{2}\right)^2 + \left(\frac{a_2 + b_2}{2}\right)^2 \\ d &= \left(\frac{a_1 - b_1}{2}\right)^2 + \left(\frac{a_2 - b_2}{2}\right)^2. \end{aligned}$$

Since $2(c + d) = a_1^2 + a_2^2 + b_1^2 + b_2^2$ and $c - d = a_1 b_1 + a_2 b_2$, we can rewrite

$$\mathcal{L}_\lambda(\mathbf{W}) = (1 - c + d)^2 + 1 + 2\lambda(c + d).$$

The above is minimized (over the set of positive c, d) at $c = 1 - \lambda$ and $d = 0$, since it is the unique point of the quarterplane $\left\{ \begin{pmatrix} c \\ d \end{pmatrix} : c, d \geq 0 \right\}$ where the gradient

$$\nabla \mathcal{L}_\lambda(\mathbf{W}) = \begin{pmatrix} \partial_c \mathcal{L}_\lambda(\mathbf{W}) \\ \partial_d \mathcal{L}_\lambda(\mathbf{W}) \end{pmatrix} = \begin{pmatrix} 0 \\ 4\lambda \end{pmatrix}$$

points toward the inside of the quarterplane.

The set weights which optimal amongst the set of positive weights equals the set of positive weights such that $c = 1 - \lambda$ and $d = 0$. Such weights a_1, a_2, b_1, b_2 must satisfy $a_1 = b_1$ and $a_2 = b_2$ (since $d = 0$) and $a_1^2 + a_2^2 = 1 - \lambda$ (since $c = 1 - \lambda$). In other terms, the weights of the form

$$W_1 = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, W_2 = \begin{pmatrix} a_1 & a_2 \end{pmatrix}$$

for any choice of positive a_1, a_2 s.t. $a_1^2 + a_2^2 = 1 - \lambda$ (we have assumed that $\lambda < 1$). For any choice of a_1, a_2 that are both strictly positive, the above weights lie in the inside of the set of positive weights, which implies that these weights form a local minimum.

To prove that the corresponding covariances (K_1, K_1^σ) are not a local minimum of the reformulation, it is sufficient to find a pre-image of these covariances which is not a local minimum. We will show that the extrema of the segment of local minima that we identified are not local minima. Since all weights on the segment have the same covariances, it follows from Proposition 2 that if one of those points is not a local minimum, the covariances cannot be a local minimum of the reformulation.

Let us consider one of the extrema:

$$W_1 = \begin{pmatrix} \sqrt{1-\lambda} \\ 0 \end{pmatrix}, W_2 = \begin{pmatrix} \sqrt{1-\lambda} & 0 \end{pmatrix}.$$

This extremum can be approached by the following weights as $\epsilon \searrow 0$

$$W_1^\epsilon = \begin{pmatrix} \sqrt{1-\lambda} \\ -\epsilon \end{pmatrix}, W_2^\epsilon = \begin{pmatrix} \sqrt{1-\lambda} & -\epsilon \end{pmatrix}.$$

We simply need to show that for small enough ϵ, we have $\mathcal{L}_\lambda(\mathbf{W}^\epsilon) < \mathcal{L}_\lambda(\mathbf{W})$. Let us first compute the activations

$$\begin{aligned} Z_1 &= \begin{pmatrix} \sqrt{1-\lambda} & -\sqrt{1-\lambda} \\ -\epsilon & \epsilon \end{pmatrix} \\ \sigma(Z_1) &= \begin{pmatrix} \sqrt{1-\lambda} & 0 \\ 0 & \epsilon \end{pmatrix} \\ Z_2 &= \begin{pmatrix} 1-\lambda & \epsilon^2 \end{pmatrix}. \end{aligned}$$

Therefore the cost $\mathcal{L}_\lambda(\mathbf{W}^\epsilon)$ takes the form

$$\mathcal{L}_\lambda(\mathbf{W}^\epsilon) = (1-\lambda-1)^2 + (\epsilon^2-1)^2 + 2\lambda\left((1-\lambda)+\epsilon^2\right).$$

Clearly for small enough $\epsilon > 0$, we have $\mathcal{L}_\lambda(\mathbf{W}^\epsilon) < \mathcal{L}_\lambda(\mathbf{W}) = \mathcal{L}_\lambda(\mathbf{W}^{\epsilon=0})$. □

D Description of the Plateau

Proposition 5 (Proposition 8 of the main)**.** *Let $(\mathbf{K}, Z_L) \in \mathcal{K}(X)$, then there are parameters $\mathbf{W}$ of a width $\mathbf{n}$ network with covariances and outputs $\mathbf{K}$ if and only if $n_\ell \geq \mathrm{Rank}_\sigma(K_\ell, K_\ell^\sigma)$ for all $\ell = 1, \ldots, L-1$.*

Proof. To prove that the constraints $n_\ell \geq \mathrm{Rank}_\sigma(K_\ell, K_\ell^\sigma)$ are sufficient, we construct the parameters $\mathbf{W}$ recursively from the first layer to the last. Since $n_1 \geq \mathrm{Rank}_\sigma(K_1, K_1^\sigma)$, there is a hidden representation $Z_1 \in \mathbb{R}^{n_\ell \times N}$ such that $K_\ell = Z_\ell^T Z_\ell$ and $K_\ell^\sigma = (Z_\ell^\sigma)^T Z_\ell^\sigma$ (there is a representation of dimension $\mathrm{Rank}_\sigma(K_1, K_1^\sigma) \times N$, but one can add some zero lines to it to obtain Z_1 without changing the resulting K_ℓ and K_ℓ^σ). Since $\mathrm{Im}Z_1 = \mathrm{Im}K_1 \subset \mathrm{Im}K_0^\sigma = \mathrm{Im}Z_0^\sigma$, we can choose the parameters of the first layer as $W_1 = Z_1 (Z_0^\sigma)^+$. All other weight matrices W_ℓ are then constructed in the same manner.

The fact that the constraints $n_\ell \geq \mathrm{Rank}_\sigma(K_\ell, K_\ell^\sigma)$ are necessary follows from the fact that for any network of width $\mathbf{n}$ with parameters $\mathbf{W}$ we have that $\mathrm{Rank}_\sigma(K_\ell(\mathbf{W}), K_\ell^\sigma(\mathbf{W})) \leq n_\ell$ since $K_\ell(\mathbf{W}) = (Z_\ell(\mathbf{W}))^T Z_\ell(\mathbf{W})$ and $K_\ell^\sigma(\mathbf{W}) = (Z_\ell^\sigma(\mathbf{W}))^T Z_\ell^\sigma(\mathbf{W})$. □

D.1 Tightness of the upper bound

Let us first prove the Proposition on the CP-rank of matrices resulting from graphs without cliques:

Proposition 6 (Proposition 13 of the main)**.** *Given a graph G with N vertices and k edges, consider the $k \times N$ matrix E with entries $E_{ev} = 1$ if the vertex v is an endpoint of the edge e and $E_{ev} = 0$ otherwise. The matrix $A = E^T E$ is completely positive and if the graph G contains no cliques of 3 or more vertices then $\mathrm{Rank}_{\mathrm{cp}}(A) = k$.*

Proof. The fact that $A = E^T E$ implies $\mathrm{Rank}_{\mathrm{cp}}(A) \leq k$, we only need to show $\mathrm{Rank}_{\mathrm{cp}}(A) \geq k$. Let assume that there is another decomposition $E^T E = B^T B$ for some $m' \times N$ matrix B with positive entries, we will now show that $k' \geq k$.

First, we show that the absence of cliques of 3 or more vertices implies that each line B_e has at most 2 non-zero entries. The absence of cliques implies that for all sets $\Omega = \{v_1, \ldots, v_r\}$ of 3 or more vertices, there must be a pair of vertices $v, w \in \Omega$ which are not connected, i.e. $(E^T E)_{vw} = 0$. If

one line B_e contains more than two non-zero entries, corresponding to the vertices $\Omega = \{v_1, \dots, v_r\}$ then for all $v \neq w \in \Omega$, we have

$$\left(B^T B\right)_{vw} \geq \left(B_e B_e^T\right)_{vw} = 1.$$

Now if all lines B_e have at most two non-zero entries it implies that $B_e B_e^T$ has at most two non-zero off-diagonal entries. We know that $E^T E$ has $2k$ non-zero off-diagonal entries. Since

$$E^T E = \sum_{e=1}^{m'} B_e B_e^T$$

it follows that $k' \geq k$, otherwise we could not recover all the off-diagonal entries. □

We may now prove the tightness of the upper bound on the σ-rank of the hidden representation in shallow ReLU networks without bias:

Proposition 7 (Proposition 14 of the main)**.** *Consider a width-*n *shallow network ($L = 2$) with ReLU activation, no bias $\beta = 0$, $n_0 = N$, $n_1 \geq N(N+1)$, input dataset $X_N = I_N$, and any output dataset Y_N such that $\left(Y_N^T Y_N\right)^{\frac{1}{2}}$ is a completely positive matrix with CP-rank k.*

At any global minimum of $R_{\rm n}(X_N, Y_N)$, we have $\mathrm{Rank}_\sigma\left(K_1, K_1^\sigma\right) = k$*. Furthermore for λ small enough, at any global minimum of $\mathcal{L}_{\lambda,{\rm n}}^{\rm MSE}(\mathbf{W}) = \frac{1}{N}\left\|Y(X_N;\mathbf{W}) - Y_N\right\|_F^2 + \lambda\left\|\mathbf{W}\right\|^2$, we have* $\mathrm{Rank}_\sigma\left(K_1, K_1^\sigma\right) \geq k$.

Proof. The proof is in two steps, we first show that the minimizer $\mathbf{K}$ of the representation cost has rank k, and then use this to show that for small enough λs the rank must be at least k.

Representation Cost: We first show that at a minimizer (K_1, K_1^σ) of the cost $\mathrm{Tr}\left[K_1\right] + \mathrm{Tr}\left[YY^T\left(K_1^\sigma\right)^+\right]$, we have $K_1 = K_1^\sigma$. This follows from the fact that if $K_1 \neq K_1^\sigma$, then the pair (K_1^σ, K_1^σ) has a strictly lower cost than the pair (K_1, K_1^σ): for any Z_1 such that $K_1 = Z_1^T Z_1$ and $K_1^\sigma = \sigma(Z_1)^T\sigma(Z_1)$, we have that $\mathrm{Tr}\left[K_\ell\right] = \left\|Z_1\right\|_F^2 \geq \left\|\sigma\left(Z_1\right)\right\|_F^2 = \mathrm{Tr}\left[K_\ell^\sigma\right]$ and the inequality is strict if $Z_1 \neq \sigma(Z_1)$ (which happens iff $K_1 \neq K_1^\sigma$).

The optimization of the previous cost over pairs (K_1, K_1^σ) in S is therefore equivalent to the optimization of the cost $K \mapsto \mathrm{Tr}\left[K\right] + \mathrm{Tr}\left[Y^T Y K^+\right]$ over completely positive matrices K such that $\mathrm{Im}Y \subset \mathrm{Im}K$. If we remove the complete positiveness constraint on K, then the unique minimizer of the above is $K = \left(Y^T Y\right)^{\frac{1}{2}}$. Now since $\left(Y^T Y\right)^{\frac{1}{2}}$ is completely positive, it is also the unique minimizer over complete positive matrices.

We therefore have $\mathrm{Rank}_\sigma\left(K_1, K_1^\sigma\right) = \mathrm{Rank}_{cp}\left(\left(Y^T Y\right)^{\frac{1}{2}}\right) = k$.

Regularized Loss: Let us consider the regularized loss

$$\frac{1}{N}\left\|Z_2 - Y\right\|_F^2 + \lambda \mathrm{Tr}\left[K_1\right] + \lambda \mathrm{Tr}\left[Z_2^T Z_2\left(K_1^\sigma\right)^+\right].$$

The minimizer $\mathbf{K}(\lambda) = (K_1(\lambda), K_1^\sigma(\lambda), Z_2(\lambda))$ converges as $\lambda \searrow 0$ to the pair (K_1, K_1^σ, Y) where $K_1 = K_1^\sigma$ is the minimizer of the representation cost.

Let us now assume that there is no λ_0 such that for all $\lambda < \lambda_0$, any minimizer $\mathbf{K}$ of the loss $\mathcal{L}_\lambda$ satisfies $\mathrm{Rank}_\sigma\left(K_1, K_1^\sigma\right) \geq k$. This would imply that there is a sequence $\lambda_1, \lambda_2, \dots$ of ridges with $\lim_{n\to\infty}\lambda_n = 0$ and corresponding minimizers $\mathbf{K}_1, \mathbf{K}_2, \dots$ (where $\mathbf{K}_n$ is a minimizer of the loss $\mathcal{L}_{\lambda_n}$) such that $\mathrm{Rank}_\sigma\left(K_{n,1}, K_{n,1}^\sigma\right) < k$. Now by Proposition 5 for all n there are parameters $\mathbf{W}_n$ of shallow ReLU network with $n_1 = k-1$ neurons in the hidden layer with covariances equal $\mathbf{K}_n$. The sequence $\mathbf{W}_1, \mathbf{W}_2, \dots$ is uniformly bounded in norm by the representation cost $R(X_N, Y_N)$, there is therefore a converging subsequence $\mathbf{W}_{n_1}, \mathbf{W}_{n_2}, \dots$ which converges to some parameters $\mathbf{W}$. The covariances and outputs (K_1, K_1^σ, Y) at these limiting parameters $\mathbf{W}$ must minimize the representation cost, i.e. $K_1 = K_1^\sigma = \left(Y^T Y\right)^{\frac{1}{2}}$, but this yields a contradiction, since $\mathrm{Rank}_\sigma\left(K_1, K_1^\sigma\right) = k$ but $\mathbf{W}$ are parameters of network with $n_1 = k-1$ neurons in the hidden layer, which would imply $\mathrm{Rank}_\sigma\left(K_1, K_1^\sigma\right) \leq k-1$. □

To show the tightness (up to constant factor) of the upper bound, one can simply apply this proposition to the special case $Y_N = E^T E$, where E is the edge-vertex incidence matrix of the complete bipartite graph, in which case $k = \frac{N^2}{4}$.

We could also consider an output dataset $Y_N \in \mathbb{R}^{n_L \times N}$ whose lines are one-hot vectors, corresponding to a classification task. If we reorder the training set by class, the covariance $Y_N^T Y_N$ is a block diagonal matrix, with all ones blocks corresponding to each class. The square root $\left(Y_N^T Y_N\right)^{\frac{1}{2}}$ is also block-diagonal but the block of a class i has value $\frac{1}{m_i}$ where m_i is the number of datapoints in the class i. The matrix $\left(Y_N^T Y_N\right)^{\frac{1}{2}}$ is completely positive and has rank k equal to the number of classes. This implies a much earlier plateau, which could explain why in real-world classification tasks, we observe a very early plateau.

Remark 8. The representation cost for $Y = E^T E$ is $2\|E\|_F^2 = 4\frac{N^2}{4} = N^2$. We can obtain an almost optimal representation with $n_1 = N$ neurons by taking the weights $W_1 = \sqrt{\frac{N}{2}} I$ and $W_2 = \sqrt{\frac{2}{N}} E^T E$, with norm $\left\|\sqrt{\frac{N}{2}} I\right\|_F^2 + \left\|\sqrt{\frac{2}{N}} E^T E\right\|_F^2 = \frac{N^2}{2} + \frac{2}{N}(N\frac{N^2}{4} + 2\frac{N^2}{4}) = \frac{N^2}{2} + \frac{N^2}{2} + N = N^2 + N$.

D.2 One Dimensional Shallow Network

We now prove an upper bound on the start of the plateau for shallow networks with one-dimensional inputs and outputs:

Proposition 9 (Proposition 15 of the main)**.** *Consider shallow networks ($L = 2$) with scalar inputs and outputs ($n_0 = n_2 = 1$), a ReLU nonlinearity, and a dataset $X, Y \in \mathbb{R}^{1 \times N}$. Both the representation cost $R_{\mathbf{n}}(X, Y)$ and global minimum $\min_{\mathbf{W}} \mathcal{L}_{\lambda,\mathbf{n}}(\mathbf{W})$ for any $\lambda > 0$ are constant as long as $n_1 \geq 4N$.*

Proof. We show that if there is a network with depth $L = 2$ and $n_1 > 4N$ hidden neurons, we can construct a network with strictly less neurons with the same outputs on the dataset and a smaller parameter norm.

The network function can be written in the form

$$f_{\mathbf{W}}(x) = b + \sum_{k=1}^{n_1} a_k \sigma\left(c_k x + d_k\right).$$

We may assume that for all neuron i, we have $a_k^2 = c_k^2 + d_k^2$ since if this is not the case, one can multiply a_k by a scalar and divide c_k and d_k by the same scalar to satisfy this constraint while reducing the norm of the parameters.

For each neuron i, we define the cusp of the neuron the value $-\frac{d_k}{c_k}$, which is the point where the neuron goes from dead to active.

If there are neurons that are inactive on the whole training set, they can simply be removed without changing the outputs and reducing the norm.

If there are more $4N$ neurons, we either have:

1. There are more than 4 neurons whose cusp lies between two inputs x_i and x_{i+1} (w.l.o.g. we assume $x_1 < \cdots < x_N$).
2. There are more than 2 neurons whose cusp lies to the left or right of the data.

We will now show how in the case 1, one can remove a neuron while keeping the same outputs on the training data and reducing the norm of the parameters. The second case is analogous.

If there are five or more neurons with a cusp between x_i and x_{i+1}, then two of those neurons k, m must have the same signs $\text{sign} a_k = \text{sign} a_m$ and $\text{sign} c_k = \text{sign} c_m$ (w.l.o.g. we assume they are all

positive). We will replace these two neurons by a single neuron $\tilde{a}\sigma(\tilde{c}x + \tilde{d})$ where $\tilde{a}, \tilde{c}, \tilde{d}$ are chosen as the unique positive values ($\tilde{d}$ may be negative) to satisfy

$$\begin{aligned}\tilde{a}\tilde{c} &= a_k c_k + a_m c_m \\ \tilde{a}\tilde{d} &= a_k d_k + a_m d_m \\ \tilde{a}^2 &= \tilde{c}^2 + \tilde{d}^2.\end{aligned}$$

First note this new neurons contributes $\tilde{a}^2 + \tilde{c}^2 + \tilde{d}^2 = 2\tilde{a}^2$ to the norm of the parameters which is less than the two previous neurons $2a_k^2 + 2a_m^2$, since

$$\begin{aligned}\tilde{a}^2 &= \sqrt{\tilde{a}^2\left(\tilde{c}^2 + \tilde{d}^2\right)} \\ &= \sqrt{\left(a_k d_k + a_m d_m\right)^2 + \left(a_k c_k + a_m c_m\right)^2} \\ &= (a_k + a_m)\sqrt{\left(\frac{a_k}{a_k + a_m} d_k + \frac{a_m}{a_k + a_m} d_m\right)^2 + \left(\frac{a_k}{a_k + a_m} c_k + \frac{a_m}{a_k + a_m} c_m\right)^2} \\ &\leq (a_k + a_m)\left(\frac{a_k}{a_k + a_m}\sqrt{d_k^2 + c_k^2} + \frac{a_m}{a_k + a_m}\sqrt{d_m^2 + c_m^2}\right) \\ &= a_k^2 + a_m^2\end{aligned}$$

where the inequality follows from the convexity of the norm function $(c, d) \mapsto \sqrt{c^2 + d^2}$.

For any x with $x \leq x_i$ or $x \geq x_{i+1}$, one can check that

$$\tilde{a}\sigma(\tilde{c}x + \tilde{d}) = a_k\sigma(c_k x + d_k) + a_m\sigma(c_m x + d_m),$$

which implies that replacement has not changed the values of the network on the training set. □

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998.