Bottleneck Structure in Learned Features: Low-Dimension vs Regularity Tradeoff

Anonymous Author(s)
Affiliation
Address
email

1 Corrections

1.1 First Correction

Theorem 1 (Theorem 3 from the main)**.** *For all inputs x where $\mathrm{Rank}Jf(x) = R^{(0)}(f;\Omega)$, we have $R^{(1)}(f) \geq 2\log|Jf(x)|_+$, furthermore:*

1. *If $R^{(0)}(f\circ g) = R^{(0)}(f) = R^{(0)}(g)$, then $R^{(1)}(f\circ g) \leq R^{(1)}(f) + R^{(1)}(g)$.*
2. *If $R^{(0)}(f+g) = R^{(0)}(f) + R^{(0)}(g)$, then $R^{(1)}(f+g) \leq R^{(1)}(f) + R^{(1)}(g)$.*
3. *If $P_{\mathrm{Im}A^T}\Omega$ and $A\Omega$ are $k = \mathrm{Rank}A$ dimensional and completely positive (i.e. they can be embedded with an isometric linear map into $\mathbb{R}^m_+$ for some m), then $R^{(1)}(x\mapsto Ax;\Omega) = 2\log|A|_+$.*

Proof. For the first bound, we remember that $R(f;\Omega,L) \geq L\left\|Jf\right\|_{2/L}^{2/L}$, therefore

$$R^{(1)}(f;\Omega) = \lim_{L\to\infty} R(f;\Omega,L) - LR^{(0)}(f;\Omega) \geq \lim_{L\to\infty} L\sum_{i=1}^{\mathrm{Rank}Jf(x)} s_i(Jf(x))^{\frac{2}{L}} - 1 \geq \sum_{i=1}^{\mathrm{Rank}Jf(x)} 2\log s_i(Jf(x))$$

where we used $s^{\frac{2}{L}} - 1 = e^{\frac{2}{L}\log s} - 1 \geq \frac{2}{L}\log s$.

(1) Since $R(f\circ g;\Omega,L_1+L_2) \leq R(f;L_1) + R(g;L_2)$, we have

$$\begin{aligned} R^{(1)}(f\circ g;\Omega) &= \lim_{L_1+L_2\to\infty} R(f\circ g;\Omega,L_1+L_2) - (L_1+L_2)R^{(0)}(f\circ g;\Omega) \\ &\leq \lim_{L_1\to\infty} R(f;\Omega,L_1) - L_1R^{(0)}(f;\Omega) + \lim_{L_2\to\infty} R(g;\Omega,L_2) - L_2R^{(0)}(f;\Omega) \\ &= R^{(1)}(f;\Omega) + R^{(1)}(g;\Omega). \end{aligned}$$

(2) Since $R(f+g;\Omega,L) \leq R(f;\Omega,L) + R(g;\Omega,L)$, we have

$$\begin{aligned} R^{(1)}(f+g;\Omega) &= \lim_{L\to\infty} R(f+g;\Omega,L) - LR^{(0)}(f+g;\Omega) \\ &\leq \lim_{L\to\infty} R(f;\Omega,L) - LR^{(0)}(f;\Omega) + \lim_{L\to\infty} R(g;\Omega,L) - LR^{(0)}(g;\Omega) \\ &= R^{(1)}(f;\Omega) + R^{(1)}(g;\Omega). \end{aligned}$$

(3) By the first bound, we know that $R^{(1)}(x\mapsto Ax;\Omega) \geq 2\log|A|_+$, we now need to show $R^{(1)}(x\mapsto Ax;\Omega) \leq 2\log|A|_+$. Let us define the set of completely positive representations as the

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

set of bilinear kernels $K(x,y) = x^T B^T B y$ such that Bx has non-negative entries for all $x \in \Omega$ (we say that a kernel K is completely positive over Ω if it can be represented in this way for some choice of B). The set of completely positive representations is convex, since for $K(x,y) = x^T B^T B y$ and $\tilde{K}(x,y) = x^T \tilde{B}^T \tilde{B} y$, we have

$$\frac{K(x,y)+\tilde{K}(x,y)}{2} = x^T \begin{pmatrix} \frac{1}{\sqrt{2}} B \\ \frac{1}{\sqrt{2}} \tilde{B} \end{pmatrix}^T \begin{pmatrix} \frac{1}{\sqrt{2}} B \\ \frac{1}{\sqrt{2}} \tilde{B} \end{pmatrix} y.$$

The conditions that there are O_{in} and O_{out} with $O_{in}^T O_{in} = P_{\mathrm{Im} A^T}$ and $O_{out}^T O_{out} = P_{\mathrm{Im} A}$ such that $O_{in}\Omega \in \mathbb{R}_+^{k_1}$ and $O_{out} A \Omega \in \mathbb{R}_+^{k_2}$ is equivalent to saying that the kernels $K_{in}(x,y) = x^T P_{\mathrm{Im} A^T} y$ and $K_{out}(x,y) = x^T A^T A x$ are completely positive over Ω.

By the convexity of completely positive representations, the interpolation $K_p = pK_{in} + (1-p)K_{out}$ is completely positive for all $p \in [0,1]$. Now choose for all depths L and all layers $\ell = 1, \dots, L-1$ a matrix $B_{L,\ell}$ such that $K_{p=\frac{\ell}{L}}(x,y) = x^T B_{L,\ell}^T B_{L,\ell} y$ and then choose the weights W_ℓ of the depth L network as

$$W_\ell = B_{L,\ell} B_{L,\ell-1}^+,$$

using the convention $B_{L,0} = I_{d_{in}}$ and $B_{L,L} = I_{out}$. By induction, we show that for any input $x \in \Omega$ the activation of the ℓ-th hidden layer is $B_{L,\ell} x$. This is true for $\ell = 1$, since $W_1 = B_{L,1}$ and therefore $p^{(1)}(x) = B_{L,1} x$ which has positive entries so that $q^{(1)}(x) = \sigma\left(p^{(1)}(x)\right) = B_{L,1} x$. Then by induction

$$p^{(\ell)}(x) = W_\ell q^{(\ell-1)}(x) = B_{L,\ell} B_{L,\ell-1}^+ B_{L,\ell-1} x = B_{L,\ell} x,$$

which has positive entries, so that again $q^{(\ell)}(x) = \sigma\left(p^{(\ell)}(x)\right) = B_{L,\ell} x$. In the end, we get $p^{(L)}(x) = Ax$ as needed.

Let us now compute the Frobenius norms of the weight matrices $\|W_\ell\|_F^2 = \mathrm{Tr}\left[B_{L,\ell}^T B_{L,\ell} \left(B_{L,\ell-1}^T B_{L,\ell-1}\right)^+\right]$ as $L \to \infty$, remember that $B_{L,\ell}^T B_{L,\ell} = \frac{\ell}{L} P_{\mathrm{Im} A^T} + (1-\frac{\ell}{L}) A^T A$, therefore the matrices $B_{L,\ell}^T B_{L,\ell}$ and $B_{L,\ell-1}^T B_{L,\ell-1}$ converge to each other, so that at first order $B_{L,\ell}^T B_{L,\ell} \left(B_{L,\ell-1}^T B_{L,\ell-1}\right)^+$ converges to $P_{\mathrm{Im} A^T}$, so that $\|W_\ell\|_F^2 \to \mathrm{Rank} A$, so that $\sum_{\ell=1}^L \|W_\ell\|_F^2 - L\mathrm{Rank} A$ converges to a finite value as $L \to \infty$. To obtain this finite limit, we need to study approximate the next order

$$\begin{aligned}
\|W_\ell\|_F^2 - \mathrm{Rank} A &= \sum_{i=1}^{\mathrm{Rank} A} 2 \log s_i(W_i) + O(L^{-2}) \\
&= \log \left| B_{L,\ell}^T B_{L,\ell} \left(B_{L,\ell-1}^T B_{L,\ell-1}\right)^+ \right|_+ + O(L^{-2}) \\
&= \log \left| B_{L,\ell}^T B_{L,\ell} \right|_+ - \log \left| B_{L,\ell-1}^T B_{L,\ell-1} \right|_+ + O(L^{-2}).
\end{aligned}$$

But as we sum all these second order terms, they cancel out, and we are left with

$$\sum_{\ell=1}^L \|W_\ell\|_F^2 - L\mathrm{Rank} A = 2 \log |A|_+ - 2 \log |I_{\mathrm{Im} A^T}|_+ + O(L^{-1}).$$

We have therefore build parameters θ that represent the function $x \mapsto Ax$ with parameter norm $\|\theta\|^2 = L\mathrm{Rank} A + 2 \log |A|_+ + O(L^{-1})$, which upper bounds the representation cost, thus implying that $R^{(1)}(x \mapsto Ax; \Omega) \leq 2 \log |A|_+$ as needed. □

1.2 Identity

Proposition 2 (Proposition 4 from the main)**.** *For a domain with* $\mathrm{Rank}_J(id;\Omega) = \mathrm{Rank}_{BN}(id;\Omega) = k$*, then* Ω *is* k*-planar if* $R^{(1)}(id;\Omega) = 0$*.*

Proof. We will show that for any two points $x, y \in \Omega$ with k-dim tangent spaces, their tangent spaces must match if $R^{(1)}(id;\Omega) = 0$.

Let $A = J\alpha^{(L-1)}(x)_{|T_x\Omega}$ and $B = J\alpha^{(L-1)}(y)_{|T_y\Omega}$ be the be the Jacobian of the last hidden activations restricted to the tangent spaces, we know that

$$P_{T_x\Omega} = W_L A$$
$$P_{T_y\Omega} = W_L B$$

so that given any weight matrix W_L whose image contains $T_x\Omega$ and $T_y\Omega$, we can write

$$A = W_L^+ P_{T_x\Omega}$$
$$B = W_L^+ P_{T_y\Omega}.$$

Without loss of generality, we may assume that the span of $T_x\Omega$ and $T_y\Omega$ is full output space, and therefore that $W_L W_L^T$ is invertible.

Now we now that any parameters that represent the identity on Ω and has $A = J\alpha^{(L-1)}(x)_{|T_x\Omega}$ and $B = J\alpha^{(L-1)}(y)_{|T_y\Omega}$ must have parameter norm at least

$$\|W_L\|_F^2 + k(L-1) + \max\left\{2\log|A|_+, 2\log|B|_+\right\}.$$

Subtracting kL and taking $L \to \infty$, we obtain that

$$R^{(1)}(id;\Omega) \geq \min_{W_L} \|W_L\|_F^2 - k + \max\left\{2\log\left|W_L^+ P_{T_x\Omega}\right|_+, 2\log\left|W_L^+ P_{T_y\Omega}\right|_+\right\}.$$

If we optimize W_L only up to scaling (i.e. optimize aW_L over a) we see that at the optimum, we always have $\|W_L\|_F^2 = k$. This allows us to rewrite the optimization as

$$R^{(1)}(id;\Omega) \geq \min_{\|W_L\|_F^2 = k,} \max\left\{2\log\left|W_L^+ P_{T_x\Omega}\right|_+, 2\log\left|W_L^+ P_{T_y\Omega}\right|_+\right\}.$$

The only way to put the first term inside the maximum to 0 is to have $W_L W_L^T = P_{T_x\Omega}$, but this leads to an exploding second term if $P_{T_x\Omega} \neq P_{T_y\Omega}$. □

Under the assumption of uniform Lipschitzness, one can show a stronger version of the above:

Proposition 3. *For a C-uniformly Lipschitz sequence of ReLU networks representing the function f, we have*

$$R^{(1)}(f) \geq \log|Jf(x)|_+ + \log|Jf(y)|_+ + C^{-2}\|Jf_\theta(x) - Jf_\theta(y)\|_*.$$

Proof. The decomposition of the difference

$$Jf_\theta(x) - Jf_\theta(y) = \sum_{\ell=1}^{L-1} W_L D_{L-1}(y)\cdots W_{\ell+1}\left(D_\ell(x) - D_\ell(y)\right) W_\ell D_{\ell-1}(x)\cdots D_1(x) W_1,$$

for the $w_\ell \times w_\ell$ diagonal matrices $D_\ell(x) = \mathrm{diag}(\dot\sigma(\tilde\alpha_\ell(x)))$, implies the bound

$$\begin{aligned}\|Jf_\theta(x) - Jf_\theta(y)\|_* &\leq \sum_{\ell=1}^{L-1} \|W_L D_{L-1}(y)\cdots D_{\ell+1}(y)\|_{op} \|W_{\ell+1}\left(D_\ell(x) - D_\ell(y)\right) W_\ell\|_* \|D_{\ell-1}(x)\cdots D_1(x) W_1\|_{op} \\ &\leq \frac{C^2}{2}\sum_{\ell=1}^{L-1}\left(\|W_{\ell+1}\left(D_\ell(x) - D_\ell(y)\right)\|_F^2 + \|\left(D_\ell(x) - D_\ell(y)\right) W_\ell\|_F^2\right)\end{aligned}$$

since $\|AB\|_* \leq \frac{\|A\|_F^2 + \|B\|_F^2}{2}$.

Now since

$$L\,\|Jf_\theta(x)\|_{2/L}^{2/L} \leq \frac{1}{2}\sum_{\ell=1}^{L}\|W_\ell D_{\ell-1}(x)\|_F^2$$
$$L\,\|Jf_\theta(x)\|_{2/L}^{2/L} \leq \frac{1}{2}\sum_{\ell=1}^{L}\|D_\ell(x) W_\ell\|_F^2$$

with the convention $D_0(x) = I_{d_{in}}$ and $D_L(x) = I_{d_{out}}$. We obtain that

$$L \left\| Jf_\theta(x) \right\|_{2/L}^{2/L} + L \left\| Jf_\theta(y) \right\|_{2/L}^{2/L} \leq \frac{1}{2} \sum_{\ell=1}^{L} \left\| W_\ell D_{\ell-1}(x) \right\|_F^2 + \left\| W_\ell D_{\ell-1}(y) \right\|_F^2 + \left\| D_\ell(x) W_\ell \right\|_F^2 + \left\| D_\ell(y) W_\ell \right\|_F^2$$

$$\leq \sum_{\ell=1}^{L} 2 \left\| W_\ell \right\|_F^2 - \frac{1}{2} \left\| W_\ell \left(D_{\ell-1}(x) - D_{\ell-1}(y) \right) \right\|_F^2 - \frac{1}{2} \left\| \left(D_\ell(x) - D_\ell(y) \right) W_\ell \right\|_F^2 .$$

This implies the bound

$$\left\| \theta \right\|^2 \geq \frac{L \left\| Jf_\theta(x) \right\|_{2/L}^{2/L} + L \left\| Jf_\theta(y) \right\|_{2/L}^{2/L}}{2} + C^{-2} \left\| Jf_\theta(x) - Jf_\theta(y) \right\|_*$$

and thus

$$R^{(1)}(f) \geq \log \left| Jf(x) \right|_+ + \log \left| Jf(y) \right|_+ + C^{-2} \left\| Jf_\theta(x) - Jf_\theta(y) \right\|_* .$$

□

1.3 Second Correction

Proposition 4 (Proposition 11 from the main)**.** *If there is a limiting representation as $L \to 0$ in the optimal representation of f, then $R^{(2)}(f) \geq 0$. Furthermore:*

1. *If $R^{(0)}(f \circ g) = R^{(0)}(f) = R^{(0)}(g)$ and $R^{(1)}(f \circ g) = R^{(1)}(f) + R^{(1)}(g)$, then $\sqrt{R^{(2)}(f \circ g)} \leq \sqrt{R^{(2)}(f)} + \sqrt{R^{(2)}(g)}$.*
2. *If $R^{(0)}(f+g) = R^{(0)}(f) + R^{(0)}(g)$ and $R^{(1)}(f+g) = R^{(1)}(f) + R^{(1)}(g)$, then $R^{(2)}(f+g) \leq R^{(2)}(f) + R^{(2)}(g)$.*
3. *If $A^p\Omega$ is $k = \mathrm{Rank}A$-dimensional and completely positive for all $p \in [0,1]$, where A^p has its non-zero singular taken to the p-th power, then $R^{(2)}(x \mapsto Ax; \Omega) = \frac{1}{2} \left\| \log_+ A^T A \right\|^2$.*

Proof. We start from the inequality

$$R(f \circ g; \Omega, L_f + L_g) \leq R(f; g(\Omega), L_f) + R(g; \Omega, L_g).$$

We subtract $(L_f + L_g) R^{(0)}(f \circ g) + R^{(1)}(f \circ g)$ divide by $L_f + L_g$ and take the limit of increasing depths L_f, L_g with $\lim_{L_g, L_f \to \infty} \frac{L_f}{L_f + L_g} = p \in (0,1)$ to obtain

$$R^{(2)}(f \circ g; \Omega) \leq \frac{1}{1-p} R^{(2)}(f; g(\Omega)) + \frac{1}{p} R^{(2)}(g; \Omega). \tag{1}$$

If K_p is the limiting representation at a ratio $p \in (0,1)$, we have $R^{(2)}(f; \Omega) = \frac{1}{p} R^{(2)}(K_p; \Omega) + \frac{1}{1-p} R^{(2)}(K_p \to f; \Omega)$ and p must minimize the RHS since if it was instead minimized at a different ratio $p' \neq p$, one could find a lower norm representation by mapping to K_p in the first $p'L$ layers and then back to the outputs. Now there are two possiblities, either $R^{(2)}(K_p; \Omega)$ and $R^{(2)}(K_p \to f; \Omega)$ are non-negative in which case the minimum is attained at some $p \in (0,1)$ and $R^{(2)}(f; \Omega) \geq 0$, or one or both is negative in which case the above is minimized at $p \in \{0,1\}$ and $R^{(2)}(f; \Omega) = -\infty$. Since we assumed $p \in (0,1)$, we are in the first case.

(1) To prove the first property, we optimize the RHS of 1 over all possible choices of p (and assuming that $R^{(2)}(f; g(\Omega)), R^{(2)}(g; \Omega) \geq 0$) we obtain

$$\sqrt{R^{(2)}(f \circ g; \Omega)} \leq \sqrt{R^{(2)}(f; g(\Omega))} + \sqrt{R^{(2)}(g; \Omega)}.$$

(2) This follows from the inequality $R(f+g; \Omega, L) \leq R(f; g(\Omega), L) + R(g; \Omega, L)$ after subtracting the $R^{(0)}$ and $R^{(1)}$ terms, dividing by L and taking $L \to \infty$.

(3) If $A = USV^T$, one chooses $W_\ell = U_\ell S^{\frac{1}{L}} U_{\ell-1}^T$ with $U_0 = V$, $U_L = U$ and U_ℓ chosen so that $U_\ell S^{\frac{\ell}{L}} V^T \Omega \in \mathbb{R}_+^{n_\ell}$, choosing large enough widths n_ℓ. This choice of representation of A is optimal,

i.e. its parameter norm matches the representation cost $L\mathrm{Tr}\left[S^{\frac{2}{L}}\right] = L\mathrm{Rank}A + 2\log|A|_+ + \frac{1}{2L}\left\|\log_+ A^TA\right\|^2 + O(L^{-2})$.

We know that

$$\lim_{L\to\infty} R^{(1)}(\alpha_{\ell_1} \to \alpha_{\ell_2}; \Omega) = R^{(1)}(f_\theta; \Omega) \lim_{L\to\infty} \frac{\ell_2 - \ell_1}{L}$$

$$\frac{1}{p}R^{(2)}(\alpha; \Omega) + \frac{1}{1-p}R^{(2)}(\alpha \to f; \Omega) \geq R^{(2)}(f; \Omega)$$

$$\frac{1}{p}R^{(2)}(\alpha; \Omega) + \frac{1}{1-p}R^{(2)}(\alpha \to f; \Omega) \geq R^{(2)}(f; \Omega)$$

□

2 Local Minima Stability

In this section we motivate the assumption that the Jacobian $J\tilde{\alpha}_\ell(x)$ is uniformly bounded in operator norm as $L \to \infty$. The idea is that solutions with a blowing up Jacobian $J\tilde{\alpha}_\ell(x)$ correspond to very narrow local minima.

The narrowness of a local minimum is related to the Neural Tangent Kernel (or Fisher matrix). We have that

$$\mathrm{Tr}\left[\Theta^{(L)}(x,x)\right] = \sum_{\ell=1}^{L} \|\alpha_{\ell-1}(x)\|^2 \|J(\tilde{\alpha}_\ell \to f_\theta)(x)\|_F^2 .$$

If the Jacobian $J\tilde{\alpha}_\ell(x)$ blows up, it is reasonable to expect both $\|\alpha_{\ell-1}(x)\|^2$ and $\|J(\tilde{\alpha}_\ell \to f_\theta)(x)\|_F^2$ to blow up too:

1. If $\|J\tilde{\alpha}_{\ell-1}(x)\|_{op}$ is very large then small variation of the inputs x can lead to very large preactivations $\|\tilde{\alpha}_{\ell-1}(x)\|^2$ and thus activations $\|\alpha_{\ell-1}(x)\|^2$ (if we use the ReLU one needs to also assume $\tilde{\alpha}_{\ell-1}(x)$ does not have mostly negative entries).
2. Since $\det J\tilde{\alpha}_\ell(x)$ is bounded, a blowing up top singular value $\|J\tilde{\alpha}_\ell(x)\|_{op}$ implies that the k-th singular value must implode for $k = \mathrm{Rank}_{BN}(f_\theta; \Omega)$. More precisely, if $\|J\tilde{\alpha}_\ell(x)\|_{op}$ is of order L^γ for $\gamma > 1$, then the k-th singular value is of order $L^{-\frac{\gamma}{k-1}}$ or less, and since

$$Jf_\theta(x) = J(\tilde{\alpha}_\ell \to f_\theta)(x) J\tilde{\alpha}_\ell(x)$$

we need $\|J(\tilde{\alpha}_\ell \to f_\theta)(x)\|_{op}$ to be at least of order $L^{\frac{\gamma}{k-1}}$ so that the k-th singular value of $Jf_\theta(x)$ remains of order 1.

This suggests that in non-uniformly Lipschitz sequences, the NTK would blow up at a rate faster than L, whereas uniformly Lipschitz sequence have a NTK of order L. In simpler terms non-uniformly Lipschitz sequences of local minima are infinitely narrower than their uniformly Lipschitz counterpart, which suggests that finite learning rate GD is naturally biased towards uniformly Lipschitz local minima.

Theorem 5 (Theorem 5 from the main)**.** *For any point x, we have*

$$\left\|\partial^2_{xy}\Theta(x,x)\right\|_{op} \geq 2L \left\|Jf_\theta(x)\right\|_{op}^{2-2/L}$$

where $\partial^2_{xy}\Theta(x,x)$ is understood as a $d_{in}d_{out} \times d_{in}d_{out}$ matrix.

Furthermore, for any two points x, y such that the pre-activations of all neurons of the network remain constant on the segment $[x, y]$, then either $\|\Theta(x,x)\|_{op}$ or $\|\Theta(y,y)\|_{op}$ is lower bounded by $\frac{L}{4}\|x-y\|^2 \left\|Jf_\theta(x)\frac{y-x}{\|x-y\|}\right\|_2^{2-2/L}$.

Proof. (1) For any point x, we have

$$\partial_{x,y}\left(v^T\Theta(x,x)v\right)[u,u] = \sum_{\ell=1}^{L} u^T W_1^T D_1(x)\cdots D_{\ell-1}(x)^2\cdots D_1(x)W_1 u v^T W_L D_{L-1}(x)\cdots D_\ell(x)^2\cdots D_{L-1}(x)W_L^T v$$
$$= \sum_{\ell=1}^{L} \|D_{\ell-1}(x)\cdots D_1(x)W_1 u\|_2^2 \, \|D_\ell(x)\cdots D_{L-1}(x)W_L v\|_2^2 .$$

On the other hand, we have

$$\left|v^T Jf_\theta(x)u\right| = \left|v^T W_L D_{L-1}(x)\cdots D_1(x)W_1 u\right|$$
$$\leq \|D_\ell(x)\cdots D_1(x)W_1 u\|_2 \, \|D_\ell(x)\cdots D_{L-1}(x)W_L v\|_2 ,$$

where we used the fact that $D_\ell(x)D_\ell(x) = D_\ell(x)$. This applies to the case $\ell = L$ and $\ell = 1$ too, using the definition $D_L(x) = I_{d_{out}}$ and $D_0(x) = I_{d_{in}}$. This implies

$$\partial_{xy}^2\left(v^T\Theta(x,x)v\right)[u,u] \geq \left|v^T Jf_\theta(x)u\right|^2 \sum_{\ell=1}^{L} \frac{\|D_{\ell-1}(x)\cdots D_1(x)W_1 u\|_2^2}{\|D_\ell(x)\cdots D_1(x)W_1 u\|_2^2}$$
$$\geq \left|v^T Jf_\theta(x)u\right|^2 L\left(\frac{\|u\|_2^2}{\|W_L D_{L-1}(x)\cdots D_1(x)W_1 u\|_2^2}\right)^{\frac{1}{L}}$$
$$\geq L\frac{\left|v^T Jf_\theta(x)u\right|^2}{\|Jf_\theta(x)u\|_2^{2/L}}.$$

where we used the geometric/arithmetic mean inequality for the second inequality.

If u, v are right and left singular vectors of $Jf_\theta(x)$ with singular value s, then the above bound equals $Ls^{2-\frac{2}{L}}$.

(2) Now let us consider a segment $\gamma(t) = (1-t)x + ty$ between two points x, y with no changes of activations on these paths (i.e. $D_\ell(\gamma(t))$ is constant for all $t \in [0,1]$). Defining $u = \frac{y-x}{\|y-x\|}$ and $v = \frac{Jf_\theta(x)u}{\|Jf_\theta(x)u\|}$, we have

$$\partial_t v^T\Theta(\gamma(t),\gamma(t))v = \|x-y\|\,\partial_x\left(v^T\Theta(\gamma(t),\gamma(t))v\right)[u] + \|x-y\|\,\partial_y\left(v^T\Theta(\gamma(t),\gamma(t))v\right)[u]$$

and since $\partial_{xx}\Theta(\gamma(t),\gamma(t)) = 0$ and $\partial_{yy}\Theta(\gamma(t),\gamma(t)) = 0$ for all $t \in [0,1]$, we have

$$\partial_t^2\left(v^T\Theta(\gamma(t),\gamma(t))v\right) = 2\|x-y\|^2\,\partial_{xy}^2\left(v^T\Theta(\gamma(t),\gamma(t))v\right)[u,u] \geq 2L\|x-y\|^2\|Jf_\theta(x)u\|_2^{2-2/L}.$$

Since $v^T\Theta(\gamma(t),\gamma(t))v \geq 0$ for all $t \in [0,1]$ then either

$$v^T\Theta(x,x)v \geq \frac{L}{4}\|x-y\|^2\|Jf_\theta(x)u\|_2^{2-2/L}$$

or

$$v^T\Theta(y,y)v \geq \frac{L}{4}\|x-y\|^2\|Jf_\theta(x)u\|_2^{2-2/L}.$$

□

Now for the proof that rank-underestimating functions require exploding Jacobians:

Proposition 6 (Proposition 7 from the main)**.** *Let $f^* : \Omega \to \mathbb{R}^{d_{in}}$ for a bounded domain Ω and with $\mathrm{Rank}_J(f^*;\Omega) = m$, then with high prob. over the sampling of N i.i.d. inputs $x_1,\ldots,x_N \in \Omega$, any BN-rank function $\hat{f} = h\circ g$ that fits the data $\hat{f}(x_i) = f^*(x_i)$ must satisfy*

$$\sup_{x\in\Omega}\|Jg(x)\|_{op}\sup_{z\in g(\Omega)}\|Jh(z)\|_{op} = \Omega\left(N^{\frac{1}{m}-\frac{1}{k}}\right).$$

Proof. Writing $z_i = g(x_i) \in g(\Omega) \subset \mathbb{R}^k$, the shortest path through $z_1,\ldots,z_N$ has length at most of order $\mathrm{diam}g(\Omega)N^{1-\frac{1}{k}}$ [2]. But this path is mapped by h to a path through the random points $y_1,\ldots,y_N$ which must have length of order at least $N^{1-\frac{1}{m}}$ with high probability [1] since the support of the distribution of the y_is is m-dimensional. This implies the proposition. □

3 Representation Geodesics

Theorem 7 (Theorem 9 from the main)**.** *Consider a sequence $(\theta_L)_L$ of representations of a function $f : \Omega \to \mathbb{R}^{d_{out}}$ with $R^{(0)}(f;\Omega) = \mathrm{Rank}_J(f;\Omega)$ and $R^{(1)}(f;\Omega) < \infty$. Then any accumulating representation K_p at a ratio p is k-planar, i.e. there are k features $\phi_1, \dots, \phi_k$ such that $K_p(x,y) = \phi(x)^T\phi(y)$.*

Proof. Let K_p be an accumulating representation, then it must satisfy $R^{(1)}(K_p \to K_p;\Omega) = 0$: for any M we can find M sequences $\left(\ell_L^{(1)}\right)_L, \dots, \left(\ell_L^{(M)}\right)_L$ each separated by an infinite amount of layers as $L \to \infty$ that all converge to the representation K_p; if $R^{(1)}(K_p \to K_p) > 0$ then the overall first correction would be infinite, since $R^{(1)}(\lim_{L\to\infty} f_{\theta_L};\Omega) \geq MR^{(1)}(K_p \to K_p;\Omega)$ for all M. Then by Proposition 2 we obtain that K_p must be k-planar. □

4 Technical Results

4.1 Regularity Counterexample

We guve here an example of a simple function whose optimal representation geodesic does not converge, due to it being not uniformly Lipschitz:

Example 8. The function $f : \Omega \to \mathbb{R}^3$ with $\Omega = [0,1]^3$ defined by

$$f(x,y,z) = \begin{cases} (x,y,z) & \text{if } x \leq y \\ (x,y,z + a(x-y)) & \text{if } x > y \end{cases}$$

satisfies $R^{(0)}(f;\Omega) = 3$ and $R^{(1)}(f;\Omega) = 0$. The optimal representations of f are not uniformly Lipschitz as $L \to \infty$.

Proof. While we are not able to identify exactly the optimal representation geodesic for the function f, we will first show that $R^{(1)}(f;\Omega) = 0$, and then show that the uniform Lipschitzness of the optimal representations would contradict with Proposition 3.

(1) Since the Jacobian takes two values inside $\mathbb{R}^3_+$, either the identity I_3 or $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$, we know by Theorem 1 that $R^{(1)}(f;\Omega) \geq 2\log|I_3|_+ = 0$. We therefore only need to construct a sequence of parameters of different depths that represent f with a squared parameter norm of order $3L + o(1)$. For simplicity, we only do this construction for even depths (the odd case can be constructed similarly). We define:

$$W_\ell = \begin{pmatrix} e^\epsilon & 0 & 0 \\ 0 & e^\epsilon & 0 \\ 0 & 0 & e^{-2\epsilon} \end{pmatrix} \text{ for } \ell = 1, \dots, \frac{L}{2} - 1$$

$$W_{\frac{L}{2}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ e^{-\frac{L-2}{2}\epsilon} & -e^{-\frac{L-2}{2}\epsilon} & 0 \end{pmatrix} \text{ for } \ell = \frac{L}{2} + 2, \dots, L$$

$$W_\ell = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & e^{-(L-2)\epsilon} \end{pmatrix} \text{ for } \ell = \frac{L}{2} + 2, \dots, L$$

$$W_\ell = \begin{pmatrix} e^{-\epsilon} & 0 & 0 \\ 0 & e^{-\epsilon} & 0 \\ 0 & 0 & e^{2\epsilon} \end{pmatrix} \text{ for } \ell = \frac{L}{2} + 2, \dots, L$$

We have for all $x \in \mathbb{R}^3_+$

$$\alpha_{\frac{L}{2}-1}(x) = \begin{pmatrix} e^{\frac{L-2}{2}\epsilon} x_1 \\ e^{\frac{L-2}{2}\epsilon} x_2 \\ e^{-(L-2)\epsilon} x_3 \end{pmatrix}$$

and

$$\alpha_{\frac{L}{2}}(x) = \begin{pmatrix} e^{\frac{L-2}{2}\epsilon} x_1 \\ e^{\frac{L-2}{2}\epsilon} x_2 \\ e^{-(L-2)\epsilon} x_3 \\ \sigma(x_1 - x_2) \end{pmatrix}$$

and

$$\alpha_{\frac{L}{2}+1}(x) = \begin{pmatrix} e^{\frac{L-2}{2}\epsilon} x_1 \\ e^{\frac{L-2}{2}\epsilon} x_2 \\ e^{-(L-2)\epsilon} \left(x_3 + \sigma(x_1 - x_2)\right) \end{pmatrix}$$

and

$$f_\theta(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_3 + \sigma(x_1 - x_2) \end{pmatrix}.$$

The norm of the parameters is

$$\begin{aligned} &\frac{L-2}{2}(2e^{2\epsilon} + e^{-4\epsilon}) + (3 + 2e^{-(L-2)\epsilon}) + (3 + e^{-2(L-2)\epsilon}) + \frac{L-2}{2}(2e^{-2\epsilon} + e^{4\epsilon}) \\ &= 3L + 2\left(e^{2\epsilon} - 1\right) + (e^{-4\epsilon} - 1) + 2e^{-(L-2)\epsilon} + e^{-2(L-2)\epsilon} + 2\left(e^{-2\epsilon} - 1\right) + (e^{4\epsilon} - 1) \end{aligned}$$

If we take $\epsilon = L^{-\gamma}$ for $\gamma \in (\frac{1}{2}, 1)$, then the terms $2e^{-(L-2)\epsilon}$ and $e^{-2(L-2)\epsilon}$ decay exponentially (at a rate of $e^{L^{1-\gamma}}$), in addition the terms $2\left(e^{2\epsilon} - 1\right) + (e^{-4\epsilon} - 1)$ and $2\left(e^{-2\epsilon} - 1\right) + (e^{4\epsilon} - 1)$ are of order $L^{-2\gamma}$. This proves that $R^{(1)}(f; \Omega) = 0$.

(2) Let us now assume that the optimal representation of f is C-uniform Lipschitz for some constant C, then by Proposition 3, we have that

$$R^{(1)}(f; \Omega) \geq \log |I_3|_+ + \log \left| \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \right|_+ + C^{-2} \left\| I_3 - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \right\|_* > 0,$$

which contradicts with the fact that $R^{(1)}(f; \Omega) = 0$. □

4.2 Extension outside FPLFs

Since all functions represented by finite depth and width networks are FPLFs, the representation cost of any such function is infinite. But we can define the representation cost of a function f that is the limit of a sequence of FPLF as the infimum over all sequences $f_i \to f$ converging of $\lim_{i\to\infty} R(f_i; \Omega)$ (for some choice of convergence type that implies convergence of the Jacobians $Jf_i(x) \to Jf(x)$). Note that since the representation cost $R(f; \Omega)$ is lower semi-continuous, i.e. $\liminf_{f\to f_0} R(f; \Omega) \geq R(f_0; \Omega)$, this does not change the definition of the representation cost on the space of FPLFs.

The definitions of the decomposition $R^{(0)}, R^{(1)}, R^{(2)}$ can also be similarly extended, and one can check that the properties described in Theorem 1 of [3] and Theorems 1 and [] of this paper all extend as well.

References

[1] Jillian Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

[2] L. Few. The shortest path and the shortest road through n points. *Mathematika*, 2(2):141–144, 1955.

[3] Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2023.