IMPLICIT BIAS OF LARGE DEPTH NETWORKS: A NOTION OF RANK FOR NONLINEAR FUNCTIONS

Anonymous authors
Paper under double-blind review

A NOTIONS OF RANK

Claim 1. From properties (2),(4) follows:

1. For any function f, one has $\mathrm{Rank} f \leq \min\{d_{in}, d_{out}\}$.
2. For any bijection ϕ on $\mathbb{R}^d$, $\mathrm{Rank}\phi = \mathrm{Rank}\phi^{-1} = d$.
3. For any two bijections ϕ, ψ on $\mathbb{R}^{d_{in}}$ and $\mathbb{R}^{d_{out}}$ resp. one has $\mathrm{Rank}\,(\psi \circ f \circ \phi) = \mathrm{Rank} f$.

Proof. 1. By property 4, one has that $\mathrm{Rank} id = d$ for the identity $id : \mathbb{R}^d \to \mathbb{R}^d$. By property (4), one has $\mathrm{Rank} f = \mathrm{Rank}(id \circ f \circ id) \leq \min\{d_{in}, \mathrm{Rank} f, d_{out}\} \leq \min\{d_{in}, d_{out}\}$.

2. We have $d = \mathrm{Rank}\left(\phi \circ \phi^{-1}\right) \leq \min\left\{\mathrm{Rank}\phi, \mathrm{Rank}\phi^{-1}\right\}$ and $\mathrm{Rank}\phi \leq d$ as well as $\mathrm{Rank}\phi^{-1} \leq d$. Therefore $\mathrm{Rank}\phi = \mathrm{Rank}\phi^{-1} = d$.

3. Let us only show $\mathrm{Rank}\,(f \circ \phi) = \mathrm{Rank} f$, the other side follows from the same argument. We have $\mathrm{Rank}\,(f \circ \phi) \leq \min\{\mathrm{Rank} f, d_{in}\} = \mathrm{Rank} f$ and $\mathrm{Rank} f = \mathrm{Rank}\left(f \circ \phi \circ \phi^{-1}\right) \leq \min\{\mathrm{Rank}\,(f \circ \phi), d_{in}\} = \mathrm{Rank}\,(f \circ \phi)$, thus proving $\mathrm{Rank}\,(f \circ \phi) = \mathrm{Rank} f$. □

Proposition 1 (Proposition 1 in the main)**.** *We have*

$$\mathrm{Rank}_J(f;\Omega) \leq \mathrm{Rank}_{BN}(f;\Omega).$$

Proof. Since $f = g \circ h$ with an inner dimension of $\mathrm{Rank}_{BN}(f;\Omega)$ then at any point x where f is differentiable, we have by the chain rule

$$Jf(x) = Jg(h(x))Jh(x).$$

Clearly the rank of $Jf(x)$ is bounded by the inner dimension $\mathrm{Rank}_{BN}(f;\Omega)$. □

Let us now give an example of a function f where the above inequality is strict:

Example 1. Consider the piecewise linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$ which maps $x = (x_0, x_1)$ to $(x_0, \mathrm{sign}(x_1)\,|x_0|)$ if $|x_0| \geq |x_1|$ and to $(\mathrm{sign}(x_0)\,|x_1|\,, x_1)$ if $|x_0| < |x_1|$.

Proof. One can easily check that this function is continuous and equals the identity on the x-cross $X = \{(x_0, x_1) : |x_0| = |x_1|\}$. Inside the linear regions (i.e. outside of the x-cross and the $+$-cross made up of the union of both axis) the Jacobian is rank 1, as a result the function f satisfies $\mathrm{Rank}_J(f;\mathbb{R}^2) = 1$, on the other hand $\mathrm{Rank}_{BN}(f;\mathbb{R}^2) > 1$ since $\mathrm{Rank}_{BN}(f;\mathbb{R}^2) \geq \mathrm{Rank}_{BN}\,(f;X)$ and since there are no continuous functions $g : X \to \mathbb{R}$ and $h : \mathbb{R} \to X$ such that $h \circ g = id_X$ we know that $\mathrm{Rank}_{BN}\,(f;X) > 1$. We therefore know that $\mathrm{Rank}_{BN}(f;\mathbb{R}^2) = 2$ (since $1 < \mathrm{Rank}_{BN}(f;\mathbb{R}^2) \leq 2$). □

Finally, one can easily check that both Rank_J and Rank_{BN} satisfy properties 1-4.

B REPRESENTATION COST

Proposition 2 (Proposition 3 in the main)**.** *Let f be a piecewise linear function, then at any differentiable point x, we have*

$$\|Jf(x)\|_{2/L}^{2/L} := \sum_{k=1}^{\mathrm{Rank}Jf_{\mathbf{W}}(x)} s_k\left(Jf(x)\right)^{\frac{2}{L}} \leq \frac{1}{L} R(f;\Omega,\sigma_a,L),$$

where $s_k\left(Jf_{\mathbf{W}}(x)\right)$ is the k-th singular value of the Jacobian $Jf_{\mathbf{W}}(x)$.

Proof. For any weights $\mathbf{W}$ of a depth L network such that $f_{\mathbf{W}} = f$ we have

$$Jf(x) = W_L D_{L-1}(x) W_{L-1} \cdots W_2 D_1(x) W_1$$

where $D_\ell(x)$ is a $n_\ell \times n_\ell$ diagonal matrix with diagonal vector equal to $\dot{\sigma_a}\left(\tilde{\alpha}_\ell(x)\right)$.

We know from (Soudry et al., 2018) that the representation cost of linear fully connected networks equals $L\left\|A\right\|_p^p$ for $\left\|A\right\|_p^p = \lambda_1^p + \cdots + \lambda_k^p$ is the L_p-Schatten norm with $p = \frac{2}{L}$. In other terms, we have for any matrices $\tilde{W}_1, \ldots, \tilde{W}_L$

$$L\left\|\tilde{W}_L \cdots \tilde{W}_1\right\|_p^p \leq \left\|\tilde{W}_L\right\|_F^2 + \cdots + \left\|\tilde{W}_1\right\|_F^2.$$

Applying it to $\tilde{W}_L = W_L$ and $\tilde{W}_\ell = D_\ell(x) W_\ell$ for $\ell = 1, \ldots, L-1$, we obtain

$$\begin{aligned}\|Jf(x)\|_p^p &\leq \frac{\|W_L\|_F^2 + \|D_{L-1}(x)W_{L-1}\|_F^2 + \cdots + \|D_1(x)W_1\|_F^2}{L} \\ &\leq \frac{\|W_L\|_F^2 + \|W_{L-1}\|_F^2 + \cdots + \|W_1\|_F^2}{L}\end{aligned}$$

since $\left\|D_\ell(x)\right\|_{op} \leq 1$.

Note that this result applies for any widths $n_1, \ldots, n_{L-1}$. □

Theorem 1 (first part of Theorem 1 in the main)**.** *We have*

$$\mathrm{Rank}_J(f;\Omega) \leq \lim_{L\to\infty} \frac{R(f;\Omega,\sigma_a,L)}{L} \leq \mathrm{Rank}_{BN}(f;\Omega).$$

Proof. **First inequality:** Take a point x such that $\mathrm{Rank}(Jf(x)) = \mathrm{Rank}_J(f;\Omega)$, then Proposition 2 implies that $\frac{R(f;\Omega,\sigma_a,L)}{L} \geq \|Jf(x)\|_p^p$. Letting $L \to \infty$ on both sides leads to the bound $\lim_{L\to\infty} \frac{R(f;\Omega,\sigma_a,L)}{L} \geq \mathrm{Rank}(Jf(x)) = \mathrm{Rank}_J(f;\Omega)$ as needed.

This lower bound applies to any widths $n_1(L), \ldots, n_{L-1}(L)$, of course if the widths are too small, it might be impossible to represent f, in which case $R(f;\Omega,\sigma_a,L) = \infty$.

Second Inequality: Fix a decomposition $f = g \circ h$ with minimal inner dimension and such that $h(\Omega) \subset \mathbb{R}_+^{\mathrm{Rank}(f;\Omega)}$ (we need to be in the upper quadrant to represent the identity on $h(\Omega)$ efficiently, and since Ω is bounded, one can always translate the output of h to be in the upper quadrant).

Corollary 1 tells us that there are two networks of finite depths L_h and L_g (with parameters $\mathbf{W}_h$ and $\mathbf{W}_g$) which represent h and g, for any depth L larger than $L_h + L_g$ we can construct a network of depth L which represents f by concatenating the network that the represents h, followed by $L - L_h - L_g$ identity weight matrices of dimension $\mathrm{Rank}(f;\Omega) \times \mathrm{Rank}(f;\Omega)$ and finally the network representing g. The norm of the parameters of this network is $\|\mathbf{W}_h\|^2 + (L - L_h - L_g)\mathrm{Rank}(f;\Omega) + \|\mathbf{W}_g\|^2$. We therefore have the bound

$$R(f;\Omega,\sigma_a,L) \leq \|\mathbf{W}_h\|^2 + (L - L_h - L_g)\mathrm{Rank}_{BN}(f;\Omega) + \|\mathbf{W}_g\|^2$$

divinding both sides by L and letting L grow to infinity, we obtain the inequality $\lim_{L\to\infty} \frac{R(f;\Omega,\sigma_a,L)}{L} \leq \mathrm{Rank}_{BN}(f;\Omega)$.

For the upper bound to apply, the widths n_ℓ of the network in the first part must be larger than some threshold that depends on the number of linear regions in h, in the middle part the widths must be larger than k and in the last part they must be above a threshold that depends on the number of linear regions in g (He et al., 2018). Note that in each of these regions the minimal with required does not depend on the depth. □

Let us now show that the limiting rescaled representation cost $R_\infty(f;\Omega,\sigma_a) := \lim_{L\to\infty} \frac{R(f;\Omega,\sigma_a,L)}{L}$ satisfies all properties of rank except the first one (though it might actually satisfy it):

Theorem 2 (second part of Theorem 1 in the main)**.** *We have for any piecewise linear functions f, g:*

1. $R_\infty(f\circ g;\Omega,\sigma_a) \leq \min\{R_\infty(f;g(\Omega),\sigma_a), R_\infty(g;\Omega,\sigma_a)\}$.
2. $R_\infty(f+g;\Omega,\sigma_a) \leq R_\infty(f;\Omega,\sigma_a) + R_\infty(g;\Omega,\sigma_a)$.
3. *If f is affine ($f(x)=Ax+b$) then* $R_\infty(f;\Omega,\sigma_a) = \mathrm{Rank}A$.

Proof. **1.** Without loss of generality, we can translate the output of g and the input of f (keeping the same composition $f\circ g$) so that $g(\Omega)$ lies in the upper quadrant $\mathbb{R}^m_+$ where m is the inner dimension. This translation changes the parameter norm by a value which is constant in L, it therefore does not matter in the $L\to\infty$ limit of $\|\mathbf{W}\|^2/L$.

Assume $R_\infty(f;\Omega,\sigma_a) \leq R_\infty(g;\Omega,\sigma_a)$ (the other case can be proved with the same argument) and fix a network of depth L_0 and parameters $\mathbf{W}_0$ that represents the function g. For any L sufficiently large we consider the network made up of the composition of the fixed network followed by a network of depth $L-L_0$ with weigths $\mathbf{W}'$ which represents f with minimal parameter norm, i.e. $\|\mathbf{W}'\|^2 = R(f;g(\Omega),\sigma_a,L-L_0)$. The norm of this composed network is $\|\mathbf{W}_0\|^2 + R(f;g(\Omega),\sigma_a,L-L_0)$, in the $L\to\infty$ limit, this implies $R_\infty(f\circ g;\Omega,\sigma_a) \leq R_\infty(f;g(\Omega),\sigma_a)$ as needed.

2. For any sufficiently large depth L consider two networks of depth L with parameters $\mathbf{W}_f$ and $\mathbf{W}_g$ which represent the functions f and g with minimal parameter norms, i.e. $\|\mathbf{W}_f\|^2 = R(f;\Omega,\sigma_a,L)$ and $\|\mathbf{W}_g\|^2 = R(g;\Omega,\sigma,L)$. We then consider the network obtained by putting the two network in 'parallel', i.e. the first weight matrix is given by the concatenation $\begin{pmatrix} W_{f,1} \\ W_{g,1}\end{pmatrix}$, the weight matrices of the middle layers are of the form $\begin{pmatrix} W_{f,\ell} & 0 \\ 0 & W_{g,\ell}\end{pmatrix}$ for all $\ell = 2,\dots,L-1$ and the last weight matrix is given by $\begin{pmatrix} W_{f,L} & W_{g,L}\end{pmatrix}$. This new network represents the function $f+g$ and has parameter norm $\|\mathbf{W}_f\|^2 + \|\mathbf{W}_g\|^2$, which implies the bound $R(f+g;\Omega,\sigma_a,L) \leq \|\mathbf{W}_f\|^2 + \|\mathbf{W}_g\|^2 = R(f;\Omega,\sigma_a,L) + R(g;\Omega,\sigma_a,L)$ and in the limit $R_\infty(f+g;\Omega,\sigma_a) \leq R_\infty(f;\Omega,\sigma_a) + R_\infty(g;\Omega,\sigma_a)$.

3. This point follows from Theorem 1, since for affine functions both notions of rank agree $\mathrm{Rank}_J(f;\Omega) = \mathrm{Rank}_{BN}(f;\Omega) = \mathrm{Rank}A$, the same must be true for the limiting rescaled representation cost $R_\infty(f;\Omega,\sigma_a)$.

Regarding the widths required for these results to apply, the minimal widths are the one described by the construction in the proofs. Since these constructions can be included into any wider network by adding zero neurons (neurons with zero incoming weights and zero outcoming weights), these results also apply to any larger widths. □

C Global Minima are Almost Rank 1

Consider a global minimizer $\hat{\mathbf{W}}$ of the regression problem $\mathcal{L}(\mathbf{W}) = \frac{1}{N}\sum (f_{\mathbf{W}}(x_i)-y_i)^2 + \frac{\lambda}{L}\|\mathbf{W}\|^2$ we will now show that if the depth L is large enough, then the function $f_{\hat{\mathbf{W}}}$ is in a sense almost rank 1 w.r.t. both notions of rank (Rank_J and Rank_{BN}).

Proposition 3 (Proposition 2 in the main)**.** *For the regression problem* $\mathcal{L}_\lambda(\mathbf{W}) = \frac{1}{N}\sum \left(f_{\mathbf{W}}(x_i) - y_i\right)^2 + \frac{\lambda}{L}\left\|\mathbf{W}\right\|^2$ *there is a constant* C_N *(which depends only on the inputs* x_i *and outputs* y_i*) such that for* $L \geq \lceil \log_2(n_0+1)\rceil + 2$*, we have*

$$\inf_{\mathbf{W}} \mathcal{L}_\lambda(\mathbf{W}) \leq \lambda\left(1 + \frac{C_N}{L}\right).$$

Proof. There is a BN-rank 1 function $f = h \circ g$ (with $g : \mathbb{R}^{n_0} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}^{n_L}$) which fits the data perfectly $f(x_i) = y_i$ for all i. Much like in the second inequality of Theorem 1, by Corollary 1, there is a depth $\lceil \log_2(n_0+1)\rceil$ networks which represents g and a depth 2 network which represents the function f. For any depth $L > \lceil \log_2(n_0+1)\rceil + 2$, we compose the network representing g, followed by a number of identity layers, followed by the network representing h. The function represented by this network has zero loss and the regularization term is of the form $\lambda\left(1 + \frac{C_N}{L}\right)$, since

$$\frac{1}{L}\left\|\mathbf{W}\right\|^2 = \frac{\left\|\mathbf{W}_g\right\|^2}{L} + \left(1 - \frac{\lceil \log_2(n_0+1)\rceil + 2}{L}\right) + \frac{\left\|\mathbf{W}_h\right\|^2}{L}.$$

Note that the minimal width required for this result to apply might depend on the number of datapoints N, but not the depth. □

Let us now show that the function $f_{\hat{\mathbf{W}}}$ is close to a BN-rank 1 function:

Proposition 4 (Proposition 4 in the main)**.** *For any global minimum* $\hat{\mathbf{W}}$ *of the* L_2*-regularized loss* $\mathcal{L}_\lambda$ *with* $\lambda > 0$ *and any set of* $\tilde{N}$ *datapoints* $\tilde{X} \in \mathbb{R}^{d_{in} \times \tilde{N}}$ *(which do not have to be the training set* X*) with non-constant outputs, there is a layer* ℓ_0 *such that the first two singular values* s_1, s_2 *of the hidden representation* $Z_{\ell_0} \in \mathbb{R}^{n_\ell \times N}$ *(whose columns are the activations* $\alpha_{\ell_0}(x_i)$ *for all the inputs* x_i *in* $\tilde{X}$*) satisfies* $\frac{s_2}{s_1} = O(L^{-\frac{1}{4}})$.

Proof. We need to prove a lower bound on the first eigenvalue of $\frac{1}{\tilde{N}} Z_\ell^T Z_\ell$ and an upper bound on the second one. For both parts, we will rely on the balanced property described in Proposition 7: at any local minimum of the loss the weights satisfy $\left\|W_{\ell+1}\right\|_F^2 = \left\|W_\ell\right\|_F^2 + \left\|b_\ell\right\|^2$. This implies that $\left\|W_{\ell'}\right\|_F^2 \geq \left\|W_\ell\right\|_F^2$ for all $\ell' \geq \ell$. Since the overall norm of the parameters is bounded by $L + C_N$ this implies a bound

$$\left\|W_{\ell+1}\right\|_F^2 \leq \frac{L + C_N}{L - \ell} = 1 + \frac{C_N + \ell}{L - \ell}$$

for all ℓ .

Assuming by contradiction that for all layers $\frac{\lambda_2\left(\frac{1}{\tilde{N}} Z_\ell^T Z_\ell\right)}{\lambda_1\left(\frac{1}{\tilde{N}} Z_\ell^T Z_\ell\right)} > \delta$, one should intuitively think of $1 + \frac{C_N+\ell}{L-\ell}$ as a 'ressource' with which the ℓ-th layer has to do two tasks: (1) keep the top eigenvalue of $\frac{1}{\tilde{N}} Z_\ell^T Z_\ell$ close to 1 to keep enough information to represent the outputs; and (2) keep the second eigenvalue above δ to keep the contradiction. However the ressource cost of (1) is roughly 1 and the cost of (2) is roughly δ which is above the ressource allowance $1 + \frac{C_N+\ell}{L-\ell}$ for large L and constant ℓ. This leads to a contradiction.

Upper bound on λ_2**:** Let ℓ_0 be the first time where $\lambda_2 < \delta$, we will show that ℓ_0 exists and is upper bounded by $\frac{2\left\|\frac{1}{\tilde{N}} \tilde{X}^T \tilde{X}\right\|_{op}}{\delta}$.

For all $\ell < \ell_0$ we have the following for bound the operator norm $\left\|\frac{1}{\tilde{N}} Z_\ell^T Z_\ell\right\|_{op}$:

$$\begin{aligned}
\left\|\frac{1}{\tilde{N}} Z_\ell^T Z_\ell\right\|_{op} &\leq \mathrm{Tr}\left[\frac{1}{\tilde{N}} Z_\ell^T Z_\ell\right] - \lambda_2\left(\frac{1}{\tilde{N}} Z_\ell^T Z_\ell\right) \\
&\leq \mathrm{Tr}\left[\frac{1}{\tilde{N}} \sigma_a\left(W_\ell Z_{\ell-1} + b_\ell\right)^T \sigma_a\left(W_\ell Z_{\ell-1} + b_\ell\right)\right] - \delta \\
&\leq \left\|\frac{1}{\tilde{N}} Z_{\ell-1}^T Z_{\ell-1}\right\|_{op} \left\|W_\ell\right\|_F^2 + \left\|b_\ell\right\|_F^2 - \delta.
\end{aligned}$$

Assuming $\left\|\frac{1}{\tilde{N}}Z_\ell^T Z_\ell\right\|_{op} \leq \max\left\{\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}, 1\right\}$ (we will later show that this is true for all $\ell \leq \ell_0$ as long as L is sufficiently large) we obtain:

$$\left\|\frac{1}{\tilde{N}}Z_\ell^T Z_\ell\right\|_{op} \leq \left\|\frac{1}{\tilde{N}}Z_{\ell-1}^T Z_{\ell-1}\right\|_{op} + \max\left\{\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}, 1\right\}\left(\|W_\ell\|_F^2 + \|b_\ell\|_F^2 - 1\right) - \delta. \quad (1)$$

Since $\|W_\ell\|_F^2 + \|b_\ell\|_F^2 = \|W_{\ell+1}\|_F^2 \leq 1 + \frac{\ell+C_N}{L-\ell}$, if $L > \frac{2\max\left\{\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op},1\right\}\left(C_N+\left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil\right)}{\delta} + \left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil \geq \frac{\kappa}{\delta^2}$ for some κ (which depends on $\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}$ and C_N only) and L large enough, then for all $\ell \leq \min\left\{\left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil, \ell_0\right\}$, we obtain that

$$\begin{aligned}\left\|\frac{1}{\tilde{N}}Z_\ell^T Z_\ell\right\|_{op} &\leq \left\|\frac{1}{\tilde{N}}Z_{\ell-1}^T Z_{\ell-1}\right\|_{op} - \frac{\delta}{2} \\ &\leq \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op} - \ell\frac{\delta}{2}.\end{aligned}$$

Therefore for all $\ell \leq \min\left\{\left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil, \ell_0\right\}$ we have $\left\|\frac{1}{\tilde{N}}Z_\ell^T Z_\ell\right\|_{op} \leq \max\left\{\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}, 1\right\}$, as needed. Furthermore this implies that $\ell_0 \leq \left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil$, otherwise, we would get a contradiction when taking $\ell = \left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil$:

$$\left\|\frac{1}{\tilde{N}}Z_\ell^T Z_\ell\right\|_{op} \leq \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op} - \left\lceil\frac{2\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta}\right\rceil\frac{\delta}{2} < 0.$$

We have now proven that for large enough L, there is a κ (which depends on $\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}$ and C_N only) such there is a $\ell_0 \leq \left\lceil\frac{2\sqrt{L}\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\delta\sqrt{\kappa}}\right\rceil$ where $\lambda_2\left(\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right) < \sqrt{\frac{\kappa}{L}}$.

Lower bound on λ_1: We now need to lower bound the first eigenvalue $\lambda_1\left(\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right) = \left\|\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right\|_{op}$ at this same layer ℓ_0 . We denote the means $m_\ell = \frac{1}{\tilde{N}}\sum_{i=1}^{\tilde{N}}\alpha_\ell(x_i)$ and have the bounds

$$\left\|\frac{1}{\tilde{N}}\left(Z_\ell - m_\ell\right)^T\left(Z_\ell - m_\ell\right)\right\|_{op} \leq \left\|\frac{1}{\tilde{N}}\left(Z_{\ell-1} - m_{\ell-1}\right)^T\left(Z_{\ell-1} - m_{\ell-1}\right)\right\|_{op}\|W_\ell\|_{op}^2.$$

This implies that

$$\left\|\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right\|_{op} \geq \left\|\frac{1}{\tilde{N}}\left(Z_\ell - m_{\ell_0}\right)^T\left(Z_\ell - m_{\ell_0}\right)\right\|_{op} \geq \frac{\left\|\frac{1}{\tilde{N}}\left(Z_\ell - m_{\ell_0}\right)^T\left(Z_\ell - m_{\ell_0}\right)\right\|_{op}}{\|W_{\ell_0+1}\|_{op}^2\cdots\|W_L\|_{op}^2}. \quad (2)$$

We now need to lower bound the norm of the parameters in the layers up to ℓ_0, to upper bound the norm of the parameters of the layers ℓ_0+1 to L. Iterating Equation (1) leads to the equation

$$\left\|\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right\|_{op} \leq \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op} + \max\left\{\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}, 1\right\}\left(\sum_{\ell=1}^{\ell_0}\|W_\ell\|_F^2 + \|b_\ell\|_F^2 - 1\right) - \ell_0\delta.$$

which implies that

$$\sum_{\ell=1}^{\ell_0} \|W_\ell\|_F^2 + \|b_\ell\|_F^2 \geq \ell_0 + \frac{\ell_0\delta - \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}}{\max\left\{\left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}, 1\right\}} \geq \ell_0 - \min\left\{1, \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}^{-1}\right\}$$

and therefore

$$\sum_{\ell=\ell_0+1}^{L} \|W_\ell\|_F^2 + \|b_\ell\|_F^2 \leq L - \ell_0 + C_N + \min\left\{1, \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}^{-1}\right\}.$$

Applying the arithmetic/geometric mean inequality to Equation (2), we obtain a lower bound

$$\begin{aligned}
\left\|\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right\|_{op} &\geq \frac{\left\|\frac{1}{\tilde{N}}\left(Z_\ell - m_{\ell_0}\right)^T\left(Z_\ell - m_{\ell_0}\right)\right\|_{op}}{\left(\frac{1}{L-\ell_0}\sum_{\ell=\ell_0+1}^{L}\|W_\ell\|_{op}^2\right)^{L-\ell_0}} \\
&\geq \frac{\left\|\frac{1}{\tilde{N}}\left(Z_\ell - m_{\ell_0}\right)^T\left(Z_\ell - m_{\ell_0}\right)\right\|_{op}}{\left(1 + \frac{C_N + \min\left\{1, \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}^{-1}\right\}}{L-\ell_0}\right)^{L-\ell_0}} \\
&\geq e^{-C_N - \min\left\{1, \left\|\frac{1}{\tilde{N}}\tilde{X}^T\tilde{X}\right\|_{op}^{-1}\right\}}\left\|\frac{1}{\tilde{N}}\left(Z_\ell - m_{\ell_0}\right)^T\left(Z_\ell - m_{\ell_0}\right)\right\|_{op}.
\end{aligned}$$

Putting it all together, we have shown that for large enough L, there is a $\ell_0 = O(\sqrt{L})$ such that $\lambda_1\left(\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right) = \Omega(1)$ and $\lambda_2\left(\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right) = O\left(\frac{1}{\sqrt{L}}\right)$, which together imply that

$$\frac{\lambda_1\left(\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right)}{\lambda_2\left(\frac{1}{\tilde{N}}Z_{\ell_0}^T Z_{\ell_0}\right)} = O\left(\frac{1}{\sqrt{L}}\right)$$

and therefore

$$\frac{s_1(Z_{\ell_0})}{s_2(Z_{\ell_0})} = O(L^{-\frac{1}{4}}).$$

Finally note that this result does not require anything more than the widths be nonzero. Of course , if one of the widths is 1, the result is trivial. □

D RANK RECOVERY

Consider a finite dataset X, Y of size N, with x_i sampled i.i.d. for a distribution p with support equal to Ω and with $y_i = f^*(x_i)$ for a true function $f^* : \Omega \to \mathbb{R}^{d_{out}}$ with $\mathrm{Rank}_J(f^*; \Omega) = k > 1$. For any function g which fits the data $g(x_i) = y_i$ with a BN-Rank of 1 (there always exists at least one such function), then if the depth L is large enough we have $R(g; \Omega, \sigma_a, L) < R(f^*; \Omega, \sigma_a, L)$.

This is problematic as it suggests that for large depths, minimizing the representation cost will always lead to fitting the data with a function with a BN rank of 1 instead of the rank of the true function k. However, if we instead fix a depth L and let the number of datapoints N grow, the representation cost required to fit the data with a rank 1 function (or any rank lower than k) increases to infinity, whereas the representation cost of the true function remains constant. This suggest that if one increases the depth L and the number of datapoints N simultaneously with the right scaling, minimizing the representation cost over fitting functions should recover a function h with the right rank k.

Proposition 5 (Theorem 2 in the main)**.** *Let f satisfy $f(x_i) = y_i$ and $\mathrm{Rank}_{BN}(f;\Omega) = 1$ for some Ω which contains the convex hull of $x_1, \dots, x_N$. There is a point $x \in \Omega$, such that*

$$\|Jf(x)\|_{op} \geq \frac{\mathrm{TSP}(y_1, \dots, y_N)}{\mathrm{diam}(x_1, \dots, x_N)},$$

for the Traveling Salesman Problem $\mathrm{TSP}(y_1, \dots, y_N)$, i.e. the length of the shortest path passing through every points $y_1, \dots, y_m$, and for the diameter $\mathrm{diam}(x_1, \dots, x_N)$ of the points $x_1, \dots, x_N$. As a result any rank 1 interpolator with parameters $\mathbf{W}$ satisfies $\|\mathbf{W}\|^2 \geq L \left(\frac{\mathrm{TSP}(y_1,\dots,y_N)}{\mathrm{diam}(x_1,\dots,x_N)}\right)^{\frac{2}{L}}$.

Proof. The lower bound on the nom of the paramaters $\|\mathbf{W}\|^2$ follows directly from the first bound and Proposition 2.

Let us now prove the first bound. Since $\mathrm{Rank}_{BN}(f;\Omega) = 1$, there are piecewise linear functions $g : \mathbb{R}^{d_{in}} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}^{d_{out}}$ such that $f = h \circ g$. We define $z_i = g(x_i)$ and w.l.o.g. we assume that $z_1 \leq \dots \leq z_N$.

The image of the segment $[x_1, x_N]$ under f is a path that connects y_1 to y_N, passing through the points $y_2, \dots, y_{N-1}$ (since the segment $[x_1, x_N]$ is mapped by g to a path from z_1 to z_N on the line, which must pass through $z_2, \dots, z_{N-1}$). This implies that the function f maps a path of length $\|x_1 - x_N\| \leq \mathrm{diam}(x_1, \dots, x_N)$ to a path of length at least $\mathrm{TSP}(y_1, \dots, y_N)$, as a result there must be a point x on the segment $[x_1, x_N]$ whose Jacobian has operator norm at least $\frac{\mathrm{TSM}(y_1,\dots,y_N)}{\mathrm{diam}(x_1,\dots,x_N)}$. □

Let us now prove that the global minima are approximately rank k in deep networks:

Proposition 6 (Proposition 5 in the main)**.** *Let the 'true function' $f^* : \Omega \to \mathbb{R}^{d_{out}}$ be piecewise linear with $\mathrm{Rank}_{BN}(f^*) = k$, then there is a constant C which depends on f^* only such that any global minimum $\hat{\mathbf{W}}$ of the loss $\mathcal{L}_\lambda(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^N \|f_{\mathbf{W}}(x_i) - f^*(x_i)\|^2 + \frac{\lambda}{L}\|\mathbf{W}\|^2$ for a sufficiently wide network satisfies*

$$\frac{R(f_{\hat{\mathbf{W}}};\Omega,\sigma_a,L)}{L} \leq k + \frac{C}{L}.$$

Proof. The true function f^* equals the composition of two piecewise linear functions $h \circ g$ with $g : \Omega \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}^{d_{out}}$ which can be represented by networks of depth $\lceil \log_2 d_{in} + 1 \rceil + 1$ (resp. $\lceil \log_2 k + 1 \rceil + 1$) and with parameters $\mathbf{W}_g$ (resp. $\mathbf{W}_h$) using Corollary 1. For $L > \lceil \log_2 d_{in} + 1 \rceil + \lceil \log_2 k + 1 \rceil + 2$, consider the network made up of the concatenation of the network representing g, and the network h at the end, with identity layers in the middle. This concatenated network has parameters norm

$$\|W_g\|^2 + k(L - \lceil \log_2 d_{in} + 1 \rceil - \lceil \log_2 k + 1 \rceil - 2) + \|W_h\|^2 .$$

Since this network recovers the true function, we have that for any global minimum $\hat{\mathbf{W}}$:

$$\begin{aligned}
\frac{R(f_{\hat{\mathbf{W}}};\Omega,\sigma_a,L)}{L} &\leq \frac{1}{L}\left\|\hat{\mathbf{W}}\right\|^2 \\
&\leq \frac{1}{\lambda}\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) \\
&\leq \frac{1}{L}\left(\|W_g\|^2 + k(L - \lceil \log_2 d_{in} + 1 \rceil - \lceil \log_2 k + 1 \rceil - 2) + \|W_h\|^2\right) \\
&= k + \frac{C}{L}.
\end{aligned}$$

The minimal widths required for this result only depends on the decomposition $f = h \circ g$ chosen, it does not depend on the depth. □

D.1 Rank of Kernel Ridge Regression

Consider a translation- and rotation-invariant kernel $K(x,y) = k(\|x-y\|)$ then the Kernel Ridge Regression (KRR) predictor with ridge parameter λ and on inputs X and outputs Y is of the form

$$\hat{f}_K(x) = K(x,X)\left(K(X,X) + \lambda I_N\right)^{-1} Y.$$

The Jacobian of $\hat{f}_K(x)$ equals $J\hat{f}_K(x) = JK(x,X)\left(K(X,X) + \lambda I_N\right)^{-1} Y$, where

$$JK(x,X) = (X-x)\mathrm{diag}\left(\frac{k'(\|x-X\|)}{\|x-X\|}\right),$$

where $X - x$ is the $n_0 \times N$ dimension matrix with entries $(X-x)_{ki} = X_{ki} - x_k$ and $\mathrm{diag}\left(\frac{k'(\|x-X\|)}{\|x-X\|}\right)$ is the $N \times N$ diagonal matrix with diagonal entries $\frac{k'(\|x-x_i\|)}{\|x-x_i\|}$. Since $\mathrm{diag}\left(\frac{k'(\|x-X\|)}{\|x-X\|}\right)$ is invertible, we have $\mathrm{Rank}\left(JK(x,X)\right) = \mathrm{Rank}(X-x)$. For almost all choices of x (i.e. as long as x does not belong to a zero Lebesgue measure set) one has $\mathrm{Rank}\left(JK(x,X)\right) = \mathrm{Rank}(X-x) = \min\{d_{in}, N, \mathrm{Rank}X + 1\}$.

Assuming that Y conditioned on X is sampled from a distribution with full support (as is the case when there is i.i.d. noise on the entries of Y for example), then $\mathrm{Rank}Y = \min\{d_{out}, N\}$ with prob. 1. As a result, the rank of $J\hat{f}_K(x)$ will be $\min\{\mathrm{Rank}(X-x), \mathrm{Rank}Y\} = \min\{\mathrm{Rank}X + 1, N, d_{out}\}$ with prob. 1.

Assuming N to be larger than the input and output dimensions and X to be full rank, we obtain that the Jacobian $J\hat{f}_K(x)$ is almost surely full rank

$$\mathrm{Rank}\left(J\hat{f}_K(x)\right) = \min\{d_{in}, d_{out}\}.$$

E Technical Results

E.1 Representation of Piecewise Linear Functions

Let us prove a generalization of the Theorem 2.1 from (Arora et al., 2018):

Corollary 1. *For any $L \geq \lceil \log_2(n_0+1) \rceil + 1$, and any piecewise linear function $f : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ there are widths $\mathbf{n}$ and parameters $\mathbf{W}$ such that $f_{\mathbf{W}} = f$.*

Proof. This result was proven in (Arora et al., 2018) for ReLU networks, we therefore simply need to show that given a ReLU network with widths $\mathbf{n}$ and parameters $\mathbf{W}$ such $f_{\mathbf{W}} = f$, there are widths $\mathbf{n}'$ and parameters $\mathbf{W}'$ such that $f_{\mathbf{W}'} = f$ for a network with a nonlinearity σ_a.

Notice that for any $a \in (-1,1)$ $\frac{\sigma(x) - a\sigma(-x)}{1-a^2} = \max\{0,x\}$. By doubling the number of neurons in each hidden layer, i.e. $n'_\ell = 2n_\ell$, we can represent the same output function f with the nonlinearity σ_a. □

E.2 Weak Balancedness Property

In the analysis of linear networks a widely used tool is the notion of balancedness, which is an invariant of linear networks during training. Furthermore any at any local minimum of the L_2-regularized loss the weights of the network must be balanced. While no direct equivalent of this notion exists for nonlinear DNNs, for homogeneous nonlinearities a weaker notion exists, which we describe now.

Proposition 7. *Let $\mathbf{W}$ be a local minimum of the L_2-regularized loss $\mathcal{L}_\lambda(\mathbf{W}) = C(f_{\mathbf{W}}) + \lambda \|\mathbf{W}\|^2$ for some $\lambda > 0$. Then $\mathbf{W}$ satisfies*

$$\|W_\ell\|_F^2 + \|b_\ell\|^2 = \|W_{\ell+1}\|_F^2 .$$

Proof. Given a local minimum $\mathbf{W}$ of the L_2-regularized loss, one can change the the weights of the network to new weights $\mathbf{W}(\alpha_1, \dots, \alpha_L)$ with the same outputs for any set of scalars $\alpha_1, \dots, \alpha_L$ such that $\alpha_1 \cdots \alpha_L = 1$:

$$W_\ell(\alpha_1, \dots, \alpha_L) \mapsto \alpha_\ell W_\ell$$
$$b_\ell(\alpha_1, \dots, \alpha_L) \mapsto \alpha_1 \cdots \alpha_\ell b_\ell$$

Since $\mathbf{W}$ is a local minimum, the derivatives of the norm of the parameters $\|\mathbf{W}(\alpha_1, \dots, \alpha_L)\|^2$ w.r.t. to $\alpha_1, \dots, \alpha_L$ at $\alpha_1 = \cdots = \alpha_L = 1$ have to be orthogonal to the constraint space $\alpha_1 \cdots \alpha_L = 1$. At $\alpha_1 = \cdots = \alpha_L = 1$the normal space (orthogonal to the tangent space) is the space of constant vectors, since it is spanned by the gradient of the product $\alpha_1 \cdots \alpha_L$. This implies that the values

$$\begin{aligned}\partial_{\alpha_\ell} \left(\|\mathbf{W}(\alpha_1, \dots, \alpha_L)\|^2 \right) (1, \dots, 1) &= \partial_{\alpha_\ell} \left(\sum_{\ell=1}^{L} \|\alpha_\ell W_\ell\|_F^2 + \|\alpha_1 \cdots \alpha_\ell b_\ell\|^2 \right) (1, \dots, 1) \\ &= 2 \|W_\ell\|_F^2 + 2 \|b_\ell\|^2 + \cdots + 2 \|b_L\|^2\end{aligned}$$

must all be equal. This equality for two consecutive layers implies that

$$\|W_\ell\|_F^2 + \|b_\ell\|^2 + \cdots + \|b_L\|^2 = \|W_{\ell+1}\|_F^2 + \|b_{\ell+1}\|^2 + \cdots + \|b_L\|^2$$

and therefore that at any local minimum

$$\|W_\ell\|^2 + \|b_\ell\|^2 = \|W_{\ell+1}\|^2 .$$

□

Remark 1. A yet stronger notion of balancedness can be obtained by observing that this rescaling of the weights can be done neuron by neuron instead of layer by layer, but we do not need this notion for our proofs.

F Experimental Setup

All our experiments were done on fully-connected ReLU networks with biases. We used diagonal networks, i.e. $n_1 = n_2 = \cdots = n_{L-1} = w$ for some width w. We trained the network using Adam with weight decay, in some cases we used traditional gradient descent at the end of training to make sure to converge as close as possible to a local minimum. When the ridge λ is small, we often observe two phases in learning: in the first phase, the cost $C(f_{\mathbf{W}})$ goes down very fast as the network fits the data, in the second part the cost remains close to zero and the parameter of the network slowly goes down. This second part is very slow and we did in most case stop before the parameter norm had completely stabilized. Note that even with this 'early stopping' we observed results consistent with our theory.

For Figure 1, the inputs $x \in \mathbb{R}^{50}$ and outputs $y \in \mathbb{R}^{50}$ were generated from a 15-dimensional latent representation $z \in \mathbb{R}^{15}$ sampled with i.i.d. $\mathcal{N}(0, 1)$ entries. The inputs x then equal $x = g(x)$ for a function $g : \mathbb{R}^{15} \to \mathbb{R}^{50}$ and the outputs equal $y = h(x_1, \dots, x_5)$ for a function $h : \mathbb{R}^5 \to \mathbb{R}^{50}$ which depends only on the first 5 coordinates of the latent space. Both functions g, h are represented by random shallow ReLU networks with inner width $n_1 = 100$.

For Figure 2 the data points from the 4 classes were using the same inverted S-shape distribution and translated on the x axis according to their class.

For Figure 3, we used the MNIST dataset on the left and on the right data of the form $g(z)$ for z random 1D Gaussian scalars and a function $g : \mathbb{R}^1 \to \mathbb{R}^2$ represented y a random ReLU network.

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=B1J_rgWRW`.

Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. Relu deep neural networks and linear finite elements. *arXiv preprint arXiv:1807.03973*, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.