Supplementary: Towards a Mathematical Understanding of the Difficulty in Learning with Feedforward Neural Networks

Hao Shen
fortiss - The Research Institute of the Free State of Bavaria, Germany
Guerickestr. 25, 80805 Munich, Germany
hao.shen@fortiss.org

1. Tracy-Singh product and Khatri-Rao product

Given two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$. Let us partition $\mathbf{A}$ into blocks $A_{ij} \in \mathbb{R}^{m_i \times n_j}$, and $\mathbf{B}$ into blocks $B_{kl} \in \mathbb{R}^{p_k \times q_l}$. The *Tracy-Singh product* of $\mathbf{A}$ and $\mathbf{B}$ [4] is defined as

$$\mathbf{A} \circledast \mathbf{B} = (A_{ij} \circledast \mathbf{B})_{ij} = ((A_{ij} \otimes B_{kl})_{kl})_{ij}, \qquad (1)$$

where the notion $(\cdot)_{ij}$ follows the convention of referring to the (i, j)-th block of a partitioned matrix. The matrix $\mathbf{A} \circledast \mathbf{B}$ is of the dimension $(mp) \times (nq)$, and its rank shares the same property as the *Kronecker product* of matrices as

$$\operatorname{rank}(\mathbf{A} \circledast \mathbf{B}) = \operatorname{rank}(\mathbf{A}) \operatorname{rank}(\mathbf{B}). \qquad (2)$$

If $\mathbf{A}$ and $\mathbf{B}$ are partitioned identically, then the *Khatri-Rao product* of the two matrices is defined as

$$\mathbf{A} \odot \mathbf{B} = (A_{ij} \otimes B_{ij})_{ij}. \qquad (3)$$

The matrix $\mathbf{A} \odot \mathbf{B}$ is of the dimension $(\sum_i m_i p_i) \times (\sum_i n_i q_i)$. The connection between the *Tracy-Singh product* and the *Khatri-Rao product* is given as

$$\mathbf{A} \odot \mathbf{B} = Z_1^\top (\mathbf{A} \circledast \mathbf{B}) Z_2, \qquad (4)$$

where $Z_1 \in \mathbb{R}^{(mp) \times (\sum_i m_i p_i)}$ and $Z_2 \in \mathbb{R}^{(nq) \times (\sum_i n_i q_i)}$ are two selection matrices, satisfying $Z_1^\top Z_1 = I_{\sum_i m_i p_i}$ and $Z_2^\top Z_2 = I_{\sum_i n_i q_i}$. We refer to [2] for concrete constructions of matrices Z_1 and Z_2, and more technical details regarding the *Khatri-Rao product*. It is then trivial to conclude the following corollary.

Corollary 1. *Given two identically partitioned matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, the rank of the* Tracy-Singh product *and the rank of the* Khatri-Rao product *of both matrices fulfils the following inequality*

$$\operatorname{rank}(\mathbf{A} \odot \mathbf{B}) \leq \operatorname{rank}(\mathbf{A} \circledast \mathbf{B}). \qquad (5)$$

Furthermore, it is clear that

$$\operatorname{rank}(Z_1) = \sum_i m_i p_i, \qquad (6)$$

and

$$\operatorname{rank}(Z_2) = \sum_i n_i q_i. \qquad (7)$$

Now, we recall the *Frobenius' rank inequality* [3], i.e., given three matrices A, B, C that have compatible dimensions, then

$$\operatorname{rank}(ABC) + \operatorname{rank}(B) \geq \operatorname{rank}(AB) + \operatorname{rank}(BC). \qquad (8)$$

A special case of the *Frobenius' rank inequality* is the so-called *Sylvester's rank inequality*, i.e., given two matrices $A \in \mathbb{R}^{m \times n}$ and $C \in \mathbb{R}^{n \times p}$, then the rank of the product of U and V is bounded by

$$\operatorname{rank}(AC) \geq \operatorname{rank}(A) + \operatorname{rank}(C) - n. \qquad (9)$$

Let $A \in \mathbb{R}^{m \times n_1}$, $B \in \mathbb{R}^{n_1 \times n_2}$, and $C \in \mathbb{R}^{n_2 \times p}$. By combining both the Frobenius' rank inequality and the Sylvester's rank inequality, we have

$$\begin{aligned} \operatorname{rank}(ABC) &\geq \operatorname{rank}(AB) + \operatorname{rank}(BC) - \operatorname{rank}(B) \\ &\geq \operatorname{rank}(A) + \operatorname{rank}(B) - n_1 - n_2 + \\ &\quad + \operatorname{rank}(B) + \operatorname{rank}(C) - \operatorname{rank}(B) \\ &= \operatorname{rank}(A) + \operatorname{rank}(B) + \operatorname{rank}(C) - \\ &\quad - n_1 - n_2. \end{aligned} \qquad (10)$$

If the *Tracy-Singh product* $\mathbf{A} \circledast \mathbf{B}$ has full rank, denoted by

$$R_{ts} := \operatorname{rank}(\mathbf{A} \circledast \mathbf{B}), \qquad (11)$$

then the rank of the *Khatri-Rao product* $\mathbf{A} \odot \mathbf{B}$ is bounded from below by

$$\begin{aligned} \operatorname{rank}(\mathbf{A} \odot \mathbf{B}) \geq & \sum_i m_i p_i + R_{ts} + \sum_j n_j q_j - \\ & - mp - nq. \end{aligned} \qquad (12)$$

Note, that the above lower bound is not guaranteed to be positive. Hence, nothing is conclusive about the rank of the *Khatri-Rao product* of two arbitrary full rank matrices.

2. Proof of Proposition 2, 3, 5, and 6

Proposition 2. *Given a collection of matrices* $\Psi_i \in \mathbb{R}^{n_l \times n_L}$ *and a collection of vectors* $\phi_i \in \mathbb{R}^{n_{l-1}}$*, for* $i = 1, \ldots, T$*, let* $\mathbf{\Psi} := [\Psi_1, \ldots \Psi_T] \in \mathbb{R}^{n_l \times (n_L T)}$ *and* $\mathbf{\Phi} = [\phi_1, \ldots, \phi_T] \in \mathbb{R}^{n_{l-1} \times T}$*. Then the rank of the Khatri-Rao product* $\mathbf{\Psi} \odot \mathbf{\Phi}$ *is bounded from below by*

$$\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq n_l \operatorname{rank}(\mathbf{\Phi}) + \sum_{i=1}^{T} \operatorname{rank}(\Psi_i) - T n_l. \quad (13)$$

If all matrices Ψ_i*'s and* $\mathbf{\Phi}$ *are of full rank, then the rank of* $\mathbf{\Psi} \odot \mathbf{\Phi}$ *has the following properties:*

(1) If $n_l \leq n_L$*, then* $\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq n_l \operatorname{rank}(\mathbf{\Phi})$*;*

(2) If $n_l > n_L$ *and* $n_{l-1} \geq T$*, then* $\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq T n_L$*;*

(3) If $n_l > n_L$ *and* $n_{l-1} < T$*, then* $\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq n_L$*.*

Proof. We can trivially rewrite the Kronecker product for each partition as

$$\begin{aligned} \Psi_i \otimes \phi_i &= (I_{n_l} \Psi_i) \otimes (\phi_i 1) \\ &= (I_{n_l} \otimes \phi_i) \Psi_i. \end{aligned} \quad (14)$$

Then, the *Khatri-Rao product* of $\mathbf{\Psi}$ and $\mathbf{\Phi}$ can be computed as the product of two matrices, i.e.,

$$\mathbf{\Psi} \odot \mathbf{\Phi} = \underbrace{[I_{n_l} \otimes \phi_1, \ldots, I_{n_l} \otimes \phi_T]}_{=:(I_{n_l} \circledast \mathbf{\Phi}) \in \mathbb{R}^{(n_l n_{n-1}) \times (n_l T)}} \underbrace{\operatorname{diag}(\Psi_1, \ldots, \Psi_T)}_{=:\widetilde{\mathbf{\Psi}} \in \mathbb{R}^{(n_l T) \times (n_L T)}}, \quad (15)$$

where $I_{n_l} \circledast \mathbf{\Phi}$ denotes the *Tracy-Singh product* of the identity matrix I_{n_l} and T column-wised partitioned matrix $\mathbf{\Phi}$, and the operator $\operatorname{diag}(\cdot)$ puts a sequence of matrices into a block diagonal matrix. By the rank property of the *Tracy-Singh product*, the rank of matrix $I_{n_l} \circledast \mathbf{\Phi}$ is equal to $n_l \operatorname{rank}(\mathbf{\Phi})$. Further, by the *Sylvester's rank inequality*, the rank of $\mathbf{\Psi} \odot \mathbf{\Phi}$ is bounded from below

$$\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq n_l \operatorname{rank}(\mathbf{\Phi}) + \sum_{i=1}^{T} \operatorname{rank}(\Psi_i) - T n_l. \quad (16)$$

Specifically, if all matrices Ψ_i's and $\mathbf{\Phi}$ are of full rank, we have the following properties.

(1) If $n_l \leq n_L$, then the rank of the block diagonal matrix $\widetilde{\mathbf{\Psi}}$ is equal to $n_l T$. By the Sylvester's rank inequality [3], we have

$$\begin{aligned} \operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) &\geq n_l \operatorname{rank}(\mathbf{\Phi}) + n_l T - n_l T \\ &= n_l \operatorname{rank}(\mathbf{\Phi}). \end{aligned} \quad (17)$$

(2) If $n_l > n_L$ and $n_{l-1} \geq T$, then the rank of $\widetilde{\mathbf{\Psi}}$ is equal to $n_L T$, and the rank of $(I_{n_l} \circledast \mathbf{\Phi})$ is equal to $n_l T$. By the Sylvester's rank inequality, we have

$$\begin{aligned} \operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) &\geq n_l T + n_L T - n_l T \\ &= n_L T. \end{aligned} \quad (18)$$

(3) If $n_l > n_L$ and $n_{l-1} < T$, then the rank of $(I_{n_l} \circledast \mathbf{\Phi})$ is equal to $n_l n_{l-1}$. By the same argument, we have

$$\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq n_l n_{l-1} + n_L T - n_l T. \quad (19)$$

It is clear that such a lower bound can be even negative, i.e., practically useless. However, since matrix $\mathbf{\Phi}$ is of full rank, there must exist a non-zero vector ϕ_i, so that $\operatorname{rank}(\Psi_i \otimes \phi_i) = n_L$. Then we have the result $\operatorname{rank}(\mathbf{\Psi} \odot \mathbf{\Phi}) \geq n_L$. □

Proposition 3. *For an MLP architecture* $\mathcal{F}$*, the rank of* $\mathbf{P}(\mathbf{W})$ *as defined in Eq. (22) (in the manuscript) is bounded from below by*

$$\begin{aligned} \operatorname{rank}\big(\mathbf{P}(\mathbf{W})\big) \geq & \sum_{l=1}^{L} n_l \operatorname{rank}\left(\mathbf{\Phi}_{l-1}\right) - \sum_{l=1}^{L-1} T n_l \\ & + \sum_{l=1}^{L} \sum_{i=1}^{T} \operatorname{rank}\left(\Psi_l^{(i)}\right) - L T n_L. \end{aligned} \quad (20)$$

Proof. By stacking all row blocks $\mathbf{\Psi}_l \odot \mathbf{\Phi}_{l-1}$ for $l = 1, \ldots, L$ together, we have $\mathbf{P}(\mathbf{W})$ as in Eq. (22) (in the manuscript). We can rewrite $\mathbf{P}(\mathbf{W})$ as

$$\begin{aligned} \mathbf{P}(\mathbf{W}) = & \operatorname{diag}\left(I_{n_1} \circledast \mathbf{\Phi}_0, \ldots, I_{n_L} \circledast \mathbf{\Phi}_{L-1}\right) \cdot \\ & \cdot \operatorname{diag}\left(\widetilde{\mathbf{\Psi}}_1, \ldots, \widetilde{\mathbf{\Psi}}_L\right) \cdot \mathbf{I}_{T n_l}^{L}, \end{aligned} \quad (21)$$

where $\mathbf{I}_{T n_l}^{L} := [I_{T n_L}, \ldots, I_{T n_L}]^\top \in \mathbb{R}^{L T n_L \times T n_L}$ is a matrix of stacking L copies of the identity matrix $I_{T n_L}$ on top of each other. Then, by applying Eq. (10), it is straightforward to get

$$\begin{aligned} \operatorname{rank}\big(\mathbf{P}(\mathbf{W})\big) \geq & \sum_{l=1}^{L} n_l \operatorname{rank}\left(\mathbf{\Phi}_{l-1}\right) - \sum_{l=1}^{L} T n_l \\ & + \sum_{l=1}^{L} \sum_{i=1}^{T} \operatorname{rank}\left(\Psi_l^{(i)}\right) - L T n_L \\ & + T n_L. \end{aligned} \quad (22)$$

The result follows directly. □

It is clear that such a bound in Proposition 3 is still very problem-dependent, and hard to control. Nevertheless, due to the special structure of $\mathbf{I}_{T n_l}^{L}$, the actual rank bound is given practically by the largest bound of each individual row block as characterised in Proposition 2, i.e.,

$$\operatorname{rank}\big(\mathbf{P}(\mathbf{W})\big) \geq \max_{1 \leq l \leq L} \operatorname{rank}\left(\mathbf{\Psi}_l \odot \mathbf{\Phi}_{l-1}\right). \quad (23)$$

Proposition 5. *Let an MLP architecture with one hidden layer satisfy Principle 1, 3, and 4. Then, for a learning task with T unique training samples, if the following two conditions are fulfilled:*

(1) There are T units in the hidden layer, i.e., $n_1 = T$,

(2) T unique samples produce a basis in the output space of the hidden layer for all $W_1 \in \mathbb{R}^{n_0 \times n_1}$,

then a finite exact approximator $\widehat{g}$ is realised at a global minimum $\mathbf{W}^* \in \boldsymbol{\mathcal{W}}$, i.e., $F(\mathbf{W}^*, \cdot) = \widehat{g}$, and the loss function $\mathcal{J}$ is free of suboptimal local minima.

Proof. We feed samples $X := [x_1, \ldots, x_T] \in \mathbb{R}^{n_0 \times T}$ through the MLP to generate the outputs in the hidden layer $\Phi_1 := [\phi_1^{(1)}, \ldots, \phi_1^{T}] \in \mathbb{R}^{T \times T}$, which is invertible due to Condition (2). It can be achieved by employing appropriate activation functions as suggested in [1], such as the *Sigmoid* and the $\tanh$. Then in the output layer, we have $\Phi_2 := [\phi_2^{(1)}, \ldots, \phi_2^{(T)}] = W_2^\top \Phi_1 \in \mathbb{R}^{n_2 \times T}$. Let us denote by $Y := [g^*(x_1), \ldots, g^*(x_T)] \in \mathbb{R}^{n_2 \times T}$ the desired outputs. Then, every pair $(W_1, (Y\Phi_1^{-1})^\top)$ is a global minimum of the total loss function.

We then compute the critical point conditions in the output layer as

$$\underbrace{\left[I_{n_2} \otimes \phi_1^{(1)} \ldots I_{n_2} \otimes \phi_1^{(T)} \right]}_{:= P_2 \in \mathbb{R}^{(Tn_2) \times (Tn_2)}} \begin{bmatrix} \nabla_E(\phi_2^{(1)}) \\ \vdots \\ \nabla_E(\phi_2^{(T)}) \end{bmatrix} = 0, \tag{24}$$

where P_2 is a square matrix. By following case (1) in Proposition 2, we get $\operatorname{rank}(P_2) = Tn_2$. The result simply follows. □

Note, that in Proposition 5, we do not consider the dummy units introduced by the scalar-valued bias $b_{l,k}$. Nevertheless, using similar arguments, the statements in Proposition 5 also hold true for the case with free variables $b_{l,k}$.

The following result is simply a special case of Proposition 3 to a two-layer MLP.

Proposition 6. *Let a two-layer MLP architecture $\boldsymbol{\mathcal{F}}(n_0, n_1, n_2)$ with dummy units and $n_2 \leq n_1 \leq T$ satisfy Principle 1, 3, and 4, and $\mathbf{1} := [1, \ldots, 1]^\top \in \mathbb{R}^T$. For a learning task with T unique samples $X \in \mathbb{R}^{n_0 \times T}$, we have*

(1) if $\operatorname{rank}([X^\top, \mathbf{1}]) = n_0$, then

$$\operatorname{rank}\big(\mathbf{P}(\mathbf{W})\big) \geq \max\left\{n_1 n_2, n_1(n_0 + n_2 - T)\right\}; \tag{25}$$

(2) if $\operatorname{rank}([X^\top, \mathbf{1}]) = n_0 + 1$, then

$$\operatorname{rank}\big(\mathbf{P}(\mathbf{W})\big) \geq \max\left\{n_1 n_2, n_1(n_0 + n_2 - T + 1)\right\}. \tag{26}$$

Proof. It is straightforward to have

$$\begin{aligned} \operatorname{rank}\big(\mathbf{P}(\mathbf{W})\big) &\geq n_1 \operatorname{rank}\left(\boldsymbol{\Phi}_0\right) + n_2 n_1 - T n_1 \\ &\quad + 2Tn_2 - 2Tn_2 \\ &= n_1\big(\operatorname{rank}\left(\boldsymbol{\Phi}_0\right) + n_2 - T\big). \end{aligned} \tag{27}$$

Proposition 2 implies

$$\operatorname{rank}(\boldsymbol{\Psi}_2 \odot \boldsymbol{\Phi}_1) \geq n_2 \operatorname{rank}(\boldsymbol{\Phi}_1), \tag{28}$$

and

$$\operatorname{rank}(\boldsymbol{\Psi}_1 \odot \boldsymbol{\Phi}_0) \geq n_2. \tag{29}$$

By the construction of $\operatorname{rank}(\boldsymbol{\Phi}_1) \geq n_1$, the result follows directly. □

References

[1] Y. Ito. Nonlinearity creates linear independence. *Advances in Computational Mathematics*, 5(1):189–203, 1996. 3

[2] S. Liu. Matrix results on the Khatri-Rao and Tracy-Singh products. *Linear Algebra and its Applications*, 289(1):267–277, 1999. 1

[3] D. A. Simovici. *Linear Algebra Tools for Data Mining*. World Scientific Publishing Company, 2012. 1, 2

[4] D. S. Tracy and R. P. Singh. A new matrix product and its applications in partitioned matrices. *Statistica Neerlandica*, 26:143–157, 1972. 1