APPENDIX

A SEQUENCE OF KRONECKER PRODUCTS

The Kronecker product between a sequence of factor tensors is given by

$$\left(\boldsymbol{\mathcal{A}}^{(1)} \otimes \cdots \otimes \boldsymbol{\mathcal{A}}^{(S)}\right)_{i_1 \cdots i_N} \triangleq \boldsymbol{\mathcal{A}}^{(1)}_{j_1^{(1)} \cdots j_N^{(1)}} \cdots \boldsymbol{\mathcal{A}}^{(S)}_{j_1^{(S)} \cdots j_N^{(S)}}, \tag{15}$$

where

$$j_n^{(k)} = \begin{cases} i_n - \sum_{t=1}^{k-2} j_n^{(t)} \prod_{l=t+1}^{S} a_n^{(l)} \bmod a_n^{(S)} & k = S, \\ \left\lfloor \frac{i_n - \sum_{t=1}^{k-1} j_n^{(t)} \prod_{l=t+1}^{S} a_n^{(l)}}{\prod_{l=k+1}^{S} a_n^{(l)}} \right\rfloor & \text{otherwise}, \end{cases} \tag{16}$$

and $\boldsymbol{\mathcal{A}}^{(k)} \in \mathbb{R}^{a_1^{(k)} \times \cdots \times a_N^{(k)}}$.

B ALTERNATIVE EXPANSION DIRECTIONS OF SEKRON

The proposed SeKron structure represents a given tensor $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{w_1 \times \cdots \times w_n}$ using a sequence of Kronecker products as follows:

$$\boldsymbol{\mathcal{W}} = \sum_{r_1=1}^{R_1} \boldsymbol{\mathcal{A}}^{(1)}_{r_1} \otimes \sum_{r_2=1}^{R_2} \boldsymbol{\mathcal{A}}^{(2)}_{r_1 r_2} \otimes \cdots \otimes \sum_{r_{S-1}=1}^{R_{S-1}} \boldsymbol{\mathcal{A}}^{(S-1)}_{r_1 \cdots r_{S-1}} \otimes \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1}}. \tag{4 revisited}$$

While this decomposition structure is obtained by recursively finding the Kronecker decomposition of the right-most tensor, many alternative sequential Kronecker structures can be obtained as illustrated in Figure 4. However, such alternative structures do not fall within our SeKron framework as they cannot make use of our convolution algorithm (Algorithm 2)

C THEOREM PROOFS

Theorem 1 (Tensor Decomposition using a Sequence of Kronecker Products)**.** *Any tensor $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{w_1 \times \cdots \times w_N}$ can be represented by a sequence of Kronecker products between $S \in \mathbb{N}$ factors:*

$$\boldsymbol{\mathcal{W}} = \sum_{r_1=1}^{R_1} \boldsymbol{\mathcal{A}}^{(1)}_{r_1} \otimes \sum_{r_2=1}^{R_2} \boldsymbol{\mathcal{A}}^{(2)}_{r_1 r_2} \otimes \cdots \otimes \sum_{r_{S-1}=1}^{R_{S-1}} \boldsymbol{\mathcal{A}}^{(S-1)}_{r_1 \cdots r_{S-1}} \otimes \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1}}, \tag{4}$$

where $R_i \in \mathbb{N}$ and $\boldsymbol{\mathcal{A}}^{(k)} \in \mathbb{R}^{R_1 \times \cdots \times R_k \times a_1^{(k)} \times \cdots \times a_N^{(k)}}$.

Proof. First, we define intermediate tensors

$$\boldsymbol{\mathcal{B}}^{(k)}_{r_1 \cdots r_k} \triangleq \sum_{r_{k+1}}^{R_{k+1}} \boldsymbol{\mathcal{A}}^{(k+1)}_{r_1 \cdots r_{k+1}} \otimes \sum_{r_{k+2}}^{R_{k+2}} \boldsymbol{\mathcal{A}}^{(k+2)}_{r_1 \cdots r_{k+2}} \otimes \cdots \otimes \sum_{r_{S-1}}^{R_{S-1}} \boldsymbol{\mathcal{A}}^{(S-1)}_{r_1 \cdots r_{S-1}} \otimes \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1}} \tag{5 revisited}$$

Then the reconstruction error can be written as

$$\left\| \boldsymbol{\mathcal{W}}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \boldsymbol{\mathcal{A}}^{(k)}_{r_1 \cdots r_k} \otimes \boldsymbol{\mathcal{B}}^{(k)}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 \tag{17}$$

where $\boldsymbol{\mathcal{W}}^{(1)}$ is the initial tensor being decomposed. As described in Section 3.2, using reshaping operations

$$\mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} = \textsc{Mat}(\textsc{Unfold}(\boldsymbol{\mathcal{W}}^{(k)}_{r_1 \cdots r_{k-1}}, \mathbf{d}_{\boldsymbol{\mathcal{B}}^{(k)}_{r_1 \cdots r_k}})), \tag{8 revisited}$$

$$\mathbf{a}^{(k)}_{r_1 \cdots r_k} = \textsc{Vec}(\textsc{Unfold}(\boldsymbol{\mathcal{A}}^{(k)}_{r_1 \cdots r_k}, \mathbf{d}_{\boldsymbol{\mathcal{I}}_{\boldsymbol{\mathcal{A}}^{(k)}_{r_1 \cdots r_k}}})), \quad \mathbf{b}^{(k)}_{r_1 \cdots r_k} = \textsc{Vec}(\boldsymbol{\mathcal{B}}^{(k)}_{r_1 \cdots r_k}), \tag{9 revisited}$$

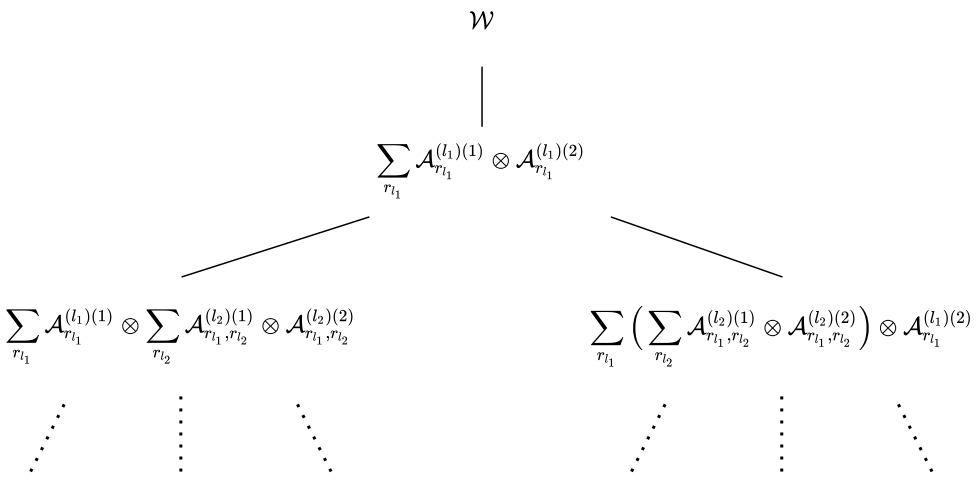

Figure 4: Illustration of alternative expansion directions using sequences of Kronecker products. SeKron structures are those which are leftmost on each level of the tree. Each node is obtained through the decomposition of a single tensor present in its parent node.

that preserve the sum of squares allows us to equivalently write the reconstruction error as

$$\left\| \mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} \mathbf{b}^{(k)\top}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2. \tag{18}$$

Now consider the singular value decomposition of matrix $\mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}}$ and let $\mathbf{u}^{(k)}_{r_1 \cdots r_k}, \mathbf{v}^{(k)}_{r_1 \cdots r_k}$ denote its left and right singular vectors, respectively (with the right singular vector scaled according to its corresponding singlar value). Set $\mathbf{a}^{(k)}_{r_1 \cdots r_k} = \mathbf{u}^{(k)}_{r_k}$ and define and define error terms

$$\delta^{(k)}_{r_1 \cdots r_k} = \mathbf{v}^{(k)}_{r_1 \cdots r_k} - \mathbf{b}^{(k)}_{r_1 \cdots r_k}, \quad \epsilon^{(k)}_{r_1 \cdots r_k} = \|\delta^{(k)}_{r_1 \cdots r_k}\|. \tag{19}$$

Expanding out equation 18 reveals its recursive form

$$\left\| \mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} \mathbf{b}^{(k)\top}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 = \left\| \mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} (\mathbf{v}^{(k)}_{r_k} - \delta^{(k)}_{r_1 \cdots r_k})^\top \right\|_{\mathrm{F}}^2 \tag{20}$$

$$= \left\| \mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} \mathbf{v}^{(k)\top}_{r_1 \cdots r_k} + \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} \delta^{(k)\top}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 \tag{21}$$

$$\leq \left\| \mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} \mathbf{v}^{(k)\top}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 + \sum_{r_k=1}^{\widehat{R}_k} \left\| \mathbf{a}^{(k)}_{r_1 \cdots r_k} \delta^{(k)\top}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 \tag{22}$$

$$\leq \left\| \mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}} - \sum_{r_k=1}^{\widehat{R}_k} \mathbf{a}^{(k)}_{r_1 \cdots r_k} \mathbf{v}^{(k)\top}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 + \sum_{r_k=1}^{\widehat{R}_k} d^{(k)} \epsilon^{(k)}_{r_1 \cdots r_k} \tag{23}$$

$$= \left(\sum_{r_k=\widehat{R}_k+1}^{R_k} \sigma^2_{r_k}(\mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}}) \right) + \left(\sum_{r_k=1}^{\widehat{R}_k} d^{(k)} \epsilon^{(k)}_{r_1 \cdots r_k} \right) \tag{24}$$

$$= \sum_{r_k=\widehat{R}_k+1}^{R_k} \sigma^2_{r_k}(\mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}}) + \sum_{r_k=1}^{\widehat{R}_k} d^{(k)} \left\| \mathbf{v}^{(k)}_{r_1 \cdots r_k} - \mathbf{b}^{(k)}_{r_1 \cdots r_k} \right\|_{\mathrm{F}}^2 \tag{25}$$

where $d^{(k)} \in \mathbb{N}$ is the number of dimensions of vector $\mathbf{a}^{(k)}_{r_1 \cdots r_k}$ and R_k is the rank of matrix $\mathbf{W}^{(k)}_{r_1 \cdots r_{k-1}}$, $\sigma_{r_k}(\boldsymbol{\mathcal{W}}^{(k)}_{r_1 \cdots r_{k-1}})$ denotes the r_k^{th} singular value of tensor $\boldsymbol{\mathcal{W}}^{(k)}_{r_1 \cdots r_{k-1}}$. By reshaping

vectors $\mathbf{v}^{(k)}_{r_1\cdots r_k}, \mathbf{b}^{(k)}_{r_1\cdots r_k}$ to matrices according to

$$\mathbf{V}^{(k)}_{r_1\cdots r_k} = \textsc{Mat}\left(\textsc{Unfold}\left(\textsc{Vec}^{-1}\left(\mathbf{v}^{(k)}_{r_1\cdots r_k}, \prod_{s=k+1}^{S}\mathbf{d}^{(s)}\right), \prod_{s=k+2}^{S}\mathbf{d}^{(s)}\right)\right), \tag{26}$$

$$\mathbf{B}^{(k)}_{r_1\cdots r_k} = \textsc{Mat}\left(\textsc{Unfold}\left(\textsc{Vec}^{-1}\left(\mathbf{b}^{(k)}_{r_1\cdots r_k}, \prod_{s=k+1}^{S}\mathbf{d}^{(s)}\right), \prod_{s=k+2}^{S}\mathbf{d}^{(s)}\right)\right), \tag{27}$$

where $\mathbf{d}^{(s)} = (a^{(s)}_1, \ldots, a^{(s)}_N)$ describes the dimensions of the s^{th} factor, we can re-write equation 25 as

$$\sum_{r_k=\widehat{R}_k+1}^{R_k}\sigma^2_{r_k}(\mathbf{W}^{(k)}_{r_1\cdots r_{k-1}}) + \sum_{r_k=1}^{\widehat{R}_k} d^{(k)}\left\|\mathbf{v}^{(k)}_{r_1\cdots r_k} - \mathbf{b}^{(k)}_{r_1\cdots r_k}\right\|^2_{\mathrm{F}} \tag{28}$$

$$= \sum_{r_k=\widehat{R}_k+1}^{R_k}\sigma^2_{r_k}(\mathbf{W}^{(k)}_{r_1\cdots r_{k-1}}) + \sum_{r_k=1}^{\widehat{R}_k} d^{(k)}\left\|\mathbf{V}^{(k)}_{r_1\cdots r_k} - \mathbf{B}^{(k)}_{r_1\cdots r_k}\right\|^2_{\mathrm{F}} \tag{29}$$

$$= \sum_{r_k=\widehat{R}_k+1}^{R_k}\sigma^2_{r_k}(\mathbf{W}^{(k)}_{r_1\cdots r_{k-1}}) + \sum_{r_k=1}^{\widehat{R}_k} d^{(k)}\left\|\mathbf{V}^{(k)}_{r_1\cdots r_k} - \sum_{r_{k+1}=1}^{\widehat{R}_{k+1}}\mathbf{a}^{(k+1)}_{r_1\cdots r_{k+1}}\mathbf{b}^{(k+1)\top}_{r_1\cdots r_{k+1}}\right\|^2_{\mathrm{F}}. \tag{30}$$

The last line reveals the recursive nature of the formula (compare with equation 20). Unrolling the recursive formula for $k = 1, \ldots, S-1$, by setting $\mathbf{W}^{(k+1)}_{r_1\cdots r_k} \leftarrow \mathbf{V}^{(k)}_{r_1\cdots r_k}$, leads to the following formula for the reconstruction error:

$$\begin{aligned}\varepsilon_{\text{SeKron}}(\mathbf{W}, \mathbf{r}, \mathbf{D}) = &\sum_{r_1=\widehat{R}_1+1}^{R_1}\sigma^2_{r_1}(\mathbf{W}^{(1)}) + d^{(1)}\sum_{r_1=1}^{\widehat{R}_1}\sum_{r_2=\widehat{R}_2+1}^{R_2}\sigma^2_{r_2}(\mathbf{W}^{(2)}_{r_1}) + \cdots \\ &+ d^{(1)}d^{(2)}\cdots d^{(S-2)}\sum_{r_1,r_2,\ldots,r_{S-2}=1}^{\widehat{R}_1,\cdots,\widehat{R}_{S-2}}\sum_{r_{S-1}=\widehat{R}_{S-1}+1}^{R_{S-1}}\sigma^2_{r_{S-1}}(\mathbf{W}^{(S-1)}_{r_1\cdots r_{S-2}})\end{aligned} \tag{31}$$

where $\mathbf{r} = (\widehat{R}_1, \ldots, \widehat{R}_{S-1})$ contains the rank values, $\mathbf{D}_s = \mathbf{d}^{(s)}$ contains the Kronecker factor shapes and is referred to as the $\mathbf{Dr}$-SeKron approximation error (note that the dependency of intermediate matrices $\mathbf{W}^{(k)}_{r_1\cdots r_{k-1}}$ on Kronecker factor shapes $\mathbf{D}$ is implied). Selecting $\widehat{R}_i = R_i\ \forall i$ in equation 31 results in zero reconstruction error. □

Theorem 2. *The factorization structure imposed by CP, Tucker, TT and TR when decomposing a given tensor $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{w_1\times\cdots\times w_N}$ can be achieved using SeKron.*

Proof. The SeKron decomposition of tensor $\boldsymbol{\mathcal{W}}$ is given by

$$\boldsymbol{\mathcal{W}}_{i_1\cdots i_N} = \sum_{r_1,\ldots,r_S=1}^{R_1,\cdots,R_S}\boldsymbol{\mathcal{A}}^{(1)}_{r_1 j^{(1)}_1\cdots j^{(1)}_N}\cdots\boldsymbol{\mathcal{A}}^{(S)}_{r_1\cdots r_{S-1}j^{(S)}_1\cdots j^{(S)}_N} \tag{32}$$

where $\boldsymbol{\mathcal{A}}^{(k)} \in \mathbb{R}^{R_1\times\cdots\times R_k\times a^{(k)}_1\times\cdots a^{(k)}_N}$ and

$$j^{(k)}_n = \begin{cases} i_n - \sum_{t=1}^{k-2} j^{(t)}_n \prod_{l=t+1}^{S} a^{(l)}_n \bmod a^{(S)}_n & k = S, \\ \left\lfloor \frac{i_n - \sum_{t=1}^{k-1} j^{(t)}_n \prod_{l=t+1}^{S} a^{(l)}_n}{\prod_{l=k+1}^{S} a^{(l)}_n} \right\rfloor & \text{otherwise,} \end{cases} \tag{16 revisited}$$

The CP decomposition of tensor $\boldsymbol{\mathcal{W}}$ in scalar form is

$$\boldsymbol{\mathcal{W}}_{i_1\cdots i_N} = \sum_{r=1}^{R^{(\text{CP})}}\boldsymbol{\mathcal{A}}^{(\text{CP}_1)}_{ri_1}\cdots\boldsymbol{\mathcal{A}}^{(\text{CP}_N)}_{ri_N} \tag{33}$$

where $\boldsymbol{\mathcal{A}}^{(\mathrm{CP}_k)} \in \mathbb{R}^{R^{(\mathrm{CP})} \times w_k}$. Configuring the SeKron decomposition in equation 32 such that $S = N;\ R_1 = R^{(\mathrm{CP})};\ R_2, \ldots, R_N = 1$ and $a_n^{(n)} = w_n$ for $n = 1, \ldots, N$ leads to the equivalent form

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} = \sum_{r_1=1}^{R^{(\mathrm{CP})}} \boldsymbol{\mathcal{A}}^{(1)}_{r_1 i_1 1 \cdots 1} \cdots \boldsymbol{\mathcal{A}}^{(N)}_{r, 1 \cdots 1 i_N}. \tag{34}$$

The Tucker decomposition of tensor $\boldsymbol{\mathcal{W}}$ is given by

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} = \sum_{r_1=1,\ldots,r_N}^{R_1^{(\mathrm{T})},\ldots R_N^{(\mathrm{T})}} \boldsymbol{\mathcal{G}}_{r_1 \cdots r_N} \boldsymbol{\mathcal{A}}^{(\mathrm{T}_1)}_{i_1 r_1} \cdots \boldsymbol{\mathcal{A}}^{(\mathrm{T}_N)}_{i_N r_N} \tag{35}$$

where $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{R_1^{(\mathrm{T})} \times \cdots \times R_N^{(\mathrm{T})}}$ and $\boldsymbol{\mathcal{A}}^{(\mathrm{T}_k)} \in \mathbb{R}^{w_k \times R_k^{(\mathrm{T})}}$. The SeKron decomposition of tensor $\boldsymbol{\mathcal{W}}$, with $S = N + 1,\ R_n = R_n^{(T)}$ and $a_n^{(n)} = w_n$ for $n = 1, \ldots, N$ yields

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} = \sum_{r_1,\ldots,r_N=1}^{R_1^{(\mathrm{T})},\cdots,R_N^{(\mathrm{T})}} \boldsymbol{\mathcal{A}}^{(1)}_{r_1 i_1 1 \cdots 1} \cdots \boldsymbol{\mathcal{A}}^{(N)}_{r_1 \cdots r_N 1 \cdots 1 i_N} \boldsymbol{\mathcal{A}}^{(N+1)}_{r_1 \cdots r_N 1 \cdots 1}, \tag{36}$$

which is equivalent to equation 35 in the special case where there are nullity constraints on some elements in the Kronecker factors, such that for $k = 2, \ldots, N$

$$\boldsymbol{\mathcal{A}}^{(k)}_{r_1 \cdots r_k 1 \cdots 1 i_k 1 \cdots 1} = 0 \quad \text{when} \quad r_j \in \{x \in \mathbb{N} \mid x \leq R_j^{(\mathrm{T})},\ x \neq R_j^{(\mathrm{T}^*)}\} \quad j = 1, \ldots, k-1 \tag{37}$$

for any choice of $R_j^{(\mathrm{T}*)} \in \{x \in \mathbb{N} \mid x \leq R_j^{(\mathrm{T})}\}$. The Tensor Ring (TR) decomposition of $\boldsymbol{\mathcal{W}}$ is given by

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} = \sum_{r_1=1,\ldots,r_N}^{R_1^{(\mathrm{TR})},\ldots R_N^{(\mathrm{TR})}} \boldsymbol{\mathcal{A}}^{(\mathrm{TR}_1)}_{i_1 r_1 r_2} \cdots \boldsymbol{\mathcal{A}}^{(\mathrm{TR}_N)}_{i_N r_N r_{N+1}} \tag{38}$$

where $\boldsymbol{\mathcal{A}}^{(\mathrm{TR}_k)} \in \mathbb{R}^{w_k \times R_k^{(\mathrm{TR})} \times R_{k+1}^{(\mathrm{TR})}}$, and $R_1^{(\mathrm{TR})} = R_{N+1}^{(\mathrm{TR})}$. As the Tensor Train decomposition can be viewed as a special case of the Tensor Ring decomposition (with $R_1^{(\mathrm{TR})} = R_{N+1}^{(\mathrm{TR})} = 1$), it suffices to show that SeKron generalizes Tensor Ring. The SeKron decomposition of tensor $\boldsymbol{\mathcal{W}}$, with $S = N+1$; $R_k = R_k^{(\mathrm{TR})}$ for $k = 1, \ldots, N-1$ and $a_n^{(n+1)} = w_n$ for $n = 1, \ldots, N$ leads to

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} = \sum_{r_1,\ldots,r_N=1}^{R_1^{(\mathrm{TR})},\ldots,R_N^{(\mathrm{TR})}} \boldsymbol{\mathcal{A}}^{(1)}_{r_1 1 \cdots 1} \boldsymbol{\mathcal{A}}^{(2)}_{r_1 r_2 i_1 1 \cdots 1} \cdots \boldsymbol{\mathcal{A}}^{(N+1)}_{r_1 \cdots r_{N+1} 1 \cdots 1 i_N}, \tag{39}$$

which is equivalent to equation 38 in the special case where some elements in the Kronecker factors are constrained, such that all elements in tensor $\boldsymbol{\mathcal{A}}^{(1)}$ are constrained to one and

$$\boldsymbol{\mathcal{A}}^{(k)}_{r_1 \cdots r_k 1 \cdots 1 i_k 1 \cdots 1} = 0 \quad \forall r_j \in \{x \in \mathbb{N} \mid x \leq R_j^{(\mathrm{TR})},\ x \neq R_j^{(\mathrm{TR}^*)}\} \tag{40}$$

for

$$j = \begin{cases} 1, \ldots, k-2 & k = 2, \ldots, N \\ 2, \ldots, k-1 & k = N+1 \end{cases} \tag{41}$$

for any choice of $R_j^{(\mathrm{TR}^*)} \in \{x \in \mathbb{N} \mid x \leq R_j^{(\mathrm{TR})}\}$. □

Theorem 3 (Linear Mappings with Sequences of Kronecker Products)**.** *Any linear mapping using a given tensor $\boldsymbol{\mathcal{W}}$ can be written directly in terms of its Kronecker factors $\boldsymbol{\mathcal{A}}^{(k)} \in \mathbb{R}^{R_1 \times \cdots R_N \times a_1^{(k)} \times \cdots \times a_N^{(k)}}$. That is:*

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} \boldsymbol{\mathcal{X}}_{i_1+z_1, \cdots, i_N+z_N} = \sum_{r_1,\ldots r_k}^{R_1,\ldots R_k} \boldsymbol{\mathcal{A}}^{(1)}_{r_1 j_1^{(1)} \cdots j_N^{(1)}} \cdots \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1} j_1^{(S)} \cdots j_N^{(S)}} \boldsymbol{\mathcal{X}}_{f(\mathbf{j}_1)+z_1, \cdots, f(\mathbf{j}_N)+z_N}$$

where $j_n^{(k)} \in \mathbb{N}$ is a function of input indices (see Appendix A) and $f(\mathbf{j}_n) = \sum_{k=1}^{S} j_n^{(k)} \prod_{l=k+1}^{S} a_n^{(l)}$

Proof. First we bring out the summations in the SeKron representaion of $\boldsymbol{\mathcal{W}}$

$$\boldsymbol{\mathcal{W}} = \sum_{r_1}^{R_1} \boldsymbol{\mathcal{A}}^{(1)}_{r_1} \otimes \sum_{r_2}^{R_2} \boldsymbol{\mathcal{A}}^{(2)}_{r_1 r_2} \otimes \cdots \otimes \sum_{r_{S-1}}^{R_{S-1}} \boldsymbol{\mathcal{A}}^{(S-1)}_{r_1 \cdots r_{S-1}} \otimes \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1}}, \tag{4 revisted}$$

such that

$$\boldsymbol{\mathcal{W}} = \sum_{r_1,\ldots,r_S=1}^{R_1,\cdots,R_{S-1}} \boldsymbol{\mathcal{A}}^{(1)}_{r_1} \otimes \cdots \otimes \boldsymbol{\mathcal{A}}^{(S)}_{r_1 r_2 \cdots r_{S-1}}. \tag{42}$$

Then, using the scalar form definition of sequences of kronecker products in equation 16

$$j_n^{(k)} = \begin{cases} i_n - \sum_{t=1}^{k-2} j_n^{(t)} \prod_{l=t+1}^{S} a_n^{(l)} \bmod a_n^{(S)} & k = S, \\ \left\lfloor \frac{i_n - \sum_{t=1}^{k-1} j_n^{(t)} \prod_{l=t+1}^{S} a_n^{(l)}}{\prod_{l=k+1}^{S} a_n^{(l)}} \right\rfloor & \text{otherwise}, \end{cases} \tag{16 revisited}$$

allows us to re-write equation 42 in scalar form as

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} = \sum_{r_1 \cdots r_S=1}^{R_1} \boldsymbol{\mathcal{A}}^{(1)}_{r_1 j_1^{(1)} \cdots j_N^{(1)}} \cdots \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1} j_1^{(S)} \cdots j_N^{(S)}} \tag{43}$$

As the $j_n^{(k)}$ terms decompose i_n into an integer weighted sum, we can recover i_n using

$$i_n = f(\mathbf{j}_n) \triangleq \sum_{k=1}^{S} j_n^{(k)} \prod_{l=k+1}^{S} a_n^{(l)}, \tag{44}$$

where $\mathbf{j}_n = (j_n^{(1)}, \ldots, j_n^{(S)})$. Thus, we can write

$$\boldsymbol{\mathcal{X}}_{i_1+z_1, \cdots i_N+z_N} = \boldsymbol{\mathcal{X}}_{f(\mathbf{j}_1)+z_1, \cdots f(\mathbf{j}_N)+z_N}. \tag{45}$$

Finally, combining equations equation 43 and equation 45 leads to

$$\boldsymbol{\mathcal{W}}_{i_1 \cdots i_N} \boldsymbol{\mathcal{X}}_{i_1+z_1, \cdots, i_N+z_N} = \sum_{r_1, \ldots r_N}^{R_1, \ldots R_k} \boldsymbol{\mathcal{A}}^{(1)}_{r_1 j_1^{(1)} \cdots j_N^{(1)}} \cdots \boldsymbol{\mathcal{A}}^{(S)}_{r_1 \cdots r_{S-1} j_1^{(S)} \cdots j_N^{(S)}} \boldsymbol{\mathcal{X}}_{f(\mathbf{j}_1)+z_1, \cdots, f(\mathbf{j}_N)+z_N}$$

□

Theorem 4. *(Universal approximation via shallow SeKron networks) Any shallow SeKron factorized neural network $\hat{f}^{(s)}$ with an L-Lipschitz activation function a, is dense in the class of continuous functions $C(X)$ for any compact subset X of $\mathbb{R}^d$*

Proof. Let $\hat{f}$ denote a shallow neural network, and $f \in C(X)$. Then,

$$\left\| f - \hat{f}^{(s)} \right\|_2^2 \triangleq \int_X \left(f(x) - \hat{f}^{(s)}(x) \right)^2 d\mu \tag{46}$$

$$= \int_X \left(f(x) - \hat{f}(x) \right)^2 d\mu \tag{47}$$

$$+ \int_X \left(\hat{f}(x) - \hat{f}^{(s)}(x) \right)^2 d\mu \tag{48}$$

$$+ 2 \int_X \left(f(x) - \hat{f}(x) \right) \left(\hat{f}(x) - \hat{f}^{(s)}(x) \right) d\mu \tag{49}$$

According to Hornik (1991), equation 47 is dense in $C(X)$; therefore, it suffices to show that equation 48 is bounded as well.

$$\int_X \left(\hat{f}(x) - \hat{f}^{(s)}(x) \right)^2 d\mu = \int_X \left(\mathbf{w}^\top \mathbf{a}(\mathbf{W}\mathbf{x}) - \mathbf{w}^\top \mathbf{a}(\mathbf{W}^{(s)}\mathbf{x}) \right)^2 d\mu \tag{50}$$

$$\leq L \left\| \mathbf{w} \right\|_2^2 \left\| X \right\|_2^2 \varepsilon_{\text{SeKron}}(\mathbf{W}, \mathbf{r}, \mathbf{D}) \tag{51}$$

where ε denotes the $\mathbf{Dr}$-SeKron approximation error as in equation 31, with matrix $\mathbf{D}$ and vector $\mathbf{r}$ describing the shapes of the Kronecker factors the ranks used in the SeKron decomposition of $\mathbf{W}$, respectively. □

D Implementation Details

In all of our experiments we use 4 NVIDIA Tesla V100 SXM2 32 GB GPUs during training and evaluate run time on a single core of Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz.

D.1 ImageNet Experiments

We train all models using stochastic gradient descent for 100 epochs using a batch size of 256. The learning rate is initially set to 0.1 and reduced by a factor of $10\times$ at epochs number 30, 60 and 90. We also use a 0.0001 weight decay.

D.2 CIFAR-10 Experiments

We train all models using using stochastic gradient descent for 200 epochs using a batch size of 128. The learning rate is initially set to 0.1 and is reduced by a factor of $5\times$ at epochs number 60, 120 and 160. We use nestrov momentum set to 0.9 and weight decay set to 0.0005.

D.3 DIV2K

We train all models using using the ADAM optimizer for 300 epochs using a batch size of 16. The optimizer's learning rate is set to 0.0001 and β_1, β_2 are set to $0.9, 0.999$ respectively.