Apppendix

A. Tensors and Higher Order Singular Value Decomposition

As a higher order analogue of Singular Value Decomposition (SVD) for matrices, Higher Order Singular Value Decomposition (HOSVD) for tensors is the main tool to develop our theorems. In this section, we give a quick review of tensors, and introduce the essential part of HOSVD which helps better understand the theorems. The contents here are based on (De Lathauwer et al., 2000) and (Kolda & Bader, 2009).

Basic Notations Below is a table for different notations.

Type	Notation	Examples
Tensor	Boldface Euler script letter	$\boldsymbol{\mathcal{A}}$
Matrix	Boldface capital letter	$\boldsymbol{A}$
Vectors	Boldface lowercase letters	$\boldsymbol{a}$
Scalars	Lowercase letters	a

The *order* of a tensor is the number of its dimensions, which is also called modes. We use subscripts on corresponding notations to denote a specific slice of a tensor. For example, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{4\times 2\times 4\times 5}$ is an order 4 tensor, and a_{ijkl} denotes the (i,j,k,l)-element of $\boldsymbol{\mathcal{A}}$. The vector coordinated along the first dimension is denoted as $\boldsymbol{a}_{:jkl}$, and the matrix coordinated along the first and the second dimension is denoted as $\boldsymbol{A}_{::kl}$.

The norm of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_K}$ is defined by

$$\|\boldsymbol{\mathcal{A}}\| = \sqrt{\sum_{i_1=1}^{I_1}\cdots\sum_{i_K=1}^{I_K} a_{i_1\cdots i_K}^2}, \tag{30}$$

Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times\cdots I_K}$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{J_1\times\cdots\times J_{K'}}$ be two tensors, then the outer product of $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$ is a tensor in $\mathbb{R}^{I_1\times\cdots\times I_K\times J_1\times\cdots\times J_{K'}}$, which is defined as

$$(\boldsymbol{\mathcal{A}}\otimes\boldsymbol{\mathcal{B}})_{i_1\ldots i_K j_1\ldots j_{K'}} = a_{i_1\ldots i_K} b_{j_1\ldots j_{K'}}. \tag{31}$$

Tensor reshaping The tensor reshaping refers that one can reshape a tensor into a matrix, or a matrix into a tensor based on some specific rules. The process of reshaping a tensor into a matrix is called tensor flattening, while reshaping a matrix into a tensor is called tensorisation. Tensor reshaping is the core idea of HOSVD.

Definition A.1. *Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_K}$ be an order K tensor. The mode-k flattening of $\boldsymbol{\mathcal{A}}$ is denoted as $\boldsymbol{A}_{(k)} \in \mathbb{R}^{I_k\times(I_1\cdots I_{k-1}I_{k+1}\cdots I_K)}$, where the $(i_1,i_2,\cdots,i_K)$-element of $\boldsymbol{\mathcal{A}}$ is mapped to the (i_k,j)-element of the matrix $\boldsymbol{A}_{(k)}$ with*

$$j = 1 + \sum_{\substack{s=1\\ s\neq k}}^{K}\left((i_s-1)\prod_{\substack{s'=1\\ s'\neq k}}^{s-1} I_{s'}\right). \tag{32}$$

That is, the columns of $\boldsymbol{A}_{(k)}$ are actually the vectors $\boldsymbol{a}_{i_1\cdots i_{k-1}:\, i_{k+1}\cdots i_K}$.

Definition A.2. *Let $\boldsymbol{A} \in \mathbb{R}^{I_K\times(I_1I_2\cdots I_{K-1})}$ be a matrix. Then $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_K}$ is the tensorisation of $\boldsymbol{A}$ if its mode-K flattening equals to $\boldsymbol{A}$.*

We illustrate the above two definitions by an example.

Example A.1. *Consider an order 3 tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{4\times3\times2}$ such that*

$$\boldsymbol{A}_{::1} = \begin{pmatrix} 1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\\ 10 & 11 & 12\end{pmatrix}, \quad \boldsymbol{A}_{::2} = \begin{pmatrix} 13 & 14 & 15\\ 16 & 17 & 18\\ 19 & 20 & 21\\ 22 & 23 & 24\end{pmatrix}. \tag{33}$$

Then

$$\begin{aligned}
\boldsymbol{A}_{(1)} &= \begin{pmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{pmatrix}, \\
\boldsymbol{A}_{(2)} &= \begin{pmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{pmatrix}, \\
\boldsymbol{A}_{(3)} &= \begin{pmatrix} 1 & 4 & 7 & 10 & \cdots & 3 & 6 & 9 & 12 \\ 13 & 16 & 19 & 22 & \cdots & 15 & 18 & 21 & 24 \end{pmatrix}.
\end{aligned} \tag{34}$$

Conversely, $\boldsymbol{\mathcal{A}}$ is the tensorisation of $\boldsymbol{A}_{(K)}$ in $\mathbb{R}^{4\times3\times2}$.

Recall the tensorisation $T_{l^K}(\boldsymbol{a})$ we defined in the paper, where $\boldsymbol{a} \in \mathbb{R}^{l^K}$ is a vector. In this case, we first rearrange the vector $\boldsymbol{a} \in \mathbb{R}^{l^K}$ into a matrix $\boldsymbol{A} \in \mathbb{R}^{l\times l^{K-1}}$ according to row major ordering, then $T_{l^K}(\boldsymbol{a})$ is the tensor in $\mathbb{R}^{l\times l\cdots\times l}$ defined as the tensorisation of $\boldsymbol{A}$.

Singular values and the rank of tensors

Definition A.3. *The singular values of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times\cdots\times I_K}$ are defined as*

$$\left\{\sigma_{i_k}^{(k)}\right\}, \quad k = 1, \ldots, K, \tag{35}$$

where $\left\{\sigma_{i_k}^{(k)}\right\}$ are the singular values of $\boldsymbol{A}_{(k)}$ arising from the SVD for matrices. That is, the singular values of $\boldsymbol{\mathcal{A}}$ is a collection of all the singular values of its mode-k flattening.

Remark A.1. *The norm of $\boldsymbol{\mathcal{A}}$ satisfies*

$$\|\boldsymbol{\mathcal{A}}\| = \|\boldsymbol{A}_{(k)}\|_F = \sqrt{\sum_{i_k=1}^{r_k}\left(\sigma_{i_k}^{(k)}\right)^2}, \qquad k = 1, 2, \ldots, K, \tag{36}$$

where r_k is the matrix rank of $\boldsymbol{A}_{(k)}$.

There are various ways to define the rank of a tensor. Here we use the following definition.

Definition A.4. *The rank of a tensor $\boldsymbol{\mathcal{A}}$ is defined as*

$$\operatorname{rank}\boldsymbol{\mathcal{A}} := \sum_{k=1}^{K} r_k. \tag{37}$$

Here r_k is the matrix rank of $\boldsymbol{A}_{(k)}$, which is also called the k-rank of $\boldsymbol{\mathcal{A}}$. Recall that the matrix rank equals to the number of its non-zero singular values, it follows from Definition A.3 that the tensor rank also equals to the number of its non-zero singular values.

Remark A.2. *Recall that the matrix SVD can be written in the form of outer product: $\boldsymbol{A} = \sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top$ with $r = \operatorname{rank}\boldsymbol{A}$.*

This can be generalised to HOSVD. For any $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_K}$, we have

$$\boldsymbol{\mathcal{A}} = \sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}\cdots\sum_{i_K=1}^{r_K} s_{i_1i_2\ldots i_K}\boldsymbol{u}_{i_1}^{(1)}\otimes\boldsymbol{u}_{i_2}^{(2)}\otimes\cdots\otimes\boldsymbol{u}_{i_K}^{(K)}, \tag{38}$$

where r_k is the k-rank of $\boldsymbol{\mathcal{A}}$, $s_{i_1i_2\ldots i_K} \in \mathbb{R}$ and $\boldsymbol{u}_{i_k}^{(k)} \in \mathbb{R}^{I_k}$.

We end up with the following approximation property of HOSVD, which can be viewed as an analogue of Eckart-Young-Mirsky theorem for matrix SVD.

Proposition A.5. *Let $\boldsymbol{\mathcal{A}}, \hat{\boldsymbol{\mathcal{A}}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ with the k-ranks denoted as $r_1, \ldots, r_K$ and $r'_1, \ldots, r'_K$ respectively. Let $\sigma_1^{(k)} \geq \sigma_2^{(k)} \geq \cdots \geq \sigma_{r_k}^{(k)} \geq 0$ be the singular values of $\boldsymbol{A}_{(k)}$, we have*

$$\inf_{\hat{\boldsymbol{\mathcal{A}}}} \|\boldsymbol{\mathcal{A}} - \hat{\boldsymbol{\mathcal{A}}}\|^2 \leq \sum_{i_1=r'_1+1}^{r_1} \left(\sigma_{i_1}^{(1)}\right)^2 + \sum_{i_2=r'_2+1}^{r_2} \left(\sigma_{i_2}^{(2)}\right)^2 + \cdots + \sum_{i_K=r'_K+1}^{r_K} \left(\sigma_{i_K}^{(K)}\right)^2, \tag{39}$$

where the infimum is taken over $\hat{\boldsymbol{\mathcal{A}}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ such that $r'_k \leq r_k$, $k = 1, \ldots K$.

Remark A.3. *Note that for matrix SVD, Eckart-Young-Mirsky theorem shows that the approximation error equals to the tail sum of singular values, but for HOSVD only the upper bound holds.*

B. Proofs

Now we get down to prove all the results shown in the text.

Recall that the input space $\mathcal{X}$ is a Hilbert space, one can apply the standard representation theorem.

Theorem B.1. (Riesz Representation Theorem) *For any continuous linear functional H defined on $\mathcal{X}$, there exists a unique $\rho \in \mathcal{X}$ such that*

$$H(\boldsymbol{x}) = \sum_{s=-\infty}^{\infty} \rho(s)^\top x(s), \tag{40}$$

and

$$\|H\| = \|\rho\|_{\mathcal{X}}. \tag{41}$$

Proof. See Bramwell & Kreyszig (1979), Theorem 3.8-1. □

Based on this, we prove Lemma 1.

Proof of Lemma 1. By Riesz Representation Theorem, for any $t \in \mathbb{Z}$ and $H_t \in \boldsymbol{H}$, there exists a unique $\boldsymbol{\rho}_t \in \mathcal{X}$ such that

$$H_t(\boldsymbol{x}) = \sum_{s=-\infty}^{\infty} \rho_t(s)^\top x(s). \tag{42}$$

With the fact that $\boldsymbol{H}$ is causal, we have

$$H_t(\boldsymbol{x}) = \sum_{s=-\infty}^{t} \rho_t(s)^\top x(s). \tag{43}$$

By the time homogeneity $H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}^{(\tau)})$ with $\tau = -t$, we get

$$H_t(\boldsymbol{x}) = \sum_{s=-\infty}^{t} \rho_t(s)^\top x(s) = \sum_{s=-\infty}^{0} \rho_0(s)^\top x(s+t). \tag{44}$$

The conclusion follows by taking $\rho^{(\boldsymbol{H})}(s) = \rho_0(-s)$. □

Recall the example in section 4.2, where we showed that for any matrix $\boldsymbol{A}$ with the rank no more than 2, there exists $\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})} \in \mathcal{H}_{\text{CNN}}^{(2,2,\{M_k\})}$ such that $T_{2^2}(\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}) = \boldsymbol{A}$. Now we extend this to the general l and K.

Proposition B.2. *For any $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K \in \mathbb{R}^l$, we have*

$$T_{l^K}\left(\boldsymbol{w}_K \underset{l^{K-1}}{\circledast} \boldsymbol{w}_{K-1} \underset{l^{K-2}}{\circledast} \cdots \underset{l}{\circledast} \boldsymbol{w}_1\right) = \boldsymbol{w}_K \otimes \boldsymbol{w}_{K-1} \otimes \cdots \otimes \boldsymbol{w}_1. \tag{45}$$

Proof. We prove this by induction. When $K = 2$, it is the usual outer product for vectors. Suppose the conclusion holds for K, we prove for $K+1$. Let $\boldsymbol{f}_K := \boldsymbol{w}_K *_{l^{K-1}} \boldsymbol{w}_{K-1} *_{l^{K-2}} \cdots *_l \boldsymbol{w}_1 = (c_1, c_2, \ldots, c_{l^K})$ and $\boldsymbol{w}_{K+1} = (w_1, w_2, \ldots, w_l)$, then

$$\begin{aligned}(\boldsymbol{w}_{K+1} \underset{l^K}{*} \boldsymbol{f}_K)(t) &= \sum_{s\in\mathbb{N}} w_{K+1}(s) f_K(t - l^K s) \\ &= w(0) f_K(t) + w(1) f_K(t - l^K) + \cdots + w(l-1) f_K(t - (l-1) l^K) \\ &= w_1 f_K(t) + w_2 f_K(t - l^K) + \cdots + w_l f_K(t - (l-1) l^K). \end{aligned} \tag{46}$$

The right hand side is non-zero only when $t = 0, 1, 2, \ldots, l^{K+1} - 1$, which gives

$$(\boldsymbol{w}_{K+1} \underset{l^K}{*} \boldsymbol{f}_K) = (w_1 c_1, w_1 c_2, \ldots, w_1 c_{l^K}, w_2 c_1, w_2 c_2, \ldots, w_2 c_{l^K}, \ldots, w_l c_1, w_l c_2 \ldots, w_l c_{l^K}) \in \mathbb{R}^{l^{K+1}}. \tag{47}$$

By the tensorisation for vectors discussed above, $\boldsymbol{w}_{K+1} *_{l^K} \boldsymbol{f}_K$ can be rearranged into a matrix according to row major ordering:

$$\begin{pmatrix} w_1 c_1 & w_1 c_2 & \cdots & w_1 c_{l^K} \\ w_2 c_1 & w_2 c_2 & \cdots & w_2 c_{l^K} \\ \vdots & \vdots & \ddots & \vdots \\ w_l c_1 & w_l c_2 & \cdots & w_l c_{l^K} \end{pmatrix} \in \mathbb{R}^{l \times l^K}. \tag{48}$$

This is in fact the mode-K flattening of $\boldsymbol{w}_{K+1} \otimes T_{l^K}(\boldsymbol{f}_K)$. By the induction hypothesis, we conclude that

$$T_{l^{K+1}}\Big(\boldsymbol{w}_{K+1} \underset{l^K}{*} \boldsymbol{f}_K\Big) = \boldsymbol{w}_{K+1} \otimes T_{l^K}(\boldsymbol{f}_K) = \boldsymbol{w}_{K+1} \otimes \boldsymbol{w}_K \otimes \boldsymbol{w}_{K-1} \otimes \cdots \otimes \boldsymbol{w}_1. \tag{49}$$

□

Proposition B.3. *Let $\mathcal{A} \in \mathbb{R}^{l \times l \times \cdots \times l}$ be an order K tensor with the k-rank r_k, $k = 1, \cdots, K$. There exists $\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})} \in \mathcal{H}^{(l)}_{CNN}$ such that $T_{l^K}(\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}) = \mathcal{A}$.*

Proof. Recall the linear CNN model

$$\begin{aligned} \boldsymbol{h}_{0,i} &= \boldsymbol{x}_i, \\ \boldsymbol{h}_{k+1,i} &= \sum_{j=1}^{M_k} \boldsymbol{w}_{kji} \underset{l^k}{*} \boldsymbol{h}_{k,j}, \\ \hat{\boldsymbol{y}} &= \boldsymbol{h}_{K+1}. \end{aligned} \tag{50}$$

By linearity, there exists $\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})} \in \mathcal{H}^{(l,K,\{M_k\})}_{\text{CNN}}$ such that

$$\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})} = \sum_{i_K=1}^{M_K} \sum_{i_{K-1}=1}^{M_{K-1}} \cdots \sum_{i_1=1}^{M_1} \boldsymbol{w}_{K,i_K} \underset{l^K}{*} \boldsymbol{w}_{K-1,i_{K-1}} \underset{l^{K-1}}{*} \cdots \underset{l}{*} \boldsymbol{w}_{1,i_1}, \tag{51}$$

where $\boldsymbol{w}_{k,i_k} \in \{\boldsymbol{w}_{kij}\}$ is a filter at layer k. According to Proposition B.2, we have

$$T_{l^K}(\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}) = \sum_{i_K=1}^{M_K} \sum_{i_{K-1}=1}^{M_{K-1}} \cdots \sum_{i_1=1}^{M_1} \boldsymbol{w}_{K,i_K} \otimes \boldsymbol{w}_{K-1,i_{K-1}} \otimes \cdots \otimes \boldsymbol{w}_{1,i_1}. \tag{52}$$

The conclusion now follows from Remark A.2 for sufficient large M_k. □

Now we can prove the first main theorem in the main text.

Proof of Theorem 2. By Lemma 1 and (12) in the main text, we have

$$\begin{aligned}\|\boldsymbol{H}-\hat{\boldsymbol{H}}\|^2=\sup_{t\in\mathbb{Z}}\|H_t-\widehat{H}_t\|^2&=\sup_{t\in\mathbb{Z}}\sup_{\|\boldsymbol{x}\|_{\mathcal{X}}\le 1}\left|\sum_{s\in\mathbb{N}}\left(\boldsymbol{\rho}^{(\boldsymbol{H})}(s)-\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}(s)\right)^{\top}x(t-s)\right|^2\\&\le\sum_{s\in\mathbb{N}}\left|\boldsymbol{\rho}^{(\boldsymbol{H})}(s)-\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}(s)\right|^2\\&=\sum_{s=0}^{l^K-1}\left|\boldsymbol{\rho}^{(\boldsymbol{H})}(s)-\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}(s)\right|^2+\sum_{s=l^K}^{\infty}\left|\boldsymbol{\rho}^{(\boldsymbol{H})}(s)\right|^2\\&=\left\|T_{l^K}(\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})})-T_{l^K}(\boldsymbol{\rho}^{(\boldsymbol{H})})\right\|^2+\sum_{s=l^K}^{\infty}\left|\boldsymbol{\rho}^{(\boldsymbol{H})}(s)\right|^2.\end{aligned}\tag{53}$$

Since $\boldsymbol{\rho}^{(\boldsymbol{H})}\in\ell^2$, we can choose K appropriately large such that the second term is less than ϵ. According to Proposition B.3, there exists $\mathcal{H}^{(l)}_{\text{CNN}}$ such that the first term is zero. The proof is completed. □

Lemma 3 follows from Proposition A.5 and the definition of rank for tensors.

This was denoted by $\|\cdot\|_{l,g}$ in the main text. Here we amend the notation as we do not discuss the norm aspects.

Proposition B.4. *Let $\boldsymbol{\rho}$ be a finitely supported sequence such that $r(\boldsymbol{\rho})\le l^K-1$. Denote the singular values of $T_{l^K}(\boldsymbol{\rho})$ by $\sigma_1\ge\sigma_2\ge\cdots\ge\sigma_{lK}$. For any sequence $\hat{\boldsymbol{\rho}}\in\mathbb{R}^{l^{K+1}}$ with $\hat{\boldsymbol{\rho}}_{[0,l^K-1]}=\boldsymbol{\rho}$ and $\hat{\boldsymbol{\rho}}_{[l^K,l^{K+1}]}=0$, the singular values of $T_{l^{K+1}}(\hat{\boldsymbol{\rho}})$ are*

$$\|T_{l^K}(\boldsymbol{\rho})\|=\sigma_{l(K+1)}\ge\sigma_1\ge\sigma_2\ge\cdots\ge\sigma_{lK}\ge 0=\sigma_{lK+1}=\sigma_{lK+1}=\cdots=\sigma_{lK+l-1}.\tag{54}$$

Proof. The singular values of $T_{l^{K+1}}(\hat{\boldsymbol{\rho}})$ arising from mode-1 to mode-K flattening are not changed, since there are only additions of zero columns. Now we consider the mode-$(K+1)$ flattening. We add zeros to an additional dimension of the tensor such that $\hat{a}_{i_1i_2\ldots i_K1}=a_{i_1i_2\ldots i_K}$ and $\hat{a}_{i_1i_2\ldots i_Kj}=0$ for $2\le j\le l$. Then the columns of $T_{l^{K+1}}(\hat{\boldsymbol{\rho}})_{(K+1)}$ are the vectors along the new dimension $(\hat{a}_{i_1i_2\ldots i_K1},0,0,\ldots,0)=(a_{i_1i_2\ldots i_K},0,0,\ldots,0)$, which gives a unique non-zero singular value $\sigma_{l(K+1)}=\|T_{l^K}(\boldsymbol{\rho})\|$, and $(l-1)$ zero singular values. □

Proposition B.5. *Suppose the function g is monotonously decreasing and strictly positive. Then for any $\boldsymbol{H}\in\mathcal{C}$ such that $\boldsymbol{\rho}^{(\boldsymbol{H})}$ is finitely supported, we have $\boldsymbol{H}\in\mathcal{C}^{(l,g)}$.*

Proof. Let $K'=\inf\{K\in\mathbb{N}_+:l^K\ge r(\boldsymbol{\rho}^{(\boldsymbol{H})})\}$. Then for any $s\ge lK'$, $\sum_{i=s+K'+k}^{l(K'+k)}|\sigma_i^{(K'+k)}|^2=0$ for all $k\in\mathbb{N}_+$ according to Proposition B.4, which completes the proof. □

Proposition B.6. *Suppose $\boldsymbol{H}\in\mathcal{C}$ with $\boldsymbol{\rho}^{(\boldsymbol{H})}$ finitely supported. Then there exists a finitely supported decreasing g such that $\boldsymbol{H}\in\mathcal{C}^{(l,g)}$.*

Proof. The proof is a straightforward application of Proposition B.4. □

Next we prove the main theorem on error bound.

Proof of theorem 4. (i) Lower bound. Since

$$\|\boldsymbol{H}-\hat{\boldsymbol{H}}\|=\sup_{t\in\mathbb{Z}}\sup_{\|\boldsymbol{x}\|_{\mathcal{X}}\le 1}\left|\sum_{s\in\mathbb{N}}\left(\boldsymbol{\rho}^{(\boldsymbol{H})}(s)-\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}(s)\right)^{\top}x(t-s)\right|,$$

by taking a specific $\boldsymbol{x}$ with $x_i(0) = \frac{1}{\sqrt{d}} \operatorname{sgn}(\boldsymbol{\rho}_i^{(\boldsymbol{H})}(t) - \boldsymbol{\rho}_i^{(\hat{\boldsymbol{H}})}(t))$ and $x_i(s) = 0$ otherwise, we have

$$\begin{aligned} &\geq \frac{1}{\sqrt{d}} \sup_{t\in\mathbb{N}} \left| \boldsymbol{\rho}^{(\boldsymbol{H})}(t) - \boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}(t) \right| \\ &\geq \frac{1}{\sqrt{d}} \sup_{t\in[l^K,\infty]} \left| \boldsymbol{\rho}^{(\boldsymbol{H})}(t) \right| \\ &\geq \frac{1}{\sqrt{d}} \sup_{t\in[l^K,\infty]} \|\boldsymbol{\rho}^{(\boldsymbol{H})}(t)\|_2, \end{aligned} \tag{55}$$

where the inequality holds by taking a specific $\boldsymbol{x}$ with $x_i(0) = 1$ and $x_i(s) = 0$ otherwise.

(ii) Upper bound. Following the proof of Theorem 2, or (53) gives

$$\|\boldsymbol{H} - \hat{\boldsymbol{H}}\| \leq \sum_{i=1}^{d} \left\| T_{l^K}(\boldsymbol{\rho}_i^{(\hat{\boldsymbol{H}})}) - T_{l^K}(\boldsymbol{\rho}_i^{(\boldsymbol{H})}) \right\| + \|\boldsymbol{\rho}_{[l^K,\infty]}^{(\boldsymbol{H})}\|_2. \tag{56}$$

The remaining task is to bound the first term. Based on Lemma 3, we only need to calculate the maximum possible rank of $T_{l^K}(\boldsymbol{\rho}_i^{(\hat{\boldsymbol{H}})})$. Let r_k denote the k-rank of $T_{l^K}(\boldsymbol{\rho}_i^{(\hat{\boldsymbol{H}})})$. From Remark A.2, by absorbing the scalar $s_{i_1 i_2 \ldots i_K}$ into any of the vector $\boldsymbol{u}_{i_k}^{(k)}$, we have the following relationship:

$$d \prod_{k=1}^{K} r_k + lK \geq d \prod_{k=1}^{K} r_k + \sum_{k=1}^{K} r_k \geq \sum_{k=2}^{K} M_k M_{k-1}, \tag{57}$$

thus we have $\prod_{k=1}^{K} r_k \geq \frac{1}{d}(\sum_{k=2}^{K} M_k M_{k-1} - lK) = M$, which implies $\operatorname{rank} T_{l^K}(\boldsymbol{\rho}_i^{(\hat{\boldsymbol{H}})}) = \sum_{k=1}^{K} r_k \geq K M^{\frac{1}{K}}$. Combined with Definition 3 gives the conclusion. □

Next we look in details the two examples about comparison between RNNs and CNNs in details.

Example where RNNs out-perform CNNs. We take a scalar input with $d = 1$. Consider a target $\boldsymbol{H} \in \mathcal{C}$ with the representation $\boldsymbol{\rho}^{(\boldsymbol{H})}(t) = \gamma^t$, where $0 < \gamma < 1$. It is easy for RNNs to approximate this target, since the representation has a power form. In fact, we have $\boldsymbol{H} \in \mathcal{H}_{\text{RNN}}^{(1)}$, i.e. a RNN with one hidden unit is sufficient to achieve an exact representation for any $\gamma \in (0, 1)$.

For any CNN model $\hat{\boldsymbol{H}} \in \mathcal{H}_{\text{CNN}}^{(l,K,\{M_k\})}$, based on the lower bound of Theorem 4, we have that

$$\|\boldsymbol{H} - \hat{\boldsymbol{H}}\|^2 \geq \frac{1}{\sqrt{d}} \sup_{t\in[l^K,\infty]} \|\boldsymbol{\rho}^{(\boldsymbol{H})}(t)\|_2. \tag{58}$$

Thus, in order to achieve an approximation error with $\|\boldsymbol{H} - \hat{\boldsymbol{H}}\| < \epsilon$, we have

$$\sup_{t\in[l^K,\infty]} \|\boldsymbol{\rho}^{(\boldsymbol{H})}(t)\|_2 = \gamma^{l^K} < \epsilon. \tag{59}$$

This implies $l^K \geq \frac{\log(\epsilon)}{\log(\gamma)}$. That is, the number of layers necessary to achieve an approximation error smaller than ϵ diverges to infinity as γ approaches 1.

Example where CNNs out-perform RNNs. We still take a scalar input with $d = 1$. Consider a target $\boldsymbol{H} \in \mathcal{C}$ with the representation

$$\rho^{(\boldsymbol{H})}(t) = \begin{cases} 1, & t = 2^K \\ 0, & \text{otherwise} \end{cases}, \quad K \in \mathbb{N}_+. \tag{60}$$

We have $\boldsymbol{H} \in \mathcal{H}_{\text{CNN}}^{(2,K,\{1\})}$. That is, a K-layer CNN with one channel per layer is sufficient to achieve an exact representation.

Recall that RNN approximates the target $\boldsymbol{\rho}^{(\boldsymbol{H})}$ with a power sum $\boldsymbol{\rho}^{(\hat{\boldsymbol{H}})}(s) = c^\top W^{s-1}U$. Suppose here $W \in \mathbb{R}^{m\times m}$ is a diagonalisable matrix with negative eigenvalues. It has some special structures which are summarised in the following theorem.

Theorem B.7. *(Borwein & Erdélyi, 1996) Let* $E_m := \left\{ u : u(t) = c_0 + \sum_{i=0}^{m} c_i \gamma_i^t,\ c_i \in \mathbb{R}, \gamma_i > 0 \right\}$, *then*

$$\sup_{u\in E_k} \frac{|u'(y)|}{\sup_{s\in[a,b]} u(s)} \leq \frac{2m-1}{\min\{y-a, b-y\}},\ \ y \in (a,b). \tag{61}$$

We rewrite this theorem into a discrete form.

Corollary B.8. *Let* $u \in E_m$. *Then*

$$|u(t+1) - u(t)| \leq \frac{2m}{t} \sup_{s\geq 0} u(s). \tag{62}$$

Proof. By the mean value theorem, there exists $y \in [t, t+1]$ such that $|u(t+1) - u(t)| = |u'(y)|$. The corollary then follows from Theorem B.7. □

For a fixed m, the changes between $u(t+1)$ and $u(t)$ approaches zero as t goes to infinity. This implies that if there is a sudden change in u far from the origin, the number of terms m must be large.

In order to achieve an approximation error with $\|\boldsymbol{H} - \hat{\boldsymbol{H}}\| < \epsilon$, by taking a specific $\boldsymbol{x}$ as the unit sample function where $x(0) = 0$ and $x(s) = 0$ otherwise , we have

$$\begin{aligned}
\epsilon > \sup_t \sup_{\|\boldsymbol{x}\|\leq 1} |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| &\geq \sup_t |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| && (63)\\
&= \sup_t |\boldsymbol{\rho}^{(\boldsymbol{H})}(t) - c^\top W^{t-1} U| && (64)\\
&= \sup_t |\boldsymbol{\rho}^{(\boldsymbol{H})}(t) - u(t)|, \quad u \in E_{m^2}. && (65)
\end{aligned}$$

Since

$$\begin{aligned}
|u(2^K + 1)| &< \epsilon, && (66)\\
|u(2^K) - 1| &< \epsilon, && (67)
\end{aligned}$$

we have

$$\begin{aligned}
|u(2^K+1) - u(2^K)| &= |u(2^K+1) - 1 - u(2^K) + 1| && (68)\\
&> 1 - |u(2^K+1)| - |u(2^K) - 1| && (69)\\
&> 1 - 2\epsilon && (70)
\end{aligned}$$

Combining with Corollary B.8 gives

$$m^2 > 2^{K-1} \frac{1-2\epsilon}{1+\epsilon}. \tag{71}$$

As K increases, the number of parameters needed for RNNs to achieve an error less than ϵ increases exponentially, while this increment is linear for CNNs.

C. Special structures of dilated convolutions

In this section, we discuss an interesting structure of dilated convolutions.

Proposition C.1. *Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K$ be K filters with the same filter size l, $\boldsymbol{s} = (s_1, s_2, \ldots, s_K)$ with $0 \le s_k \le l-1$, $k = 1, \cdots, K$. Suppose all entries of $\boldsymbol{w}_k$ are zero except $w_k(s_k) = 1$, then*

$$(\boldsymbol{w}_K \underset{l^{K-1}}{\circledast} \boldsymbol{w}_{K-1} \underset{l^{K-2}}{\circledast} \cdots \underset{l}{\circledast} \boldsymbol{w}_1)(t) = \begin{cases} 1, & t = \hat{t} \\ 0, & \textit{otherwise} \end{cases}, \tag{72}$$

where $\hat{t} = (s_K s_{K-1} \ldots s_1)_l := \sum_{i=0}^{K-1} s_{i+1} l^i$. That is, $\hat{t}$ can be written as a base l expansion with digits s_K to s_1.

Proof. We prove this by induction. When $K = 1$, the conclusion is obvious. Suppose the conclusion holds for K, then by (47),

$$(\boldsymbol{w}_{K+1} \underset{l^K}{\circledast} \boldsymbol{f}_K)(t) = (w_1 c_1, w_1 c_2, \ldots, w_1 c_{l^K}, w_2 c_1, w_2 c_2, \ldots, w_2 c_{l^K}, \ldots, \ldots, w_l c_1, w_l c_2, \ldots, w_l c_{l^K}). \tag{73}$$

Suppose $w_{K+1}(s_{K+1}) = c_m = 1$, then the position of 1 in the above vector is $s_{K+1}\, l^K + c_m$, which means the results also holds for $K+1$. □

This result allows us to construct a filter with value 1 at any specific position $\hat{t}$, by choosing filters $\{\boldsymbol{w}_k\}$ according to the base l expansion of $\hat{t}$. We illustrate this by an example.

Example C.1. *Let $l = 4, K = 3$ and $\boldsymbol{w}_1 = (0,0,0,1), \boldsymbol{w}_2 = (1,0,0,0), \boldsymbol{w}_3 = (0,1,0,0)$. The positions of value 1 are recorded in $\boldsymbol{s} = (3,0,1)$. Then*

$$(\boldsymbol{w}_3 \underset{4^2}{\circledast} \boldsymbol{w}_2 \underset{4^1}{\circledast} \boldsymbol{w}_1)(t) = \begin{cases} 1, & t = 19 = (103)_4 \\ 0, & \textit{otherwise} \end{cases} \tag{74}$$

Based on above, one can define a notion of sparsity, which gives another sufficient condition for the exact representation.

Definition C.2. *Suppose $\boldsymbol{\rho}^{(\boldsymbol{H})}$ is a finitely supported sequence. The sparsity of $\boldsymbol{\rho}^{(\boldsymbol{H})}$ is defined as the number of its non-zero elements, which is denoted by $\|\boldsymbol{\rho}^{(\boldsymbol{H})}\|_0$.*

Corollary C.3. *Suppose $\boldsymbol{\rho}^{(\boldsymbol{H})}$ is a finitely supported sequence with $r(\boldsymbol{\rho}^{(\boldsymbol{H})}) \le l^K - 1$. Then $K\|\boldsymbol{\rho}^{(\boldsymbol{H})}\|_0$ filters are sufficient to achieve an exact representation.*

Proof. This follows from Theorem C.1, where for each non-zero element, we can use K filters to generate it. □

In the main text, we use the condition that $K(M+1)^{\frac{1}{K}} \ge \operatorname{rank} T_{l^K}(\boldsymbol{\rho}^{(\boldsymbol{H})})$ to ensure an exact representation. Instead of calculating the rank of $T_{l^K}(\boldsymbol{\rho}^{(\boldsymbol{H})})$, Corollary C.3 gives us another way to decide the number of filters sufficient to have an exact representation. This gives us the insight that if a target is sparse in the sense that $\|\boldsymbol{\rho}^{(\boldsymbol{H})}\|_0$ is small, it can also be efficiently approximated by CNNs.

Remark C.1. *Notice that the rank of a tensor not only depends on its sparsity, but also depends on specific positions of the non-zero elements. To illustrate, consider the following example*

$$\boldsymbol{\rho}_1 = \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} \textit{ and } \boldsymbol{\rho}_2 = \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix}. \tag{75}$$

Both of them have a sparsity 2, but $\operatorname{rank} T_{2^2}(\rho_1) = 2$ while $\operatorname{rank} T_{2^2}(\rho_2) = 4$.