On the Memory Mechanism of Tensor-Power Recurrent Models: Supplementary Materials

Hejia Qiu[⋆]
University of Nottingham Ningbo China
University of Nottingham UK

Chao Li[⋆,§]
RIKEN-AIP
Tokyo, Japan

Ying Weng[§]
University of Nottingham Ningbo China
University of Nottingham UK

Zhun Sun
BIGO PTE. LTD.
Singapore

Xingyu He
University of Nottingham Ningbo China
University of Nottingham UK

Qibin Zhao[§]
RIKEN-AIP
Tokyo, Japan

1 ADDITIONAL DISCUSSIONS

1.1 Examples of the "symmetric" property and "index-shift" operations mentioned in the notation

The *symmetric* structure of a tensor is defined as the invariance if arbitrarily reshuffling a sub-collection of the indices. For instance, assume that we have a degree-3 tensor $\mathcal{G} \in \mathbb{R}^{10\times 10\times 10}$, of which the indices can be represented as $\mathcal{G}_{i_1,i_2,i_3}$, then the tensor $\mathbf{G}$ is symmetric among the first 2 indices implies that the value of $\mathcal{G}$ will be not changed if we arbitrarily exchange its first 2 indices, i.e., $\mathcal{G}_{i_1,i_2,i_3} = \mathcal{G}_{i_2,i_1,i_3}$.

The *index-shift* operation of a tensor is defined as rotating the indices of a tensor in counterclockwise order. For instance, assume a degree-$(p+1)$ tensor $\mathcal{G} \in \mathbb{R}^{n^{(p+1)}}$, of which the indices can be represented as $\mathcal{G}_{i_1,i_2,\ldots,i_p,i_{p+1}}$. The k-step $(0 < k < p)$ index-shift operation of $\mathcal{G}$ among the first p indices outputs a new degree-$(p+1)$ tensor $IS_k(\mathcal{G}) \in \mathbb{R}^{n^{(p+1)}}$, of which the indices are converted as $\mathcal{G}_{i_{1+k},i_{2+k},\ldots,i_{p+1},i_1,\ldots,i_k}$. Moreover, we can easily know that *the index-shift operation did not change the spectral norm of the tensor*. This claim would be used in the PROOFS sections.

1.2 Visualization of the bound given in Theorem 1

Theorem 1 in the main paper shows that the long memory property of the TP recurrent model requires a sufficiently large degree parameter p. In Figure 1, we visualize the curve the bound given in Theorem 1. As shown in Figure 1, to obtain the long memory property, the TP recurrent model requires a higher degree with decreasing the variance σ^2 . As discussed in the main paper, the weights of a well-trained RNN are generally far away from 1. In this case, the variance σ^2 would be quite small. It implies that a high model degree is necessary to obtain the long memory. On the other side, we can see that the bound decreases when increasing the dimension of the hidden state. It is because the spectral norm of the weight tensor becomes more easily larger than 1 if fixing the distribution yet increasing the dimension. This fact partially reflects why an RNN with higher dimensional hidden states can "remember" more information from data, yielding lower training loss.

[⋆]Equal Contribution.
[§]Correspondence to: Chao Li<chao.li@riken.jp>;
Ying Weng<yingweng@gmail.com>;
Qibin Zhao<qibin.zhao@riken.jp>

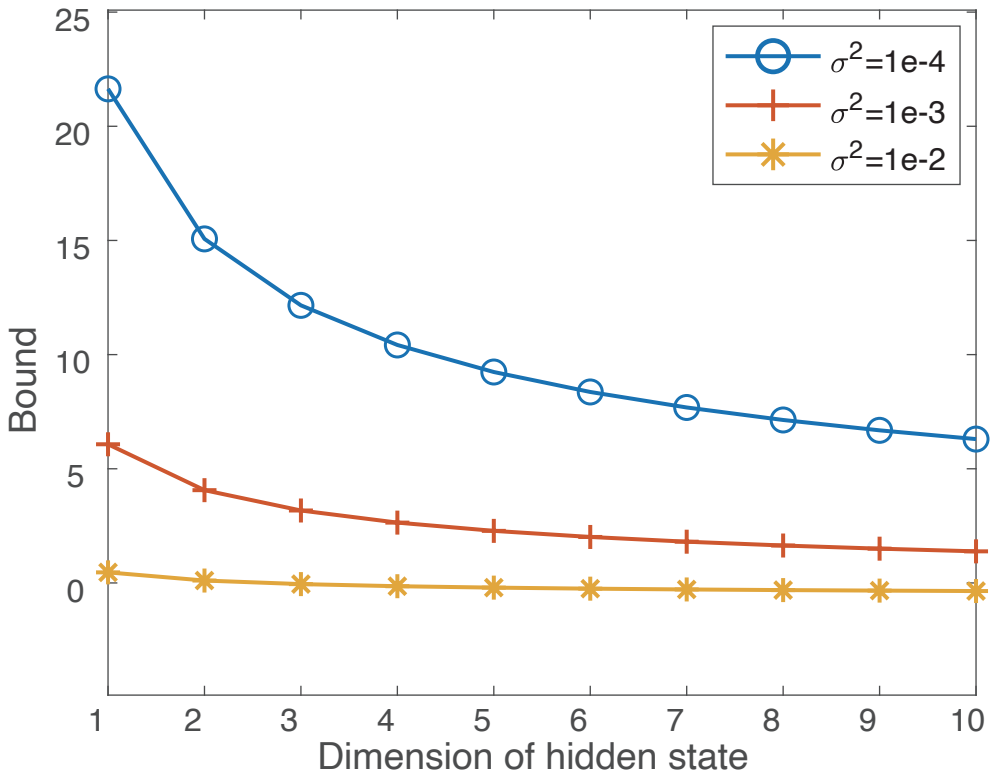

Figure 1: Illustration to the bound given in Theorem 1 of the main paper, where we set $\delta = 1e-9$.

2 PROOFS

Below, we give detailed proofs of the results that are missing in the main paper. Before the proofs, we first recall the two assumptions used in the main paper:

Assumption 1. (1) *The joint density function of* $\varepsilon(t)$ *is continuous and positive everywhere;* (2) *for some* $\kappa \geq 2$, $\mathbb{E}\|\varepsilon^{(t)}\|_2^\kappa < \infty$.

Assumption 2 (sub-Gaussian and decoupling). *The tensor* $\mathcal{M}$ *is obtained by the average over all the first* p *indices shifted variants of a tensor* $\mathcal{A} \in \mathbb{R}^{n^{(p+1)}}$*, of which each entries* $\mathcal{A}_{i_1,i_2,\ldots,i_{p+1}}$ *is independent, zero-mean and satisfied* $\mathbb{E}\left(e^{t\mathcal{A}_{i_1,i_2,\ldots,i_{p+1}}}\right) \leq e^{\sigma^2 t^2/2}$.

2.1 Proof of Lemma 1

Recall the tensor-power recurrent network process (TP-RNP):

$$\begin{aligned}\mathbf{s}^{(t)} &= \mathcal{M} \times_1 \mathbf{s}^{(t-1)} \times_2 \cdots \times_p \mathbf{s}^{(t-1)} + \mathbf{e}^{(t)} \\ &= \mathcal{M} \cdot \left(\mathbf{s}^{(t-1)}\right)^{\otimes p} + \mathbf{e}^{(t)}, \quad \forall t\end{aligned}, \tag{1}$$

Lemma 1 in the main paper shows TP-RNP has short memory if the spectral norm of the tensor $\mathcal{M}$ is bounded.

Lemma 1. *Under Assumption 1, the tensor-power recurrent network process (TP-RNP) has short memory under Def. 1 if the spectral norm of the tensor* $\mathcal{M}$ *obeys* $\|\mathcal{M}\|_2 < 1$.

Proof. Define the function $M(\mathbf{x}) := \mathcal{M} \times_1 \mathbf{x} \times \cdots \times_p \mathbf{x}$, then the operator norm $\|M\|_{op}$ obeys

$$\|M\|_{op} = \sup_{\|\mathbf{x}\|_2 \leq 1} \|M(\mathbf{x})\|_2 \leq \|\mathcal{M}\|_2 = \sup_{\|\mathbf{x}_i\|_2 \leq 1, i \in [p]} . \|\mathcal{M} \times_1 \mathbf{x}_1 \times_2 \cdots \times_p \mathbf{x}_p\|_2 \tag{2}$$

As $\|\mathcal{M}\|_2 < 1$, the function $M(\mathbf{x})$ is a bounded operator. There therefore exists two constants $0 < a < 1$ and b, such that the inequality $\|M(\mathbf{x})\|_2 \leq a\|\mathbf{x}\|_2 + b$ for all $\mathbf{x} \in \mathbb{R}^n$. Known from Theorem 1 given in the main paper that TP-RNP is geometrically ergodic, and therefore has short memory. □

2.2 Proof of Theorem 1

Theorem 1 (**Long memory requires a high model degree.**). *Under Assumptions 1 and 2, with high probability, if TP-RNP* (1) *has the long memory under Def. 1 given in the main paper, then the following inequality obeys:*

$$p \geq \frac{p_0}{2}\left(1 + \sqrt{1 + \frac{C_1}{n\sigma^2} - \frac{C_2}{n}}\right) - 1, \tag{3}$$

where $p_0 = \log(3/2)$*, and* C_1, C_2 *denote two positive constants.*

Proof. We prove the theorem from its contrapositive side. First, we prove the conditions to bound the spectral norm of the degree-$(p+1)$ tensor $\mathcal{A} \in \mathbb{R}^{n^{(p+1)}}$, which are used to generate the partially symmetric tensor $\mathcal{M}$ in TP-RNP. Known from Theorem 1 in [Tomioka and Suzuki, 2014], a non-asymptotic bound of the spectral norm $\mathcal{A}$ is given under Assumption 2:

$$\|\mathcal{A}\|_2 \leq \sqrt{8\sigma^2\left(n(p+1)\log\left((p+1)/p_0\right) + \log(2/\delta)\right)}, \tag{4}$$

with probability at least $1-\delta$. Because the index-shift operations do not change the spectral norm of the tensor, we have $\|\mathcal{M}\|_2 \leq \sqrt{8\sigma^2\left(n(p+1)\log\left((p+1)/p_0\right) + \log(2/\delta)\right)}$ by Assumption 2. Therefore, we can know by some calculation that $\|\mathcal{M}\|_2$ is upper bounded if the following inequality is held:

$$n(p+1)\log((p+1)/p_0) + \log(2/\delta) < 1/8\sigma^2. \tag{5}$$

Let $\bar{p} = (p+1)/p_0$ and use the inequality $\log(\bar{p}) \leq \bar{p} - 1$, then we know the upper bound exists if

$$\bar{p}^2 - \bar{p} < 1/8\sigma^2 p_0 n - \log(2/\delta)/np_0. \tag{6}$$

Since $\bar{p} > 0$, we can solve the above inequality by finding the positive root of the quadratic equation on the left-hand side, i.e.,

$$\bar{p} < \frac{1 + \sqrt{1 + \frac{C_1}{n\sigma^2} - \frac{C_2}{n}}}{2}, \tag{7}$$

where $C_1 = 1/(2\log(3/2))$ and $C_2 = 4\log(2/\delta - 3/2)$. Known from Lemma 1 that the model has short-term memory if the spectral norm of $\mathcal{M}$ is bounded, we take the above inequality to obtain our claim. □

2.3 Proof of Lemma 2

Recall the tensor-power (TP) recurrent model given in the main paper:

$$\begin{aligned}\mathbf{h}^{(t)} &= \mathcal{G} \times_1 \begin{pmatrix}\mathbf{x}^{(t)}\\ \mathbf{h}^{(t-1)}\end{pmatrix} \times_2 \cdots \times_p \begin{pmatrix}\mathbf{x}^{(t)}\\ \mathbf{h}^{(t-1)}\end{pmatrix}\\ &= \mathcal{G} \cdot \begin{pmatrix}\mathbf{x}^{(t)}\\ \mathbf{h}^{(t-1)}\end{pmatrix}^{\otimes p}\end{aligned}. \tag{8}$$

The following lemma gives the Jacobian of the model:

Lemma 2 (**Jacobian of the model**). *For any tensor* $\mathcal{G} \in \mathbb{R}^{n^p \times m}$ *of degree-*$(p+1)$*,* $p > 0$*, the Jacobian matrix* $\frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}}$ *with respect to Eq.* (8) *is equal to*

$$\begin{aligned}&J\left(\mathbf{h}^{(i-1)}; \mathbf{x}^{(i)}\right)\\ &= \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}} = \sum_{k=1}^{p}\left(\mathcal{G} \cdot \begin{pmatrix}\mathbf{x}^{(i)}\\ \mathbf{h}^{(i-1)}\end{pmatrix}^{\otimes p/\{k\}}\right) \times_{p+1} \begin{pmatrix}\mathbf{0}_l\\ \mathbf{I}_m\end{pmatrix},\end{aligned} \tag{9}$$

where $\mathbf{0}_l \in \mathbb{R}^{l\times m}$ *denotes the matrix filled by zeros,* $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ *is an identity matrix, and the operator* $(\cdot)^{\otimes p/\{k\}}$ *denotes the sequential "tensor-vector" product along the indices in the ordered set* $[p]/\{k\}$*. If* $\mathcal{G}$ *is symmetric among the first* p *indices, then Eq.* (9) *can be simplified as*

$$J_s(\mathbf{h}^{(i-1)};\mathbf{x}^{(i)}) = p\left(\mathcal{G}\cdot\begin{pmatrix}\mathbf{x}^{(t)}\\ \mathbf{h}^{(t-1)}\end{pmatrix}^{\otimes(p-1)}\right)\times_{p+1}\begin{pmatrix}\mathbf{0}_l\\ \mathbf{I}_m\end{pmatrix}. \tag{10}$$

Proof. At the beginning, we first give the Jacobian of Eq. (8) without the force term $\mathbf{x}^{(t)}$. Specifically, assume the function

$$\mathbf{h}^{(t)} = \mathcal{G}_{hh}\cdot\mathbf{h}^{(t-1),\otimes p}, \tag{11}$$

Then

Lemma 3 (**Jacobian without the force term**). *The Jacobian matrix* $\frac{\partial\mathbf{h}^{(i)}}{\partial\mathbf{h}^{(i-1)}}$ *with respect to Eq.* (11) *is equal to*

$$\frac{\partial\mathbf{h}^{(i)}}{\partial\mathbf{h}^{(i-1)}} = \sum_{k=1}^{p}\mathcal{G}_{hh}\left(\mathbf{h}^{(i-1)}\right)^{\otimes p/\{k\}}, \tag{12}$$

where the operator $(\cdot)^{\otimes p/\{k\}}$ *denotes the sequential "tensor-vector" product along the modes in the ordered set* $[p]/\{k\}$*.*

Proof. Construct the functions f_l as

$$f_l(\mathbf{x}_1,\ldots,\mathbf{x}_p) = \mathcal{G}_{hh,l}\times_1\mathbf{x}_1\times_2\cdots\times_p\mathbf{x}_p, \tag{13}$$

where $\mathcal{G}_{hh,l}$ denotes the sub-tensor of $\mathcal{G}_{hh}$ by fixing the last index equaling l. Therefore, the above equation can be rewritten as

$$\begin{aligned} &f_l(\mathbf{x}_1,\ldots,\mathbf{x}_p)\\ &=\underbrace{\left(\bigotimes\nolimits_{k\in[p]/\{i\}}\mathbf{x}_k\right)^{\top}[\mathcal{G}_{hh,l}]_{(i)}^{\top}}_{H_i:=}\mathbf{x}_i,\quad i\in[p], \end{aligned} \tag{14}$$

where $\bigotimes_{k\in[p]/\{i\}}\mathbf{x}_k$ denotes the sequential Kronecker product based on the inverse order, *i.e.*, $\mathbf{x}_p\otimes\mathbf{x}_{p-1}\otimes\cdots$[1], and $[\cdot]_{(i)}$ denotes unfolding a tensor along the i-th index [Cichocki et al., 2007]. Then the Jacobian of f_l is given as

$$(Jac_{*,l})_{(\mathbf{x}_1,\ldots,\mathbf{x}_p)} = \begin{pmatrix}H_1 & H_2 & \cdots & H_p\end{pmatrix}. \tag{15}$$

Then we have

$$\begin{aligned} &Jac_l(\mathbf{x}\mapsto\mathbf{x}^{\otimes p})\\ &=(Jac_{*,l})_{(\mathbf{z},\ldots,\mathbf{z})}\cdot Jac(\mathbf{x}\mapsto(\mathbf{x},\ldots,\mathbf{x}))_{\mathbf{z}}\\ &=\left(\sum_{i\in[p]}H_i\right)_{(\mathbf{x},\ldots,\mathbf{x})}\\ &=\sum_{i\in[p]}\left(\bigotimes\nolimits_{k\in[p]/\{i\}}\mathbf{x}_k\right)^{\top}[\mathcal{G}_{hh,l}]_{(i)}^{\top}\\ &=\sum_{i\in[p]}\mathcal{G}_{hh,l}\cdot\mathbf{x}^{\otimes p/\{i\}} \end{aligned}, \tag{16}$$

where the last equation holds using the basic calculation of tensor algebra. Overall the full Jacobian $\frac{\partial\mathbf{h}^{(i)}}{\partial\mathbf{h}^{(i-1)}}$ is obtained by concatenating Eq. (16) for all possible l. □

[1]We apply the symbol $\otimes$ to denote both the Kronecker product and tensor product without ambiguity.

Using the result in Lemma 2, we obtain Eq. (9) by the chain rule. Construct a dummy vector as $\mathbf{s}^{(t-1)} = \left(\mathbf{x}^{(t),\top}, \mathbf{h}^{(t-1),\top}\right)^\top \in \mathbb{R}^n$ with the function

$$\mathbf{s} = \bar{\mathcal{G}} \left(\mathbf{s}^{(t-1)}\right)^{\otimes p}, \tag{17}$$

where $\bar{\mathcal{G}} \in \mathbb{R}^{n^{(p+1)}}$ denotes a tensor by padding the $\mathcal{G}$ with zeros along the last index. Using the dummy vectors, we have the following equation by the chain rule of derivatives,

$$\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}} = \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{s}} \cdot \frac{\partial \mathbf{s}}{\partial \mathbf{s}^{(t-1)}} \cdot \frac{\partial \mathbf{s}^{(t-1)}}{\partial \mathbf{h}^{(t-1)}}. \tag{18}$$

The proposition is proved by substituting the result in Lemma 2 and

$$\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{s}} = \begin{pmatrix} \mathbf{I}_m & \mathbf{0}_l \end{pmatrix} \text{ and } \frac{\partial \mathbf{s}^{(t-1)}}{\partial \mathbf{h}^{(t-1)}} = \begin{pmatrix} \mathbf{0}_l \\ \mathbf{I}_m \end{pmatrix} \tag{19}$$

into Eq. (18). □

2.4 Proof of Theorem 2

Theorem 2 (**High degree models lead to unstable dynamics.**). *Given the tensor $\mathcal{G}$ with the partially structure of the first p indices, of which $\mathcal{G}$ also has non-zero sub-blocks with respect to $\mathbf{U} = \left(\mathbf{0}_l^\top \ \mathbf{I}_m\right)^\top$, i.e. $\|\mathcal{G} \cdot \mathbf{U}^{\otimes p}\|_2 \neq 0$, for any positive number $K > 0$ and $p > 1$, there always exist a pair of vectors $\mathbf{h} \in \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^l$, such that $\|J_s(\mathbf{h}; \mathbf{x})\|_2 > M$.*

Proof. We first have a lower bound of the Jacobian J_s. Specifically, let $\mathbf{s} := \left(\mathbf{x}^\top \ \mathbf{h}^\top\right)^\top \in \mathbb{R}^n$ and $\mathbf{s} \neq 0$, then we have

$$\begin{aligned} \|J_s(\mathbf{h}; \mathbf{x})\|_2 &= p \left\| \left(\mathcal{G} \cdot \mathbf{s}^{\otimes(p-1)}\right) \mathbf{U} \right\|_2 \\ &= p \sup_{\|\mathbf{v}\|_2 \leq 0} \left\| \left(\mathcal{G} \cdot \mathbf{s}^{\otimes(p-1)}\right) \mathbf{U}\mathbf{v} \right\|_2 . \end{aligned} \tag{20}$$

The second equation holds according to the definition of the matrix spectral norm. Next, we let $\mathbf{x} = 0$ and $\mathbf{h} \neq 0$, then the equations above can be rewritten as

$$\begin{aligned} &\|J_s(\mathbf{h}; \mathbf{x})\|_2 \\ &= p \sup_{\|\mathbf{v}\|_2 \leq 0} \left\| \underbrace{\left(\mathcal{G} \cdot \mathbf{U}^{\otimes p}\right)}_{\bar{\mathcal{G}} :=} \mathbf{h}^{\otimes(p-1)} \times_p \mathbf{v} \right\|_2 \\ &= p\|\mathbf{h}\|^{p-1} \sup_{\|\mathbf{v}\|_2 \leq 0} \left\| \bar{\mathcal{G}} \cdot \underbrace{\left(\mathbf{h}/\|\mathbf{h}\|_2\right)^{\otimes(p-1)}}_{\bar{\mathbf{h}} :=} \times_p \mathbf{v} \right\|_2 . \\ &\geq p\|\mathbf{h}\|^{p-1} \left\|\bar{\mathcal{G}} \cdot \bar{\mathbf{h}}^{\otimes p}\right\|_2 \end{aligned} \tag{21}$$

The inequality holds by $\mathbf{v} = \bar{\mathbf{h}}$. Since $\bar{\mathcal{G}} \neq 0$, there must exist $\bar{\mathbf{h}} \neq 0$ such that $\left\|\bar{\mathcal{G}} \cdot \bar{\mathbf{h}}^{\otimes p}\right\|_2 > 0$. Therefore, given arbitrary $M > 0$, we can always find $\mathbf{h} \neq 0$ and $\|\mathbf{h}\|_2 > \left(\frac{M}{p\left\|\bar{\mathcal{G}} \cdot \bar{\mathbf{h}}^{\otimes p}\right\|_2}\right)^{1/(p-1)}$ such that $\|J_s(\mathbf{h}; \mathbf{x})\|_2 > M$. □

3 ADDITIONAL EXPERIMENTAL RESULTS

3.1 The single-cell experiment

3.1.1 Datasets

In this experiment, we exploit the same datasets in [Zhao et al., 2020] to demonstrate the effectiveness of the model. The description of the datasets is shown below. The lengths for training, validation, and test sets are given in Table 1. More detailed discussions about the statistic property of data are given in [Zhao et al., 2020].

ARFIMA series (ARFIMA). The data are generated as a series of length 4001 using the model $(1 - 0.7B + 0.4B^2)(1 - B)^{0.4}Y_t = (1 - 0.2B)\varepsilon_t$ with obvious long memory effect.

Dow Jones Industrial Average (DJI). The raw dataset contains DJI daily closing prices from 2000 to 2019 obtained from Yahoo Finance. The data is converted to absolute log return for 5030 days in order to model the long memory effect in volatility.

Metro interstate traffic volume (Traffic). The raw dataset contains hourly Interstate 94 Westbound traffic volume for MN DoT ATR station 301, roughly obtained from MN Department of Transportation. The data is converted to de-seasoned daily data with length 1860.

Tree ring (Tree). Dataset contains 4351 tree ring measures of a pine from India Garden, Nevada Gt Basin obtained from R package tsdl.

Table 1: Data length of training, validation, and test sets for each dataset.

Dataset	Training	Validation	Test
ARFIMA	2000	1200	800
DJI	2500	1500	1029
Traffic	1400	200	259
Tree	2500	1000	850

3.1.2 Additional experimental results

Average RMSE and standard deviation of the experimental results are given in the main paper. Below we provide results in terms of mean absolute error (MAE, Table 2) and mean absolute percentage error (MAPE, Table 3), respectively.

3.2 The seq2seq experiment

3.2.1 Datasets and Pre-Processing

Genz. Genz functions can be used as the basic expression of higher-order functions, and can also be used for the simplified simulation of time series data sets without noise. We generated 1,000 sequences of length 100 based on Genz functions. The initial point of each sequence is randomly generated by random seeds.

TrafficLA. The Los Angeles County highway network traffic dataset provided by the California department of transportation contains speed readings collected by speed sensors. The sensor reports data every five minutes. Similar to Yu's work, we use the average speed collected by other sensors at the same time to fill in the missing values. Data collected from one district from this dataset is used in our experiment and we down-sample the sequences to every 20 minutes. In the experiment, the traffic dataset contains 1936 sequences with 72 timestamps, with the data collected by one sensor in a day taken as a sequence.

Solar. The solar dataset provided by NREL contains data points for synthetic solar photovoltaic (PV) power plants in the United States. It includes PV power generation records from PV power plants in the United States representing the year 2006. This dataset is intended for energy professionals to do the estimation of power production from hypothetical solar plants. We use the records of 137 plants in Alabama State during a week in the experiment. The data recording interval of each power station is down-sampled to 10 minutes, so those 137 sequences are with 1008 timestamps to be applied in PV power generation prediction within a week.

3.2.2 Hyper-Parameter Search

The search range of hyper-parameters is listed in Table 4. Compared with the traditional LSTM model, the order range needs to be additionally set to prevent parameter explosion in our method.

Table 2: Performance comparison in terms of MAE, where the average and standard deviation (in brackets) are reported, and the best results are highlighted in bold.

	ARFIMA	DJI(×100)	Traffic	Tree
RNN	0.9319	0.1977	233.442	0.2240
	(0.1550)	(0.0242)	(12.391)	(0.0064)
RNN2	0.9310	0.1861	233.419	0.2229
	(0.1430)	(0.0164)	(12.378)	(0.0057)
RWA	1.3330	0.2052	233.137	0.2379
	(0.0030)	(0.0164)	(7.425)	(0.0001)
MRNN	0.8710	**0.1835**	**232.794**	0.2202
	(0.0900)	(0.0165)	(12.149)	(0.0037)
LSTM	0.9070	0.1841	234.055	0.2215
	(0.0940)	(0.0182)	(11.149)	(0.0051)
MLSTM	0.9240	0.1895	233.142	0.2235
	(0.1320)	(0.0203)	(11.551)	(0.0060)
Ours	**0.8588**	0.2043	233.65	**0.2185**
	(0.0192)	(0.0360)	(4.5121)	(0.0015)

Table 3: Performance comparison in terms of MAPE, where the average and standard deviation (in brackets) are reported, and the best results are highlighted in bold.

	ARFIMA	DJI	Traffic	Tree
RNN	2.5760	1.4371	1.3943	0.2747
	(0.4030)	(0.2566)	(0.1998)	(0.0079)
RNN2	2.5570	1.4407	1.4092	0.2739
	(0.4420)	(0.2106)	(0.1789)	(0.0071)
RWA	**2.2370**	**1.2733**	1.3745	0.2939
	(0.1950)	(0.1702)	(0.1457)	(0.0005)
MRNN	2.7010	1.5031	1.4253	0.2706
	(0.2680)	(0.2045)	(0.1586)	(0.0044)
LSTM	2.5660	1.5725	1.3632	0.2727
	(0.3750)	(0.2283)	(0.1807)	(0.0060)
MLSTM	2.5500	1.3123	**1.3353**	0.2748
	(0.4370)	(0.1281)	(0.1926)	(0.0075)
Ours	2.8398	1.6034	1.4895	**0.2683**
	(0.0667)	(0.5811)	(0.1960)	(0.0015)

Table 4: Hyper-parameter search range in the experiment.

learning rate	$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
decay rate	$\{1, 0.9, 0.8, 0.7, 0.6\}$
hidden size	$\{8, 16, 32, 64, 128\}$
hidden layer	$\{1, 2, 3, 4\}$
initial order	$\{0.0, 1.0, 2.0, 3.0\}$
order range	$\{[0, 0.5], [0.5, 1], [0, 1], [1, 2]\}$

References

[Cichocki et al., 2007] Cichocki, A., Zdunek, R., and Amari, S.-i. (2007). Nonnegative matrix and tensor factorization [lecture notes]. *IEEE Signal Processing Magazine*, 25(1):142–145.

[Tomioka and Suzuki, 2014] Tomioka, R. and Suzuki, T. (2014). Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*.

[Zhao et al., 2020] Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., and Tian, G. (2020). Do RNN and LSTM have long memory? *arXiv preprint arXiv:2006.03860*.