Spectral Pruning for Recurrent Neural Networks

Takashi Furuya
Hokkaido University
takashi.furuya0101@gmail.com

Kazuma Suetake
AISIN SOFTWARE
kazuma.suetake@aisin-software.com

Koichi Taniguchi
Tohoku University
koichi.taniguchi.b7@tohoku.ac.jp

Hiroyuki Kusumoto
Nagoya University
kusumoto-108@outlook.com

Ryuji Saiin
AISIN SOFTWARE
ryuji.saiin@aisin-software.com

Tomohiro Daimon
AISIN SOFTWARE
tomohiro.daimon@aisin-software.com

Abstract

Recurrent neural networks (RNNs) are a class of neural networks used in sequential tasks. However, in general, RNNs have a large number of parameters and involve enormous computational costs by repeating the recurrent structures in many time steps. As a method to overcome this difficulty, RNN pruning has attracted increasing attention in recent years, and it brings us benefits in terms of the reduction of computational cost as the time step progresses. However, most existing methods of RNN pruning are heuristic. The purpose of this paper is to study the theoretical scheme for RNN pruning method. We propose an appropriate pruning algorithm for RNNs inspired by "spectral pruning", and provide the generalization error bounds for compressed RNNs. We also provide numerical experiments to demonstrate our theoretical results and show the effectiveness of our pruning method compared with the existing methods.

1 Introduction

Recurrent neural networks (RNNs) are a class of neural networks used in sequential tasks. However, in general, RNNs have a large number of parameters and involve enormous computational costs by repeating the recurrent structures in many time steps. These make their application difficult in edge computing devices. To overcome this difficulty, RNN compression has attracted increasing attention in recent years. It brings us more benefits in terms of the reduction of computational costs as the time step progresses, compared to deep neural networks (DNNs) without any recurrent structure. There are many RNN compression methods such as pruning (Narang et al., 2017a; Tang and Han, 2015; Zhang and Stadie, 2019; Lobacheva et al., 2017; Wang et al., 2019; Wen et al., 2020; Lobacheva et al., 2020), low rank factorization (Kliegl et al., 2017; Tjandra et al., 2017), quantization (Alom et al., 2018; Liu et al., 2018), distillation (Shi et al., 2019; Tang et al., 2016), and sparse training (Liu et al., 2021a,b; Dodge et al., 2019; Wen et al., 2017). This paper is devoted to the pruning of RNNs, and its purpose is to provide an RNN pruning method with the theoretical background.

Recently, Suzuki et al. (Suzuki et al., 2020) proposed a novel pruning method with the theoretical background, called *spectral pruning*, for DNNs such as the fully connected and convolutional neural network architectures. The idea of the proposed method is to select important nodes for each layer by minimizing the information losses (see (2) in (Suzuki et al., 2020)), which can be represented by the layerwise covariance matrix. The minimization only requires linear algebraic operations. Suzuki et al. (Suzuki et al., 2020) also evaluated generalization error bounds for networks compressed using spectral pruning (see Theorems 1 and 2 in (Suzuki et al., 2020)). It was shown that the generalization error bounds are controlled by the *degrees of freedom*, which represents the intrinsic dimensionality of a model, and is determined by the eigenvalues of the covariance matrix (Mallows, 2000; Caponnetto and De Vito, 2007). Hence, the characteristics of the eigenvalue distribution have an influence on the error bounds. We can also observe that in the generalization error bounds, there is a bias-variance tradeoff corresponding to compressibility. Numerical experiments have also demonstrated

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

the effectiveness of spectral pruning.

In this paper, we extend the theoretical scheme of spectral pruning to RNNs. Our pruning algorithm involves the selection of hidden nodes by minimizing the information losses, which can be represented by the time mean of the covariance matrix instead of the layerwise covariance matrix which appears in spectral pruning of DNNs. We emphasize that our information losses are derived from the generalization error bound. More precisely, we show that choosing compressed weight matrices which minimize the information losses reduces the generalization error bound we evaluated in Section 4.1 (see sentences after Theorem 4.5). We also remark that Suzuki et al. (Suzuki et al., 2020) has not clearly mentioned anything about how the information losses are derived. As in DNNs (Suzuki et al., 2020), we can provide the generalization error bounds for RNNs compressed with our pruning and interpret the degrees of freedom and the bias-variance tradeoff.

We also provide numerical experiments to compare our method with existing methods. We observed that our method outperforms existing methods, and gets benefits from over-parameterization (Chang et al., 2020; Zhang et al., 2021) (see Sections 5.2 and 5.3). In particular, our method can compress models with small degradation (see Remark 3.2) when we employ IRNN, which is an RNN that uses the ReLU as the activation function and initializes weights as the identity matrix and biases to zero (see (Le et al., 2015)).

The summary of our contributions is the following:

- A pruning algorithm for RNNs (Section 3) is proposed by the analysis of generalization error (Remark 4.3 and Theorem 4.8).
- The generalization error bounds for RNNs compressed with our pruning algorithm are provided (Theorem 4.8).

2 Related Works

One of the popular compression methods for RNNs is pruning that removes redundant weights based on certain criteria. For example, magnitude-based weight pruning (Narang et al., 2017a,b; Tang and Han, 2015) involves pruning trained weights that are less than the threshold value decided by the user. This method has to gradually repeat pruning and retraining weights to ensure that a certain accuracy is maintained. However, based on recent developments, the costly repetitions might not always be necessary. In one-shot pruning (Zhang and Stadie, 2019; Lee et al., 2018), weights are pruned once prior to training from the spectrum of the recurrent Jacobian. Bayesian sparsification (Lobacheva et al., 2017; Molchanov et al., 2017) induce sparse weight matrix by choosing the prior as log-uniform distribution, and the weights are also once pruned if the variance of the posterior over the weights is large.

While the above methods are referred to as weight pruning, our spectral pruning is a structured pruning where redundant nodes are removed. The advantage of the structured pruning over the weight pruning is that it more simply reduces computational costs. The implementation advantages of structured pruning are illustrated in (Wang et al., 2019). Although weight pruning from large networks to small networks is less likely to degrade accuracy, it usually requires an accelerator for addressing sparsification (see (Parashar et al., 2017)). The structured pruning method discussed in (Wang et al., 2019; Wen et al., 2020; Lobacheva et al., 2020) induces sparse weight matrices in the training process, and prunes weights close to zero, and does not repeat fine-tuning. In our pruning, weight matrices are trained by the usual way, and the compressed weight matrices consist of the multiplication of the trained weight matrix and the reconstruction matrix, and no need to repeat pruning and fine-tuning. The idea of the multiplication of the trained weight matrix and the reconstruction matrix is a similar idea to low rank factorization (Kliegl et al., 2017; Tjandra et al., 2017; Prabhavalkar et al., 2016; Grachev et al., 2019; Denil et al., 2013). In particular, the work (Denil et al., 2013) is most related to spectral pruning, and it employs the reconstruction matrix replacing the empirical covariance matrix with kernel matrix (see Section 3.1 in (Denil et al., 2013)).

In general, RNN pruning is more difficult than DNN pruning, because recurrent architectures are not robust to pruning, that is, even a little pruning causes accumulated errors and total errors increase significantly for many time steps. Such a peculiar problem for recurrent feature is also observed in dropout (see Introduction in (Gal and Ghahramani, 2016; Zaremba et al., 2014)).

Our motivation is to theoretically propose the RNN pruning algorithm. Inspired by (Suzuki et al., 2020), we focus on the generalization error bound, and we provide the algorithm so that the generalization error bound becomes smaller. Thus, the derivation of our pruning method would be theoretical, while that of existing methods such as the magnitude-based pruning (Narang et al., 2017a,b; Tang and Han, 2015; Wang et al., 2019; Wen et al., 2020) would be heuristic. For the study of the generalization error bounds for RNNs, we refer to (Tu et al., 2019; Chen et al., 2019; Akpinar et al., 2019; Joukovsky et al., 2021).

3 Pruning Algorithm

We propose a pruning algorithm for RNNs inspired by (Suzuki et al., 2020). See Appendix A for a review of spectral pruning for DNNs. Let $D = \{(X_T^i, Y_T^i)\}_{i=1}^n$ be the training data with time series sequences $X_T^i = (x_t^i)_{t=1}^T$ and $Y_T^i = (y_t^i)_{t=1}^T$, where $x_t^i \in \mathbb{R}^{d_x}$ is an input and $y_t^i \in \mathbb{R}^{d_y}$ is an output at time t. The training data are independently identically distributed. To train the appropriate relationship between input $X_T = (x_t)_{t=1}^T$ and output $Y_T = (y_t)_{t=1}^T$, we consider RNNs $f = (f_t)_{t=1}^T$ as

$$f_t = W^o h_t + b^o, \quad h_t = \sigma(W^h h_{t-1} + W^i x_t + b^{hi}),$$

for $t = 1, \dots, T$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, $h_t \in \mathbb{R}^m$ is the hidden state with the initial state $h_0 = 0$, $W^o \in \mathbb{R}^{d_y \times m}$, $W^h \in \mathbb{R}^{m \times m}$, and $W^i \in \mathbb{R}^{m \times d_x}$ are weight matrices, and $b^o \in \mathbb{R}^{d_y}$ and $b^{hi} \in \mathbb{R}^m$ are biases. Here, an element-wise activation operator is employed, i.e., we define $\sigma(x) := (\sigma(x_1), \dots, \sigma(x_m))^T$ for $x = (x_1, \dots, x_m) \in \mathbb{R}^m$.

Let $\widehat{f} = (\widehat{f}_t)_{t=1}^T$ be a trained RNN obtained from the training data D with weight matrices $\widehat{W}^o \in \mathbb{R}^{d_y \times m}$, $\widehat{W}^h \in \mathbb{R}^{m \times m}$, and $\widehat{W}^i \in \mathbb{R}^{m \times d_x}$, and biases $\widehat{b}^o \in \mathbb{R}^{d_y}$ and $\widehat{b}^{hi} \in \mathbb{R}^m$, i.e., $\widehat{f}_t = \widehat{W}^o \widehat{h}_t + \widehat{b}^o$, $\widehat{h}_t = \sigma(\widehat{W}^h \widehat{h}_{t-1} + \widehat{W}^i x_t + \widehat{b}^{hi})$ for $t = 1, \dots, T$. We denote the hidden state $\widehat{h}_t$ by

$$\widehat{h}_t = \phi(x_t, \widehat{h}_{t-1}),$$

as a function with inputs x_t and $\widehat{h}_{t-1}$. Our aim is to compress the trained network $\widehat{f}$ to the smaller network $f^\sharp$ without loss of performance to the extent possible.

Let $J \subset [m]$ be an index set with $|J| = m^\sharp$, where $[m] := \{1, \dots, m\}$, and let $m^\sharp \in \mathbb{N}$ be the number of hidden nodes for a compressed RNN $f^\sharp$ with $m^\sharp \leq m$. We denote by $\phi_J(x_t, \widehat{h}_{t-1}) = (\phi_j(x_t, \widehat{h}_{t-1}))_{j \in J}$ the subvector of $\phi(x_t, h_{t-1})$ corresponding to the index set J, where $\phi_j(x_t, \widehat{h}_{t-1})$ represents the j-th components of the vector $\phi(x_t, \widehat{h}_{t-1})$.

(i) Input information loss. The input information loss is defined by

$$L_\tau^{(A)}(J) := \min_{A \in \mathbb{R}^{m \times m^\sharp}} \left\{ \|\phi - A\phi_J\|_{n,T}^2 + \|A\|_\tau^2 \right\}, \quad (3.1)$$

where $\|\cdot\|_{n,T}$ is the empirical L^2-norm with respect to n and t, i.e.,

$$\begin{aligned}&\|\phi - A\phi_J\|_{n,T}^2 \\ &:= \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \left\| \phi(x_t^i, \widehat{h}_{t-1}^i) - A\phi_J(x_t^i, \widehat{h}_{t-1}^i) \right\|_2^2,\end{aligned}$$

where $\|\cdot\|_2$ is the Euclidean norm, $\|A\|_\tau^2 := \mathrm{Tr}[A I_\tau A^T]$ for the regularization parameter $\tau \in \mathbb{R}_+^{m^\sharp} := \{x \in \mathbb{R}^{m_l^\sharp} \mid x_j > 0,\ j = 1, \dots, m_l^\sharp\}$, and $I_\tau := \mathrm{diag}(\tau)$. Here, $\widehat{\Sigma}_{I,I'} \in \mathbb{R}^{K \times H}$ denotes the submatrix of $\widehat{\Sigma}$ corresponding to the index sets $I, I' \subset [m]$ with $|I| = K$, $|I'| = H$, i.e., $\widehat{\Sigma}_{I,I'} = (\widehat{\Sigma}_{i,i'})_{i \in I, i' \in I'}$. Based on the linear regularization theory (see e.g., (Gockenbach, 2016)), there exists a unique solution $\widehat{A}_J \in \mathbb{R}^{m \times m^\sharp}$ of the minimization problem of $\|\phi - A\phi_J\|_{n,T}^2 + \|A\|_\tau^2$, which has the form

$$\widehat{A}_J = \widehat{\Sigma}_{[m],J} \left(\widehat{\Sigma}_{J,J} + I_\tau \right)^{-1}, \quad (3.2)$$

where $\widehat{\Sigma}$ is the (noncentered) empirical covariance matrix of the hidden state $\phi(x_t, \widehat{h}_{t-1})$ with respect to n and t, i.e.,

$$\widehat{\Sigma} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \phi(x_t^i, \widehat{h}_{t-1}^i) \phi(x_t^i, \widehat{h}_{t-1}^i)^T. \quad (3.3)$$

We term the unique solution $\widehat{A}_J$ as the *reconstruction matrix*. Here, we would like to emphasize that the mean of the covariance matrix with respect to time t is employed in RNNs, while the layerwise covariance matrix is employed in DNNs (see Appendix A). By substituting the explicit formula of the reconstruction matrix $\widehat{A}_J$ into (3.1), the input information loss is reformulated as:

$$L_\tau^{(A)}(J) = \mathrm{Tr}\left[\widehat{\Sigma} - \widehat{\Sigma}_{[m],J} \left(\widehat{\Sigma}_{J,J} + I_\tau \right)^{-1} \widehat{\Sigma}_{J,[m]} \right]. \quad (3.4)$$

(ii) Output information loss. The hidden state of a RNN is forwardly propagated to the next hidden state or output, and hence, the two output information losses are defined by

$$L_\tau^{(B,o)}(J) := \sum_{j=1}^{d_y} \min_{\beta \in \mathbb{R}^{m^\sharp}} \left\{ \left\| \widehat{W}_{j,:}^o \phi - \beta^T \phi_J \right\|_{n,T}^2 + \left\| \beta^T \right\|_\tau^2 \right\}, \quad (3.5)$$

$$L_\tau^{(B,h)}(J) := \sum_{j \in J} \min_{\beta \in \mathbb{R}^{m^\sharp}} \left\{ \left\| \widehat{W}_{j,:}^h \phi - \beta^T \phi_J \right\|_{n,T}^2 + \left\| \beta^T \right\|_\tau^2 \right\}, \quad (3.6)$$

where $\widehat{W}_{j,:}^o$ and $\widehat{W}_{j,:}^h$ denote the j-th rows of the matrix $\widehat{W}^o$ and $\widehat{W}^h$, respectively. Then, the unique solutions of the minimization problems of $\|\widehat{W}_{j,:}^o \phi - \beta^T \phi_J\|_{n,T}^2 + \|\beta^T\|_\tau^2$ and $\|\widehat{W}_{j,:}^h \phi - \beta^T \phi_J\|_{n,T}^2 + \|\beta^T\|_\tau^2$ are $\widehat{\beta}^o = (\widehat{W}_{j,:}^o \widehat{A}_J)^T$ and $\widehat{\beta}_j^h = (\widehat{W}_{j,:}^h \widehat{A}_J)^T$, respectively. By substituting them into (3.5) and (3.6), the output information losses are reformulated as

$$\begin{aligned}&L_\tau^{(B,o)}(J) \\ &= \mathrm{Tr}\left[\widehat{W}^o \left(\widehat{\Sigma} - \widehat{\Sigma}_{[m],J} \left(\widehat{\Sigma}_{J,J} + I_\tau \right)^{-1} \widehat{\Sigma}_{J,[m]} \right) \widehat{W}^{o^T} \right],\end{aligned} \quad (3.7)$$

$$L_\tau^{(B,h)}(J)$$
$$= \mathrm{Tr}\left[\widehat{W}^h_{J,[m]}\Big(\widehat{\Sigma} - \widehat{\Sigma}_{[m],J}\big(\widehat{\Sigma}_{J,J} + I_\tau\big)^{-1}\widehat{\Sigma}_{J,[m]}\Big)\widehat{W}^{h^T}_{J,[m]}\right]. \tag{3.8}$$

Here, we remark that the output information losses $L_\tau^{(B,o)}(J)$ and $L_\tau^{(B,h)}(J)$ are bounded above by the input information loss $L_\tau^{(A)}(J)$ (see Remark 4.3).

(iii) Compressed RNNs. We construct the compressed RNN $f_J^\sharp$ by $f_{J,t}^\sharp = W_J^{\sharp o} h_{J,t}^\sharp + b_J^{\sharp o}$ and $h_{J,t}^\sharp = \sigma(W_J^{\sharp h} h_{J,t-1}^\sharp + W_J^{\sharp i} x_t + b_J^{\sharp hi})$ for $t = 1, \ldots, T$, where $W_J^{\sharp o} := \widehat{W}^o \widehat{A}_J$, $W_J^{\sharp h} := \widehat{W}^h_{J,[m]} \widehat{A}_J$, $W_J^{\sharp i} := \widehat{W}^i_{J,[d_x]}$, $b_J^{\sharp hi} := \widehat{b}^{hi}_J$, and $b_J^{\sharp o} := \widehat{b}^o$.

(iv) Optimization. To select an appropriate index set J, we consider the following optimization problem that minimizes the convex combination of the input and two output information losses:

$$\min_{\substack{J \subset [m] \\ s.t.\ |J| = m^\sharp}} \left\{\theta_1 L_\tau^{(A)}(J) + \theta_2 L_\tau^{(B,o)}(J) + \theta_3 L_\tau^{(B,h)}(J)\right\}, \tag{3.9}$$

for $\theta_1, \theta_2, \theta_3 \in [0,1]$ with $\theta_1 + \theta_2 + \theta_3 = 1$, where $m_l^\sharp \in [m]$ is a prespecified number. The optimal index $J^\sharp$ is obtained by the greedy algorithm. We term this method as *spectral pruning* (for a schematic diagram of spectral pruning, see Figure 1). The reason information losses are employed in the objective will be theoretically explained later, when the error bounds in Remark 4.3 and Theorem 4.5 are provided. We summarize our pruning algorithm in the following.

Algorithm 1 Spectral pruning

```
Require: Data set $D = \{(X_T^i, Y_T^i)\}_{i=1}^n$, Trained RNN $\widehat{f} = (\widehat{f}_t)_{t=1}^T$ with $\widehat{f}_t = \widehat{W}^o \widehat{h}_t + \widehat{b}^o$, $\widehat{h}_t = \sigma(\widehat{W}^h \widehat{h}_{t-1} + \widehat{W}^i x_t + \widehat{b}^{hi})$, Number of hidden nodes $m$ for trained RNN $\widehat{f}$, Number of hidden nodes $m^\sharp \leq m$ for returned compressed RNN, Regularization parameter $\tau \in \mathbb{R}_+^{m^\sharp}$, Coefficients $\theta_1, \theta_2, \theta_3 \in [0,1]$ with $\theta_1 + \theta_2 + \theta_3 = 1$.
1: Minimize $\left\{\theta_1 L_\tau^{(A)}(J) + \theta_2 L_\tau^{(B,o)}(J) + \theta_3 L_\tau^{(B,h)}(J)\right\}$ for index $J \subset [m]$ with $|J| = m^\sharp$ by the greedy algorithm where $L_\tau^{(A)}(J)$, $L_\tau^{(B,o)}(J)$, and $L_\tau^{(B,h)}(J)$ compute (3.4), (3.7), and (3.8), respectively.
2: Obtain optimal $J^\sharp$.
3: Compute $\widehat{A}_{J^\sharp}$ by (3.2).
4: Set $W_{J^\sharp}^{\sharp o} := \widehat{W}^o \widehat{A}_{J^\sharp}$, $W_{J^\sharp}^{\sharp h} := \widehat{W}^h_{J^\sharp,[m]} \widehat{A}_{J^\sharp}$, $W_{J^\sharp}^{\sharp i} := \widehat{W}^i_{J^\sharp,[d_x]}$, $b_{J^\sharp}^{\sharp hi} := \widehat{b}^{hi}_{J^\sharp}$, $b_{J^\sharp}^{\sharp o} := \widehat{b}^o$.
5: return Compressed RNN $f_{J^\sharp}^\sharp = (f_{J^\sharp,t}^\sharp)_{t=1}^T$ with $f_{J^\sharp,t}^\sharp = W_{J^\sharp}^{\sharp o} h_{J^\sharp,t}^\sharp + b_{J^\sharp}^{\sharp o}$ and $h_{J^\sharp,t}^\sharp = \sigma(W_{J^\sharp}^{\sharp h} h_{J^\sharp,t-1}^\sharp + W_{J^\sharp}^{\sharp i} x_t + b_{J^\sharp}^{\sharp hi})$.
```

Remark 3.1. In the case of the regularization parameter $\tau = 0$, spectral pruning can be applied, but the following points must be noted. In this case, the uniqueness of the minimization problem of $\|\phi - A\phi_J\|_{n,T}^2$ with respect to A does not generally hold (i.e., there might be several reconstruction matrices). One of the solutions is $\widehat{A}_J = \widehat{\Sigma}_{[m],J}\widehat{\Sigma}_{J,J}^\dagger$, which is the limit of (3.2) as $\tau \to 0$, where $\widehat{\Sigma}_{J,J}^\dagger$ is the pseudo-inverse of $\widehat{\Sigma}_{J,J}$. It should be noted that $\widehat{\Sigma}_{J,J}^\dagger$ coincides with the usual inverse $\widehat{\Sigma}_{J,J}^{-1}$, when $m^\sharp$ is smaller than or equal to the rank of the covariance matrix $\widehat{\Sigma}$.

Remark 3.2. We consider the case of the regularization parameter $\tau = 0$ and $m^\sharp \geq m_{\mathrm{nzr}}$, where m_{nzr} denotes the number of non-zero rows of $\widehat{\Sigma}$. Here, we would like to remark on the relation between m_{nzr} and pruning. Let J_{nzr} be the index set such that $[m] \setminus J_{\mathrm{nzr}}$ corresponds to zero rows of $\widehat{\Sigma}$. Then, by the definition (3.3) of $\widehat{\Sigma}$, we have for $i = 1, \cdots, n$, $t = 1, \cdots, T$, $v \in [m] \setminus J_{\mathrm{nzr}}$

$$\phi_v(x_t^i, \widehat{h}_{t-1}^i) = 0,$$

which implies that $\widetilde{A}_{J_{\mathrm{nzr}}} = I_{[m],J_{\mathrm{nzr}}}$ is a trivial solution of the minimization problem because $\|\phi - \widetilde{A}_{J_{\mathrm{nzr}}}\phi_{J_{\mathrm{nzr}}}\|_{n,T}^2 = 0$. Here, $I_{[m],J_{\mathrm{nzr}}}$ is the submatrix of the identity matrix corresponding to the index sets $[m]$ and J_{nzr}. If we choose $\widetilde{A}_{J_{\mathrm{nzr}}} = I_{[m],J_{\mathrm{nzr}}}$ as the reconstruction matrix, then the trivial compressed weights can be obtained by simply removing the columns corresponding to $[m] \backslash J_{\mathrm{nzr}}$, i.e., $W^{\sharp o} := \widehat{W}^o \widetilde{A}_{J_{\mathrm{nzr}}} = \widehat{W}^o_{[m],J_{\mathrm{nzr}}}$ and $W^{\sharp h} := \widehat{W}^h_{J_{\mathrm{nzr}},[m]} \widetilde{A}_{J_{\mathrm{nzr}}} = \widehat{W}^h_{J_{\mathrm{nzr}},J_{\mathrm{nzr}}}$, and its network $f_{J_{\mathrm{nzr}}}^\sharp$ coincides with the trained network $\widehat{f}$ for training data, i.e., for $i = 1, \cdots, n$, $t = 1, \cdots, T$

$$f_{J_{\mathrm{nzr}},t}^\sharp(X_t^i) = \widehat{f}_t(X_t^i)$$

which means that the trained RNN is compressed to size $m^\sharp$ without degradation. On the other hand, in the case of $m^\sharp < m_{\mathrm{nzr}}$, $\widetilde{A}_J = I_{[m],J}$ is not a solution of the minimization problem for any choice of the index J, which means that the compressed network using $\widehat{A}_J = \widehat{\Sigma}_{[m],J}\widehat{\Sigma}_{J,J}^\dagger$ is closer to the trained network than that using $\widetilde{A}_J = I_{[m],J}$. Therefore, spectral pruning essentially contributes to compression when $m^\sharp < m_{\mathrm{nzr}}$.

4 Generalization Error Bounds for Compressed RNNs

In this section, we discuss the generalization error bounds for compressed RNNs. In Subsection 4.1, the error bounds for general compressed RNNs are evaluated to explain the reason for deriving spectral pruning discussed in Section 3 in the error bound term. In

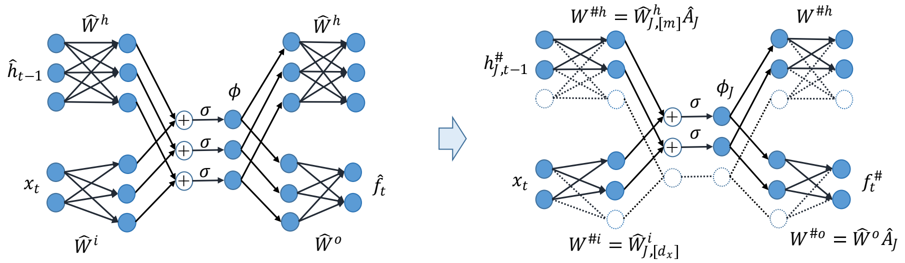


Figure 1: Spectral pruning for RNN

Subsection 4.2, the error bounds for RNNs compressed with spectral pruning are evaluated.

4.1 Error bound for general compressed RNNs

Let (X_T^i, Y_T^i) be the training data generated independently identically from the true distribution P_T, and let $f^\sharp$ be a general compressed RNN, and assume that it belongs to the following function space:

$$\begin{aligned}
&\mathcal{F}_T^\sharp = \mathcal{F}_T^\sharp(R_o, R_h, R_i, R_o^b, R_{hi}^b)\\
&\quad := \Big\{ f^\sharp \,\Big|\, f^\sharp(X_T) = (f_t^\sharp(X_t))_{t=1}^T,\\
&\quad f_t^\sharp(X_t) = (W^{\sharp o}\sigma(\cdot) + b^{\sharp o}) \circ (W^{\sharp h}\sigma(\cdot) + W^{\sharp i}x_t + b^{\sharp hi}) \circ\\
&\quad \cdots \circ (W^{\sharp h}\sigma(\cdot) + W^{\sharp i}x_2 + b^{\sharp hi}) \circ (W^{\sharp i}x_1 + b^{\sharp hi})\\
&\quad \text{for } X_T \in \mathrm{supp}(P_{X_T}),\ \left\|W^{\sharp o}\right\|_F \le R_o,\ \left\|W^{\sharp h}\right\|_F \le R_h,\\
&\quad \left\|W^{\sharp i}\right\|_F \le R_i,\ \left\|b^{\sharp o}\right\|_2 \le R_o^b,\ \left\|b^{\sharp hi}\right\|_2 \le R_{hi}^b \Big\},
\end{aligned}$$

where P_{X_T} is the marginal distribution of P_T with respect to X_T, and R_o, R_h, R_i, R_o^b, R_{hi}^b are the upper bounds of the compressed weights $W^{\sharp o} \in \mathbb{R}^{d_y \times m^\sharp}$, $W^{\sharp h} \in \mathbb{R}^{m^\sharp \times m^\sharp}$, $W^{\sharp i} \in \mathbb{R}^{m^\sharp \times d_x}$, biases $b^{\sharp o} \in \mathbb{R}^{d_y}$, and $b^{\sharp hi} \in \mathbb{R}^{m^\sharp}$, respectively. Here, $\|\cdot\|_F$ denotes the Frobenius norm.

Assumption 4.1. *The following assumptions are made:* **(i)** *The marginal distribution P_{x_t} of P_T with respect to x_t is bounded, i.e., there exists a constant R_x independent of t such that $\|x_t\|_2 \le R_x$ for all $x_t \in \mathrm{supp}(P_{x_t})$ and $t = 1, \dots, T$.* **(ii)** *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies $\sigma(0) = 0$ and $|\sigma(t) - \sigma(s)| \le \rho_\sigma |t - s|$ for all $t, s \in \mathbb{R}$.*

Under these assumptions, we obtain the following approximation error bounds between the trained network $\widehat{f}$ and compressed networks $f^\sharp$.

Proposition 4.2. *Let Assumption 4.1 hold. Let $(X_T^1, Y_T^1), \dots, (X_T^n, Y_T^n)$ be sampled i.i.d. from the distribution P_T. Then, for all $f^\sharp \in \mathcal{F}_T^\sharp$ and $J \subset [m]$ with $|J| = m^\sharp$, we have*

$$\begin{aligned}
&\left\|\widehat{f} - f^\sharp\right\|_{n,T} \lesssim \left\|\widehat{W}^o\phi - W^{\sharp o}\phi_J\right\|_{n,T}\\
&+ \left\|\widehat{W}^h_{J,[m]}\phi - W^{\sharp h}\phi_J\right\|_{n,T} + \left\|\widehat{W}^i_{J,[d_x]} - W^{\sharp i}\right\|_{op} \quad (4.1)\\
&+ \left\|\widehat{b}^{hi}_J - b^{\sharp hi}\right\|_2 + \left\|\widehat{b}^o - b^{\sharp o}\right\|_2.
\end{aligned}$$

Here, $\lesssim$ implies that the left-hand side in (4.1) is bounded above by the right-hand side times a constant independent of the trained weights and biases $\widehat{W}$, $\widehat{b}$ and compressed weights and biases $W^\sharp$, $b^\sharp$. The proof is given by direct computation. For the exact statement and proof, see Appendix B.

Remark 4.3. Let $f_J^\sharp$ be the network compressed using the reconstruction matrix (see (iii) in Section 3). By applying Proposition 4.2 as $f^\sharp = f_J^\sharp$, we obtain

$$\begin{aligned}
&\left\|\widehat{f} - f_J^\sharp\right\|_{n,T}^2 \lesssim \underbrace{\left\|\widehat{W}^o\phi - \widehat{W}^o\widehat{A}_J\phi_J\right\|_{n,T}^2 + \left\|\widehat{W}^o\widehat{A}_J\right\|_\tau^2}_{=L_\tau^{(B,o)}(J)}\\
&+ \underbrace{\left\|\widehat{W}^h_{J,[m]}\phi - \widehat{W}^h_{J,[m]}\widehat{A}_J\phi_J\right\|_{n,T}^2 + \left\|\widehat{W}^h_{J,[m]}\widehat{A}_J\right\|_\tau^2}_{=L_\tau^{(B,h)}(J)}\\
&\le \left(\left\|\widehat{W}^o\right\|_F^2 + \left\|\widehat{W}^h_{J,[m]}\right\|_F^2\right) \underbrace{\left(\left\|\phi - \widehat{A}_J\phi_J\right\|_{n,T}^2 + \left\|\widehat{A}_J\right\|_\tau^2\right)}_{=L_\tau^{(A)}(J)},
\end{aligned}$$

(4.2)

i.e., the approximation error is bounded by the input information loss.

For the RNN $f = (f_t)_{t=1}^T$, the training error with respect to the j-th component of the output is defined as

$$\widehat{\Psi}_j(f) := \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \psi(y_{t,j}^i, f_t(X_t^i)_j),$$

where $X_t = (x_t)_{t=1}^t$ and $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function. The generalization error with respect to the j-th component of the output is defined as

$$\Psi_j(f) := E\left[\frac{1}{T}\sum_{t=1}^T \psi(y_{t,j}, f_t(X_t)_j)\right],$$

where the expectation is taken with respect to $(X_T, Y_T) \sim P_T$.

Assumption 4.4. *The following assumptions are made:* **(i)** *The loss function $\psi(y_{t,j}, 0)$ is bounded, i.e., there exists a constant R_y such that $|\psi(y_{t,j}, 0)| \le R_y$ for all $y_{t,j} \in \operatorname{supp}(P_{y_{t,j}})$, $t = 1, \dots, T$, $j = 1, \dots, d_y$.* **(ii)** *ψ is ρ_ψ-Lipschitz continuous, i.e., $|\psi(y, f) - \psi(y, g)| \le \rho_\psi |f - g|$ for all $y, f, g \in \mathbb{R}$.*

We obtain the following generalization error bound for $f^\sharp \in \mathcal{F}_T^\sharp(R_o, R_h, R_i, R_o^b, R_{hi}^b)$.

Theorem 4.5. *Let Assumptions 4.1 and 4.4 hold, and let $(X_T^1, Y_T^1), \dots, (X_T^n, Y_T^n)$ be sampled i.i.d. from the distribution P_T. Then, for any $\delta \ge \log 2$, we have the following inequality with probability greater than $1 - 2e^{-\delta}$:*

$$\begin{aligned}\Psi_j(f^\sharp) &\lesssim \widehat{\Psi}_j(\widehat{f}) + \Big\{ \big\|\widehat{W}^o\phi - W^{\sharp o}\phi_J\big\|_{n,T} \\ &+ \big\|\widehat{W}^h_{J,[m]}\phi - W^{\sharp h}\phi_J\big\|_{n,T} + \big\|\widehat{W}^i_{J,[d_x]} - W^{\sharp i}\big\|_{op} \\ &+ \big\|\widehat{b}^{hi}_J - b^{\sharp hi}\big\|_2 + \big\|\widehat{b}^o - b^{\sharp o}\big\|_2 \Big\} + \frac{1}{\sqrt{n}}(m^\sharp)^{\frac{5}{4}} R_{\infty,T}^{1/2},\end{aligned} \tag{4.3}$$

for $j = 1, \dots, d_y$ and for all $J \subset [m]$ with $|J| = m^\sharp$, and $f^\sharp \in \mathcal{F}_T^\sharp$, where $R_{\infty,t}$ is defined by

$$R_{\infty,t} := R_o\rho_\sigma(R_iR_x + R^b_{hi})\left(\sum_{l=1}^{t}(R_h\rho_\sigma)^{l-1}\right) + R^b_o.$$

Here, $\lesssim$ implies that the left-hand side in (4.3) is bounded above by the right-hand side times a constant independent of the trained weights and biases $\widehat{W}$, $\widehat{b}$, compressed weights and biases $W^\sharp$, $b^\sharp$, compressed number $m^\sharp$, and the number of samples n. We remark that some omitted constants blow up as increasing T, but they can be controlled by increasing sampling number n (see Theorem C.1). The idea behind the proof is that the generalization error is decomposed into the training, approximation, and estimation errors. The approximation and estimation errors are evaluated using Proposition 4.2 and the estimation of the *Rademacher complexity*, respectively. For the exact statement and proof, see Appendix C.

The second term in (4.3) is the approximation error bound between $\widehat{f}$ and $f^\sharp$ regarded as the *bias*, which is given by Proposition 4.2, while the third term is the estimation error bound regarded as the *variance*. It can be observed that minimizing the terms $\|\widehat{W}^o\phi - W^{\sharp o}\phi_J\|_{n,T}$ and $\|\widehat{W}^h_{J,[m]}\phi - W^{\sharp h}\phi_J\|_{n,T}$ with respect to $W^{\sharp o}$ and $W^{\sharp h}$ is equivalent to the output information losses (3.5) and (3.6) with $\tau = 0$, respectively, which means that (iii) in Section 3 with $\tau = 0$ constructs the compressed RNN such that the bias term becomes smaller. Considering $\tau \neq 0$ prevents the blow up of $\|W^{\sharp o}\|_F$ and $\|W^{\sharp h}\|_F$, which means that the regularization parameter τ plays an important role in preventing the blow up of the variance term because $R_{\infty,T}$ in the variance term includes the upper bounds R_o and R_h of $\|W^{\sharp o}\|_F$ and $\|W^{\sharp h}\|_F$. Therefore, (iii) with $\tau \neq 0$ constructs the compressed RNN such that the generalization error bound becomes smaller. In addition, selecting an optimal J for minimizing the information losses (see (iv) in Section 3) further decreases the error bound.

4.2 Error bound for RNNs compressed with spectral pruning

Next, we evaluate the generalization error bounds for the RNN $f^\sharp_J$ compressed using the reconstruction matrix (see (iii) in Section 3). We define *degrees of freedom* $\widehat{N}(\lambda)$ by

$$\widehat{N}(\lambda) := \operatorname{Tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}\big] = \sum_{j=1}^{m} \frac{\widehat{\mu}_j}{\widehat{\mu}_j + \lambda},$$

where $\widehat{\mu}_j$ is an eigenvalue of $\widehat{\Sigma}$. It represents the intrinsic dimensionality of a model (Mallows, 2000; Caponnetto and De Vito, 2007). Throughout this subsection, the regularization parameter $\tau \in \mathbb{R}^{m^\sharp}_+$ is chosen as $\tau = \lambda m^\sharp \tau'$, where $\lambda > 0$ satisfies

$$m^\sharp \ge 5\widehat{N}(\lambda)\log(16\widehat{N}(\lambda)/\widetilde{\delta}), \tag{4.4}$$

for a prespecified $\widetilde{\delta} \in (0, 1/2)$. Here, $\tau' = (\tau'_j)_{j\in J} \in \mathbb{R}^{m^\sharp}$ is the *leverage score* defined by for $k \in [m]$

$$\tau'_k := \frac{1}{\widehat{N}(\lambda)}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}\big]_{k,k} = \frac{1}{\widehat{N}(\lambda)}\sum_{j=1}^{m} U^2_{k,j}\frac{\widehat{\mu}_j}{\widehat{\mu}_j + \lambda}, \tag{4.5}$$

where $U = (U_{k,j})_{k,j}$ is the orthogonal matrix that diagonalizes $\widehat{\Sigma}$, i.e., $\widehat{\Sigma} = U\operatorname{diag}\{\widehat{\mu}_1, \dots, \widehat{\mu}_m\}U^T$. The leverage score includes the information of the eigenvalues and eigenvectors of $\widehat{\Sigma}$, and indicates that the large components correspond to the important nodes from the viewpoint of the spectral information of $\widehat{\Sigma}$. Let q be the probability measure on $[m]$ defined by

$$q(v) := \tau'_v \quad \text{for } v \in [m]. \tag{4.6}$$

Proposition 4.6. *Let $v_1, \dots, v_{m^\sharp}$ be sampled i.i.d. from the distribution q in (4.6), and $J = \{v_1, \dots, v_{m^\sharp}\}$. Then, for any $\widetilde{\delta} \in (0, 1/2)$ and $\lambda > 0$ satisfying (4.4), we have the following inequality with probability greater than $1 - \widetilde{\delta}$:*

$$L^{(A)}_\tau(J) \le 4\lambda. \tag{4.7}$$

The proof is given in Appendix E. In the proof, we essentially refer to previous work (Bach, 2017). Combining (4.2) and (4.7), we conclude that

$$\left\|\widehat{f} - f_J^\sharp\right\|_{n,T}^2 \lesssim \lambda. \tag{4.8}$$

It can be observed that the approximation error bound (4.8) is controlled by the degrees of freedom. If the eigenvalues of $\widehat{\Sigma}$ rapidly decrease, then $\widehat{N}(\lambda)$ is a rapidly decreasing function as λ is large. Therefore, in that case, we can choose a smaller λ even when $m^\sharp$ is fixed. We will numerically study the relationship between the eigenvalue distribution and the input information loss in Section 5.1.

We make the following additional assumption.

Assumption 4.7. *Assume that the upper bounds for the trained weights and biases are given by* $\|\widehat{W}^o\|_F \leq \widehat{R}_o$, $\|\widehat{W}^h\|_F \leq \widehat{R}_h$, $\|\widehat{W}^i\|_F \leq \widehat{R}_i$, $\|\widehat{b}^{hi}\|_2 \leq \widehat{R}_{hi}^b$, *and* $\|\widehat{b}^o\|_2 \leq \widehat{R}_o^b$.

We have the following generalization error bound.

Theorem 4.8. *Let Assumptions 4.1, 4.4, and 4.7 hold, and let* $(X_T^1, Y_T^1), \ldots, (X_T^n, Y_T^n)$ *and* $v_1, \ldots, v_{m^\sharp}$ *be sampled i.i.d. from the distributions* P_T *and* q *in* (4.6)*, respectively. Let* $J = \{v_1, \ldots, v_{m^\sharp}\}$*. Then, for any* $\delta \geq \log 2$ *and* $\widetilde{\delta} \in (0, 1/2)$*, we have the following inequality with probability greater than* $(1 - 2e^{-\delta})\widetilde{\delta}$*:*

$$\Psi_j(f_J^\sharp) \lesssim \widehat{\Psi}_j(\widehat{f}) + \sqrt{\lambda} + \frac{1}{\sqrt{n}}(m^\sharp)^{\frac{5}{4}}, \tag{4.9}$$

for $j = 1, \ldots, d_y$ *and for all* $\lambda > 0$ *satisfying* (4.4).

Here, $\lesssim$ implies that the left-hand side in (4.9) is bounded above by the right-hand side times a constant independent of λ, $m^\sharp$, and n. We remark that some omitted constants blow up as increasing T, but they can be controlled by increasing sampling number n (see Theorem F.1). The proof is given by the combination of applying Theorem 4.5 as $f^\sharp = f_J^\sharp$ and using Proposition 4.6. For the exact statement and proof, see Appendix F. It can be observed that in (4.9), a bias-variance tradeoff relationship exists with respect to $m^\sharp$. When $m^\sharp$ is large, λ can be chosen smaller in condition (4.4), which implies that the bias term (the second term in (4.9)) becomes smaller, but the variance term (the third term in (4.9)) becomes larger. In contrast, the bias becomes larger and the variance becomes smaller when $m^\sharp$ is small. Further remarks on Theorem 4.8 are given in Appendix G.

5 Numerical Experiments

In this section, numerical experiments are detailed to demonstrate our theoretical results and show the effectiveness of spectral pruning compared with existing methods. In Sections 5.1 and 5.2, we select the pixel-MNIST as our task and employ the IRNN, which is an RNN that uses the ReLU as the activation function and initializes weights as the identity matrix and biases to zero (see (Le et al., 2015)). In Section 5.3, we select the PTB (Marcus et al., 1993) and employ the RNNLM whose RNN layer is orthodox Elman-type. For RNN training details, see Appendix H. We choose parameters $\theta_1 = 1$, $\theta_2 = \theta_3 = 0$ in (iv) of Section 3, i.e., we minimize only the input information loss. This choice is not so problematic because the bound of output information loss automatically becomes smaller with minimizing the input one (see Remark 4.3). We choose the regularization parameter $\tau = 0$, where this choice regards $\widehat{f}$ as a well-trained network and gives priority to minimizing the approximation error between $\widehat{f}$ and $f_J^\sharp$ (see below Theorem4.5).

5.1 Eigenvalue distribution and information loss

First, we numerically study the relationship between the eigenvalue distribution and the information loss. Figure 2a shows the eigenvalue distribution of the covariance matrix $\widehat{\Sigma}$ with 128 hidden nodes, which are sorted in decreasing order. In this experiment, almost half of the eigenvalues are zero, which cannot be visualized in the figure. Figure 2b shows the input information loss $L_0^{(A)}(J)$ versus the compressed number $m^\sharp$. The information losses vanish when $m^\sharp > m_{\text{nzr}}$ (see Remark 3.2). The blue and pink curves correspond to MNIST[1] and FashionMNIST[2], respectively. It can be observed that the eigenvalues for MNIST decrease more rapidly than those for FashionMNIST, and the information losses for MNIST decrease more rapidly than those for FashionMNIST. This phenomenon coincides with the interpretation on Proposition 4.6 (see the discussion below (4.8)).

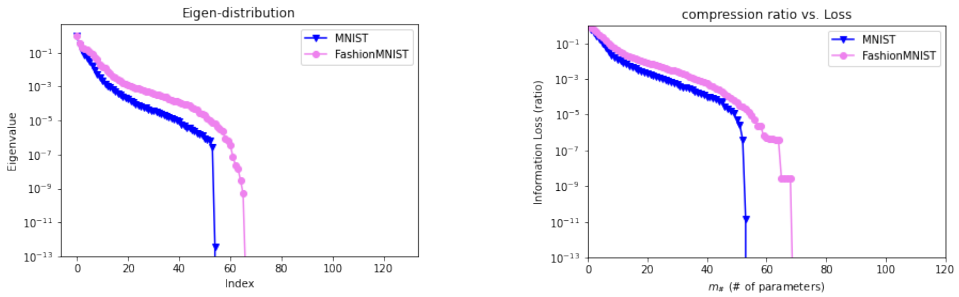
5.2 Pixel-MNIST (IRNN)

We compare spectral pruning with other pruning methods in pixel-MNIST (IRNN). Table 1 summarizes the accuracies and the number of weight parameters for different pruning methods. We consider one-third compression in the hidden state, i.e., for the node pruning, 128 hidden nodes were compressed to 42 nodes, while for weight pruning, $128^2 (= 16384)$ hidden weights were compressed to $42^2 (= 1764)$ weights.

"Baseline(128)" and "Baseline(42)" represent direct training (not pruning) with 128 and 42 hidden nodes, respectively. "Spectral w/ rec.(ours)" represents spec-

[1] http://yann.lecun.com/exdb/mnist/

[2] https://github.com/zalandoresearch/fashion-mnist



(a) Eigenvalue distribution for $\widehat{\Sigma}$

(b) Input information loss vs. $m^\sharp$

Figure 2: Relationship between the eigenvalue distribution and the input information loss

Table 1: Pixel-MNIST (IRNN)

Method	Accuracy[%] (std)	Finetuned Accuracy[%](std)	# input -hidden	# hidden -hidden	# hidden -out	total
Baseline(128)	96.80 (0.23)	-	128	16384	1280	17792
Baseline(42)	93.35 (0.75)	-	42	1764	420	2226
Spectral w/ rec.(ours)	**92.61 (2.46)**	**97.08 (0.16)**	**42**	**1764**	**420**	**2226**
Spectral w/o rec.	83.60 (8.24)	-	42	1764	420	2226
Random w/ rec.	34.72 (32.47)	-	42	1764	420	2226
Random w/o rec.	23.13 (16.09)	-	42	1764	420	2226
Random Weight	10.35 (1.38)	-	128	1764	1280	3172
Magnitude-based Weight	11.06 (0.70)	94.41 (3.02)	128	1764	1280	3172
Column Sparsification	84.80 (7.29)	-	128	5376	1280	6784
Low Rank Factorization	9.65 (3.85)	-	128	10752	1280	12160

tral pruning with the reconstruction matrix (i.e., the compressed weight is chosen as $W^{\sharp h} = \widehat{W}^h_{J,[m]}\widehat{A}_J$ with the optimal J with respect to (3.9)), while "Spectral w/o rec." represents spectral pruning without the reconstruction matrix (i.e., $W^{\sharp h} = \widehat{W}^h_{J,J}$ with the optimal J with respect to (3.9)), which idea is based on (Luo et al., 2017). "Random w/ rec." represents random node pruning with the reconstruction matrix (i.e., $W^{\sharp h} = \widehat{W}^h_{J,[m]}\widehat{A}_J$, where J is randomly chosen), while "Random w/o rec." represents random node pruning without the reconstruction matrix (i.e., $W^{\sharp h} = \widehat{W}^h_{J,J}$, where J is randomly chosen). "Random Weight" represents random weight pruning. For the reason that we compare with random pruning, see the introduction of (Zhang and Stadie, 2019). "Magnitude-based Weight" represents magnitude-based weight pruning based on (Narang et al., 2017a). "Column Sparsification" represents the magnitude-based column sparsification during training based on (Wang et al., 2019). "Low Rank Factorization" represents low rank factorization which truncates small singular values of trained weights based on (Prabhavalkar et al., 2016). "Accuracy[%](std)" and "Finetuned Accuracy[%](std)" represent their mean (standard deviation) of accuracy before and after fine-tuning, respectively. "# input-hidden", "# hidden-hidden", and "# hidden-out" represent the number of input-to-hidden, hidden-to-hidden, and hidden-to-output weight parameters, respectively. "total" represents their sum. For detailed procedures of training, pruning, and fine-tuning, see Appendix H.

We demonstrate that spectral pruning significantly outperforms other pruning methods. The reason spectral pruning can compress with small degradation is that the covariance matrix $\widehat{\Sigma}$ has a small number of non-zero rows (we observed around 50 non-zero rows). For the detail of non-zero rows, see Remark 3.2. Our method of fine-tuning outperforms "Baseline(42)", which means that the spectral pruning gets benefits from over-parameterization (Chang et al., 2020; Zhang et al., 2021). Since the magnitude-based weight pruning is the method to require the fine-tuning (e.g., see (Narang et al., 2017a)), we have also compared our method with the magnitude-based weight pruning with fine-tuning, and observed that our method outperforms the magnitude-based weight pruning as well. We also remark that our method of fine-tuning overcomes "Baseline(128)".

Table 2: PTB (RNNLM)

Method	Perplexity (std)	Finetuned Perplexity (std)	# input -hidden	# hidden -hidden	# hidden -out	total
Baseline(128)	114.66 (0.35)	-	1270016	16384	1270016	2556416
Baseline(42)	145.85 (0.74)	132.46 (0.74)	416724	1764	416724	835212
Spectral w/ rec.(ours)	**207.63 (2.19)**	**124.26 (0.39)**	**416724**	**1764**	**416724**	**835212**
Spectral w/o rec.	433.99 (10.64)	-	416724	1764	416724	835212
Random w/ rec.	243.76 (9.46)	-	416724	1764	416724	835212
Random w/o rec.	492.06 (22.40)	-	416724	1764	416724	835212
Random Weight	203.41 (2.02)	-	1270016	1764	1270016	2541796
Magnitude-based Weight	168.57 (2.57)	115.65 (0.31)	1270016	1764	1270016	2541796
Magnitude-based Weight ♢	201.41 (3.60)	126.20 (0.28)	416724	1764	416724	835212
Column Sparsification	128.98 (0.52)	-	1270016	5376	1270016	2545408
Low Rank Factorization	126.24 (1.79)	-	1270016	10752	1270016	2550784

5.3 PTB (RNNLM)

We compare spectral pruning with other pruning methods in the PTB (RNNLM). Table 2 summarizes the perplexity and the number of weight parameters for different pruning methods. As in Section 5.2, we consider one-third compression in the hidden state, and how to represent "Method" is the same as Table 1 except for "Magnitude-based Weight ♢", which represents the magnitude-based weight pruning for not only hidden-to-hidden weights but also input-to-hidden and hidden-to-out weights so that the number of resultant weight parameters is the same as Spectral w/ rec.(ours).

We demonstrate that our method of fine-tuning outperforms other pruning methods except for magnitude-based weight pruning. Even though "Low Rank Factorization" retains large number of weight parameters, its perplexity is slightly worse than our method of fine-tuning. On the other hand, our method of fine-tuning can not outperform "Magnitude-based Weight", but it can slightly under the condition of the same number of weight parameters. We also remark that our method of fine-tuning overcomes "Baseline(42)", although it does not overcome "Baseline(128)".

Therefore, we conclude that spectral pruning works well in Elman-RNN, especially in IRNN.

Limitations and Future Works

In this paper, we show the generalization error bounds for compressed RNNs, but there are some limitations. The bounds are obtained under the assumption that the length T of time series data is fixed, and they also blow up as T goes to infinity. In the future, we will analyse the generalization error without the assumption and improve the bounds to sharper ones. The other limitation is that we only treat Elman-RNNs. It would be interesting to extend our work to the long short-term memory (LSTM), which is more sophisticated and commonly used architecture than Elman-RNNs, although Elman-RNNs continue to be used today in applications such as edge computing devices. The properties of LSTMs are different from those of Elman-RNNs in that LSTMs have gated architectures including product operations, which might require more complicated analysis of the generalization error bounds as compared to Elman-RNNs. Hence, the investigation of spectral pruning for LSTMs is beyond the scope of this study and will be the focus of future work.

Acknowledgements

The authors are grateful to Professor Taiji Suzuki for useful discussions and comments on our work. The first author was supported by Grant-in-Aid for JSPS Fellows (No.21J00119), Japan Society for the Promotion of Science.

References

Nil-Jana Akpinar, Bernhard Kratzwald, and Stefan Feuerriegel. Sample complexity bounds for recurrent neural networks with application to combinatorial graph problems. *arXiv preprint arXiv:1901.10289*, 2019.

Md Zahangir Alom, Adam T Moody, Naoya Maruyama, Brian C Van Essen, and Tarek M Taha. Effective quantization approaches for recurrent neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1): 714–751, 2017.

Peter Bartlett, Dylan J Foster, and Matus Telgarsky.

Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368, 2007.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *arXiv preprint arXiv:2012.08749*, 2020.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. *arXiv preprint arXiv:1910.12947*, 2019.

Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando De Freitas. Predicting parameters in deep learning. *arXiv preprint arXiv:1306.0543*, 2013.

Jesse Dodge, Roy Schwartz, Hao Peng, and Noah A Smith. Rnn architecture learning with sparse regularization. *arXiv preprint arXiv:1909.03011*, 2019.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/076a0c97d09cf1a0ec3e19c7f2529f2b-Paper.pdf`.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40. Cambridge University Press, 2015.

Mark S Gockenbach. *Linear inverse problems and Tikhonov regularization*, volume 32. American Mathematical Soc., 2016.

Artem M Grachev, Dmitry I Ignatov, and Andrey V Savchenko. Compression of recurrent neural networks for efficient language modeling. *Applied Soft Computing*, 79:354–362, 2019.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*, 2016.

Boris Joseph Joukovsky, Tanmoy Mukherjee, Nikos Deligiannis, et al. Generalization error bounds for deep unfolding rnns. In *Proceedings of Machine Learning Research*. Journal of Machine Learning Research, 2021.

Markus Kliegl, Siddharth Goyal, Kexin Zhao, Kavya Srinet, and Mohammad Shoeybi. Trace norm regularization and faster inference for embedded speech recognition rnns. *arXiv preprint arXiv:1710.09026*, 2017.

Quoc V Le, Navdeep Jaitly, and Geoffrey E Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

Shiwei Liu, Decebal Constantin Mocanu, Yulong Pei, and Mykola Pechenizkiy. Selfish sparse rnn training. *arXiv preprint arXiv:2101.09048*, 2021a.

Shiwei Liu, Iftitahu Ni'mah, Vlado Menkovski, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Efficient and effective training of sparse recurrent neural networks. *Neural Computing and Applications*, pages 1–12, 2021b.

Xuan Liu, Di Cao, and Kai Yu. Binarized lstm language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2113–2121, 2018.

Ekaterina Lobacheva, Nadezhda Chirkova, and Dmitry Vetrov. Bayesian sparsification of recurrent neural networks. *arXiv preprint arXiv:1708.00077*, 2017.

Ekaterina Lobacheva, Nadezhda Chirkova, Alexander Markovich, and Dmitry Vetrov. Structured sparsification of gated recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4989–4996, 2020.

Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pages 5058–5066, 2017.

Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. 1993.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, pages 2498–2507. PMLR, 2017.

Sharan Narang, Greg Diamos, Shubho Sengupta, and Erich Elsen. Exploring sparsity in recurrent neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France,* *April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017a. URL https://openreview.net/forum?id=BylSPv9gx.

Sharan Narang, Eric Undersander, and Gregory Diamos. Block-sparse recurrent neural networks. *arXiv preprint arXiv:1711.02782*, 2017b.

Angshuman Parashar, Minsoo Rhu, Anurag Mukkara, Antonio Puglielli, Rangharajan Venkatesan, Brucek Khailany, Joel Emer, Stephen W Keckler, and William J Dally. Scnn: An accelerator for compressed-sparse convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 45 (2):27–40, 2017.

Rohit Prabhavalkar, Ouais Alsharif, Antoine Bruguier, and Lan McGraw. On the compression of recurrent neural networks with an application to lvcsr acoustic modeling for embedded speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5970–5974. IEEE, 2016.

Yangyang Shi, Mei-Yuh Hwang, Xin Lei, and Haoyu Sheng. Knowledge distillation for recurrent neural network language modeling with trust regularization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7230–7234. IEEE, 2019.

Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2839–2846. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

Shijian Tang and Jiang Han. A pruning based method to learn both weights and connections for lstm. *Proceedings of the Advances in Neural Information Processing Systems, NIPS, Montreal, QC, Canada*, pages 7–12, 2015.

Zhiyuan Tang, Dong Wang, and Zhiyong Zhang. Recurrent neural network training with dark knowledge transfer. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5900–5904. IEEE, 2016.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Compressing recurrent neural network with tensor train. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 4451–4458. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966420. URL https://doi.org/10.1109/IJCNN.2017.7966420.

Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. Understanding generalization in recurrent neural networks. In *International Conference on Learning Representations*, 2019.

Shaorun Wang, Peng Lin, Ruihan Hu, Hao Wang, Jin He, Qijun Huang, and Sheng Chang. Acceleration of lstm with structured pruning method on fpga. *IEEE Access*, 7:62930–62937, 2019.

Liangjiang Wen, Xueyang Zhang, Haoli Bai, and Zenglin Xu. Structured pruning of recurrent neural networks through neuron selection. *Neural networks : the official journal of the International Neural Network Society*, 123:134–141, 2020.

Wei Wen, Yuxiong He, Samyam Rajbhandari, Minjia Zhang, Wenhan Wang, Fang Liu, Bin Hu, Yiran Chen, and Hai Li. Learning intrinsic sparse structures within long short-term memory. *arXiv preprint arXiv:1709.05027*, 2017.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Matthew Shunshi Zhang and Bradly Stadie. One-shot pruning of recurrent neural networks by jacobian spectrum evaluation. *arXiv preprint arXiv:1912.00120*, 2019.

Appendix

A Review of Spectral Pruning for DNNs

Let $D = \{(x^i, y^i)\}_{i=1}^n$ be training data, where $x^i \in \mathbb{R}^{d_x}$ is an input and $y^i \in \mathbb{R}^{d_y}$ is an output. The training data are independently identically distributed. To train the appropriate relationship between input and output, we consider DNNs f as

$$f(x) = (W^{(L)}\sigma(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(1)}x + b^{(1)}),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, $W^{(l)} \in \mathbb{R}^{m_{l+1} \times m_l}$ is a weight matrix, and $b^{(l)} \in \mathbb{R}^{m_{l+1}}$ is a bias. Let $\widehat{f}$ be a trained DNN obtained from the training data D, i.e.,

$$\widehat{f}(x) = (\widehat{W}^{(L)}\sigma(\cdot) + \widehat{b}^{(L)}) \circ \cdots \circ (\widehat{W}^{(1)}x + \widehat{b}^{(1)}).$$

We denote the input with respect to l-th layer by

$$\phi^{(l)}(x) = \sigma \circ (\widehat{W}^{(l-1)}\sigma(\cdot) + \widehat{b}^{(l-1)}) \circ \cdots \circ (\widehat{W}^{(1)}x + \widehat{b}^{(1)}).$$

Let $J^{(l)} \subset [m_l]$ be an index set with $|J^{(l)}| = m_l^\sharp$, where $[m_l] := \{1, \ldots, m_l\}$ and $m_l^\sharp \in \mathbb{N}$ is the number of nodes of the l-th layer of the compressed DNN $f^\sharp$ with $m_l^\sharp \leq m_l$. Let $\phi_{J^{(l)}}^{(l)}(x) = (\phi_j^{(l)}(x))_{j \in J^{(l)}}$ be a subvector of $\phi^{(l)}(x)$ corresponding to the index set $J^{(l)}$, where $\phi_j^{(l)}(x)$ is the j-th components of the vector $\phi^{(l)}(x)$.

(i) Input information loss. The input information loss is defined by

$$L_\tau^{(A,l)}(J^{(l)}) := \min_{A \in \mathbb{R}^{m_l \times m_l^\sharp}} \left\{ \left\| \phi^{(l)} - A\phi_{J^{(l)}}^{(l)} \right\|_n^2 + \left\| A \right\|_\tau^2 \right\}, \tag{A.1}$$

where $\| \cdot \|_n$ is the empirical L^2-norm with respect to n, i.e.,

$$\left\| \phi^{(l)} - A\phi_{J^{(l)}}^{(l)} \right\|_n^2 := \frac{1}{n} \sum_{i=1}^n \left\| \phi^{(l)}(x^i) - A\phi_{J^{(l)}}^{(l)}(x^i) \right\|_2^2, \tag{A.2}$$

where $\| \cdot \|_2$ is the Euclidean norm and $\|A\|_\tau^2 := \mathrm{Tr}\,[AI_\tau A^T]$ for a regularization parameter $\tau \in \mathbb{R}_+^{m_l^\sharp}$. Here, $\mathbb{R}_+^{m_l^\sharp} := \left\{ x \in \mathbb{R}^{m_l^\sharp} \mid x_j > 0,\ j = 1, \ldots, m_l^\sharp \right\}$ and $I_\tau := \mathrm{diag}(\tau)$. By the linear regularization theory, there exists a unique solution $\widehat{A}_{J^{(l)}}^{(l)} \in \mathbb{R}^{m_l \times m_l^\sharp}$ of the minimization problem of $\|\phi^{(l)} - A\phi_{J^{(l)}}^{(l)}\|_n^2 + \|A\|_\tau^2$, and it has the form

$$\widehat{A}_{J^{(l)}}^{(l)} = \widehat{\Sigma}_{[m_l],J^{(l)}}^{(l)} \left(\widehat{\Sigma}_{J^{(l)},J^{(l)}}^{(l)} + I_\tau \right)^{-1}, \tag{A.3}$$

where $\widehat{\Sigma}^{(l)}$ is the (noncentered) empirical covariance matrix of $\phi^{(l)}(x)$ with respect to n, i.e.,

$$\widehat{\Sigma}^{(l)} = \frac{1}{n} \sum_{i=1}^n \phi^{(l)}(x^i)\phi^{(l)}(x^i)^T,$$

and $\widehat{\Sigma}_{I,I'}^{(l)} = (\widehat{\Sigma}_{i,i'}^{(l)})_{i \in I, i' \in I'} \in \mathbb{R}^{K \times H}$ is the submatrix of $\widehat{\Sigma}^{(l)}$ corresponding to index sets $I, I' \subset [m]$ with $|I| = K$ and $|I'| = H$. By substituting the explicit formula (A.2) of the reconstruction matrix $\widehat{A}_{J^{(l)}}^{(l)}$ into (A.1), the input information loss is reformulated as

$$L_\tau^{(A,l)}(J^{(l)}) = \mathrm{Tr}\left[\widehat{\Sigma}^{(l)} - \widehat{\Sigma}_{[m_l],J^{(l)}}^{(l)} \left(\widehat{\Sigma}_{J^{(l)},J^{(l)}}^{(l)} + I_\tau \right)^{-1} \widehat{\Sigma}_{J^{(l)},[m_l]}^{(l)} \right]. \tag{A.4}$$

(ii) Output information loss. For any matrix $Z^{(l)} \in \mathbb{R}^{m \times m_l}$ with an output size $m \in \mathbb{N}$, we define the output information loss by

$$L_\tau^{(B,l)}(J^{(l)}) := \sum_{j=1}^m \min_{\beta \in \mathbb{R}^{m_l^\sharp}} \left\{ \left\| Z_{j,:}^{(l)}\phi^{(l)} - \beta^T \phi_{J^{(l)}}^{(l)} \right\|_n^2 + \left\| \beta^T \right\|_\tau^2 \right\}, \tag{A.5}$$

where $Z^{(l)}_{j,:}$ denotes the j-th row of the matrix $Z^{(l)}$. A typical situation is that $Z^{(l)} = \widehat{W}^{(l)}$. The minimization problem of $\|Z^{(l)}_{j,:}\phi^{(l)} - \beta^T \phi_{J^{(l)}}\|_n^2 + \|\beta^T\|_\tau^2$ has the unique solution

$$\widehat{\beta}^{(l)}_j = (Z^{(l)}_{j,:}\widehat{A}^{(l)}_{J^{(l)}})^T,$$

and by substituting it into (A.5), the output information loss is reformulated as

$$L^{(B,l)}_\tau(J^{(l)}) = \mathrm{Tr}\Big[Z^{(l)}\big(\widehat{\Sigma}^{(l)} - \widehat{\Sigma}^{(l)}_{[m],J^{(l)}}\big(\widehat{\Sigma}^{(l)}_{J^{(l)},J^{(l)}} + I_\tau\big)^{-1}\widehat{\Sigma}^{(l)}_{J^{(l)},[m]}\big)Z^{(l)T}\Big].$$

(iii) Compressed DNN by the reconstruction matrix. We construct the compressed DNN by

$$f^\sharp_{J^{(1:L)}}(x) = (W^{\sharp(L)}_{J^{(L)}}\sigma(\cdot) + b^{\sharp(L)}) \circ \cdots \circ (W^{\sharp(1)}_{J^{(1)}}x + b^{\sharp(1)}),$$

where $J^{(1:L)} = J^{(1)} \cup \cdots \cup J^{(L)}$, and $b^{\sharp(l)} = \widehat{b}^{(l)}$ and $W^{\sharp(l)}_{J^{(l)}}$ is the compressed weight as the multiplication of the trained weight $\widehat{W}^{(l)}_{J^{(l+1)},[m_l]}$ and the reconstruction matrix $\widehat{A}_{J^{(l)}}$, i.e.,

$$W^{\sharp(l)}_{J^{(l)}} := \widehat{W}^{(l)}_{J^{(l+1)},[m_l]}\widehat{A}_{J^{(l)}}. \tag{A.6}$$

.

(iv) Optimization. To select an appropriate index set $J^{(l)}$, we consider the following optimization problem that minimizes a convex combination of input and output information losses, i.e.,

$$\min_{J^{(l)} \subset [m_l] \ s.t. \ |J^{(l)}| = m^\sharp_l} \big\{\theta L^{(A,l)}_\tau(J^{(l)}) + (1-\theta)L^{(B,l)}_\tau(J^{(l)})\big\},$$

for $\theta \in [0,1]$, where $m^\sharp_l \in [m]$ is a prespecified number. We adapt the optimal index $J^{\sharp(1:L)}$ in the algorithm. We term this method as *spectral pruning.*

In (Suzuki et al., 2020), the generalization error bounds for compressed DNNs with the spectral pruning have been studied (see Theorems 1 and 2 in (Suzuki et al., 2020)), and the parameters θ, τ, and $Z^{(l)}$ are chosen such that its error bounds become smaller.

B Proof of Proposition 4.2

We restate Proposition 4.2 in an exact form as follows:

Proposition B.1. *Suppose that Assumption 4.1 holds. Let $\{(X^i_T, Y^i_T)\}_{i=1}^n$ be sampled i.i.d. from the distribution P_T. Then,*

$$\begin{aligned}\|\widehat{f} - f^\sharp\|_{n,T} \leq \sqrt{3}\Bigg\{&\big\|\widehat{W}^o\phi - W^{\sharp o}\phi_J\big\|_{n,T} + R_o\rho_\sigma \max\{1, (R_h\rho_\sigma)^{T-2}\}T\big\|\widehat{W}^h_{J,[m]}\phi - W^{\sharp h}\phi_J\big\|_{n,T} \\ &+ R_o\rho_\sigma\left(\sum_{t=1}^T (R_h\rho_\sigma)^{t-1}\right)\left(R_x\big\|\widehat{W}^i_{J,[d_x]} - W^{\sharp i}\big\|_{op} + \|\widehat{b}^{hi}_J - b^{\sharp hi}\|_2\right) + \big\|\widehat{b}^o - b^{\sharp o}\big\|_2\Bigg\},\end{aligned} \tag{B.1}$$

for all $f^\sharp \in \mathcal{F}^\sharp_T(R_o, R_h, R_i, R^b_o, R^b_{hi})$ and $J \subset [m]$ with $|J| = m^\sharp$.

Proof. Let $\widehat{f} = (\widehat{f}_t)_{t=1}^T$ be a trained RNN and $f^\sharp \in \mathcal{F}^\sharp_T(R_o, R_h, R_i, R^b_o, R^b_{hi})$. Let us define functions ϕ and $\phi^\sharp$ by

$$\begin{aligned}\phi(x,h) &:= \sigma(\widehat{W}^h h + \widehat{W}^i x + \widehat{b}^{hi}) \quad \text{for } x \in \mathbb{R}^{d_x},\ h \in \mathbb{R}^m, \\ \phi^\sharp(x,h^\sharp) &:= \sigma(W^{\sharp h}h^\sharp + W^{\sharp i}x + b^{\sharp hi}) \quad \text{for } x \in \mathbb{R}^{d_x},\ h^\sharp \in \mathbb{R}^{m^\sharp},\end{aligned} \tag{B.2}$$

and denote the hidden states by

$$\widehat{h}_t := \phi(x_t, \widehat{h}_{t-1}), \quad h^\sharp_t := \phi^\sharp(x_t, h^\sharp_{t-1}) \quad \text{for } t = 1, 2, \cdots, T. \tag{B.3}$$

If a training data $X_T^i = (x_t^i)_{t=1}^T$ is used as input, we denote its hidden state by

$$\widehat{h}_t^i := \phi(x_t^i, \widehat{h}_{t-1}^i), \quad h_t^{\sharp i} := \phi^{\sharp}(x_t^i, h_{t-1}^{\sharp i}),$$

and its outputs at time t by

$$\widehat{f}_t(X_t^i) = \widehat{W}^o \phi(x_t^i, \widehat{h}_{t-1}^i) + \widehat{b}^o, \quad f_t^{\sharp}(X_t^i) = W^{\sharp o} \phi^{\sharp}(x_t^i, h_{t-1}^{\sharp i}) + b^{\sharp o},$$

for $t = 1, 2, \ldots, T$. Then, we have

$$\begin{aligned} \big\|\widehat{f}_t(X_t^i) - f_t^{\sharp}(X_t^i)\big\|_2 \leq{}& \big\|\widehat{W}^o \phi(x_t^i, \widehat{h}_{t-1}^i) - W^{\sharp o} \phi_J(x_t^i, \widehat{h}_{t-1}^i)\big\|_2 \\ &+ \big\|W^{\sharp o} \phi_J(x_t^i, \widehat{h}_{t-1}^i) - W^{\sharp o} \phi^{\sharp}(x_t^i, h_{t-1}^{\sharp i})\big\|_2 + \big\|\widehat{b}^o - b^{\sharp o}\big\|_2. \end{aligned} \tag{B.4}$$

If we can prove that the second term of right-hand side in (B.4) is estimated as

$$\begin{aligned} &\big\|W^{\sharp o} \phi_J(x_t^i, \widehat{h}_{t-1}^i) - W^{\sharp o} \phi^{\sharp}(x_t^i, h_{t-1}^{\sharp i})\big\|_2 \\ &\quad \leq R_o \rho_\sigma \bigg\{ \max\big\{1, (R_h \rho_\sigma)^{t-2}\big\} \sum_{l=1}^{t-1} \big\|\widehat{W}_{J,[m]}^h \phi(x_{t-l}^i, \widehat{h}_{t-l-1}^i) - W^{\sharp h} \phi_J(x_{t-l}^i, \widehat{h}_{t-l-1}^i)\big\|_2 \\ &\qquad + \sum_{l=1}^{t} (R_h \rho_\sigma)^{l-1} \big(\big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op} \|x_{t-l+1}^i\|_2 + \big\|\widehat{b}_J^{hi} - b^{\sharp hi}\big\|_2\big) \bigg\}, \end{aligned} \tag{B.5}$$

then by using the inequalities (B.4) and $(\sum_{k=1}^K a_k)^2 \leq K \sum_{k=1}^K a_k^2$, we have

$$\begin{aligned} &\big\|\widehat{f}_t(X_t^i) - f_t^{\sharp}(X_t^i)\big\|_2^2 \leq 3 \bigg\{ \big\|\widehat{W}^o \phi(x_t^i, \widehat{h}_{t-1}^i) - W^{\sharp o} \phi_J(x_t^i, \widehat{h}_{t-1}^i)\big\|_2^2 \\ &+ \bigg(R_o \rho_\sigma \max\big\{1, (R_h \rho_\sigma)^{t-2}\big\} \sum_{l=1}^{t-1} \big\|\widehat{W}_{J,[m]}^h \phi(x_{t-l}^i, \widehat{h}_{t-l-1}^i) - W^{\sharp h} \phi_J(x_{t-l}^i, \widehat{h}_{t-l-1}^i)\big\|_2 \bigg)^2 \\ &+ \bigg(R_o \rho_\sigma \sum_{l=1}^{t} (R_h \rho_\sigma)^{l-1} \big(\big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op} \|x_{t-l+1}^i\|_2 + \big\|\widehat{b}_J^{hi} - b^{\sharp hi}\big\|_2\big) + \big\|\widehat{b}^o - b^{\sharp o}\big\|_2 \bigg)^2 \bigg\}. \end{aligned}$$

Hence, by taking the average over $i = 1, \ldots, n$ and $t = 1, \ldots, T$, and by using the inequality $\sum_{t=1}^T (\sum_{l=1}^t a_l)^2 \leq T^2 \sum_{t=1}^T a_t^2$, we obtain

$$\begin{aligned} &\big\|\widehat{f} - f^{\sharp}\big\|_{n,T}^2 = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \big\|\widehat{f}_t(X_t^i) - f_t^{\sharp}(X_t^i)\big\|_2^2 \\ &\leq 3 \Bigg\{ \underbrace{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \big\|\widehat{W}^o \phi(x_t^i, \widehat{h}_{t-1}^i) - W^{\sharp o} \phi_J(x_t^i, \widehat{h}_{t-1}^i)\big\|_2^2}_{=\|\widehat{W}^o \phi - W^{\sharp o} \phi_J\|_{n,T}^2} \\ &+ \big(R_o \rho_\sigma \max\big\{1, (R_h \rho_\sigma)^{T-2}\big\} T\big)^2 \underbrace{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \big\|\widehat{W}_{J,[m]}^h \phi(x_t^i, \widehat{h}_{t-1}^i) - W^{\sharp h} \phi_J(x_t^i, \widehat{h}_{t-1}^i)\big\|_2^2}_{=\|\widehat{W}_{J,[m]}^h \phi - W^{\sharp h} \phi_J\|_{n,T}^2} \\ &+ \bigg(R_o \rho_\sigma \bigg(\sum_{t=1}^T (R_h \rho_\sigma)^{t-1} \bigg) \big(\big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op} \|x_{t-l+1}^i\|_2 + \big\|\widehat{b}_J^{hi} - b^{\sharp hi}\big\|_2\big) + \big\|\widehat{b}^o - b^{\sharp o}\big\|_2^2 \bigg)^2 \Bigg\}, \end{aligned}$$

which concludes the inequality (B.1). It remains to prove (B.5). We calculate that

$$\begin{aligned}
&\big\|W^{\sharp o}\phi_J(x_t^i,\widehat{h}_{t-1}^i) - W^{\sharp o}\phi^\sharp(x_t^i,h_{t-1}^{\sharp i})\big\|_2 \\
&\qquad \leq \big\|W^{\sharp o}\big\|_{op}\big\|\sigma\big(\widehat{W}_{J,[m]}^h\phi(x_{t-1}^i,\widehat{h}_{t-2}^i) + \widehat{W}_{J,[d_x]}^i x_t^i + \widehat{b}_J^{hi}\big) \\
&\qquad\qquad\qquad - \sigma\big(W^{\sharp h}\phi^\sharp(x_{t-1}^i,h_{t-2}^{\sharp i}) + W^{\sharp i}x_t^i + b^{\sharp hi}\big)\big\|_2 \\
&\qquad \leq R_o\rho_\sigma\Big\{\underbrace{\big\|\widehat{W}_{J,[m]}^h\phi(x_{t-1}^i,\widehat{h}_{t-2}^i) - W^{\sharp h}\phi^\sharp(x_{t-1}^i,h_{t-2}^{\sharp i})\big\|_2}_{=:H_{t-1}} \\
&\qquad\qquad\qquad + \big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op}\|x_t^i\|_2 + \big\|\widehat{b}^{hi} - b^{\sharp hi}\big\|_2\Big\},
\end{aligned} \tag{B.6}$$

where $\|\cdot\|_{op}$ is the operator norm (which is the largest singular value). Concerning the quantity H_{t-1}, we estimate

$$\begin{aligned}
H_{t-1} \leq \big\|\widehat{W}_{J,[m]}^h\phi(x_{t-1}^i,\widehat{h}_{t-2}^i) - W^{\sharp h}\phi_J(x_{t-1}^i,\widehat{h}_{t-2}^i)\big\|_2 \\
+ \big\|W^{\sharp h}\phi_J(x_{t-1}^i,\widehat{h}_{t-2}^i) - W^{\sharp h}\phi^\sharp(x_{t-1}^i,h_{t-2}^{\sharp i})\big\|_2,
\end{aligned}$$

and moreover, the second term is estimated as

$$\begin{aligned}
&\big\|W^{\sharp h}\phi_J(x_{t-1}^i,\widehat{h}_{t-2}^i) - W^{\sharp h}\phi^\sharp(x_{t-1}^i,h_{t-2}^{\sharp i})\big\|_2 \\
&\qquad \leq \big\|W^{\sharp h}\big\|_{op}\big\|\sigma\big(\widehat{W}_{J,[m]}^h\phi(x_{t-2}^i,\widehat{h}_{t-3}^i) + \widehat{W}_{J,[d_x]}^i x_{t-1}^i + \widehat{b}_J^{hi}\big) \\
&\qquad\qquad\qquad - \sigma\big(W^{\sharp h}\phi^\sharp(x_{t-2}^i,h_{t-3}^{\sharp i}) + W^{\sharp i}x_{t-1}^i + b^{\sharp hi}\big)\big\|_2 \\
&\qquad \leq R_h\rho_\sigma\Big\{H_{t-2} + \big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op}\|x_{t-1}^i\|_2 + \big\|\widehat{b}_J^{hi} - b^{\sharp hi}\big\|_2\Big\},
\end{aligned}$$

for all t. Thus, we have the recursive inequality

$$\begin{aligned}
H_{t-1} &\leq \big\|\widehat{W}_{J,[m]}^h\phi(x_{t-1}^i,\widehat{h}_{t-2}^i) - W^{\sharp h}\phi_J(x_{t-1}^i,\widehat{h}_{t-2}^i)\big\|_2 \\
&\quad + R_h\rho_\sigma\Big\{H_{t-2} + \big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op}\|x_{t-1}^i\|_2 + \big\|\widehat{b}_J^{hi} - b^{\sharp hi}\big\|_2\Big\},
\end{aligned} \tag{B.7}$$

for $t = 2, \dots, T$. By repeatedly substituting (B.7) into (B.6), we arrive at (B.5):

$$\begin{aligned}
&\big\|W^{\sharp o}\,\phi_J(x_t^i,\widehat{h}_{t-1}^i) - W^{\sharp o}\phi^\sharp(x_t^i,h_{t-1}^{\sharp i})\big\|_2 \\
&\quad \leq R_o\rho_\sigma\Big\{\sum_{l=1}^{t-1}\underbrace{(R_h\rho_\sigma)^{l-1}}_{\leq\max\{1,(R_h\rho_\sigma)^{t-2}\}}\big\|\widehat{W}_{J,[m]}^h\phi(x_{t-l}^i,\widehat{h}_{t-l-1}^i) - W^{\sharp h}\phi_J(x_{t-l}^i,\widehat{h}_{t-l-1}^i)\big\|_2 \\
&\qquad + \sum_{l=1}^{t}(R_h\rho_\sigma)^{l-1}\big(\big\|\widehat{W}_{J,[d_x]}^i - W^{\sharp i}\big\|_{op}\|x_{t-l+1}^i\|_2 + \big\|\widehat{b}_J^{hi} - b^{\sharp hi}\big\|_2\big)\Big\}.
\end{aligned}$$

Thus, we conclude Proposition B.1. □

C Proof of Theorem 4.5

We restate Theorem 4.5 in an exact form as follows:

Theorem C.1. *Suppose that Assumptions 4.1 and 4.4 hold. Let $\{(X_T^i, Y_T^i)\}_{i=1}^n$ be sampled i.i.d. from the* *distribution* P_T*. Then, for any* $\delta \geq \log 2$*, we have the following inequality with probability greater than* $1 - 2e^{-\delta}$*:*

$$
\begin{aligned}
\Psi_j(f^\sharp) \leq \widehat{\Psi}_j(\widehat{f}) + \sqrt{3}\rho_\psi \Bigg\{ & \big\| \widehat{W}^o \phi - W^{\sharp o} \phi_J \big\|_{n,T} \\
& + R_o \rho_\sigma \max\{1, (R_h\rho_\sigma)^{T-2}\} T \big\| \widehat{W}^h_{J,[m]} \phi - W^{\sharp h} \phi_J \big\|_{n,T} \\
& + R_o \rho_\sigma \left(\sum_{t=1}^{T} (R_h\rho_\sigma)^{t-1} \right) \big(R_x \big\| \widehat{W}^i_{J,[d_x]} - W^{\sharp i} \big\|_{op} + \big\| \widehat{b}^{hi}_J - b^{\sharp hi} \big\|_2 \big) + \big\| \widehat{b}^o - b^{\sharp o} \big\|_2 \Bigg\} \\
& + \frac{1}{\sqrt{n}} \left\{ \frac{\widehat{c}\rho_\psi \sqrt{m^\sharp}}{T} \left(\sum_{t=1}^{T} M_t^{1/2} R_{\infty,t}^{1/2} \right) + 3\sqrt{2\delta} (\rho_\psi R_{\infty,T} + R_y) \right\},
\end{aligned}
$$

for $j = 1, \ldots, d_y$ *and for all* $J \subset [m]$ *with* $|J| = m^\sharp$ *and* $f^\sharp \in \mathcal{F}^\sharp_T(R_o, R_h, R_i, R^b_o, R^b_{hi})$*, where* $\widehat{c} := 192\sqrt{5}$*, and* $R_{\infty,t}$ *and* M_t *are defined by*

$$
R_{\infty,t} := R_o \rho_\sigma (R_i R_x + R^b_{hi}) \left(\sum_{l=1}^{t} (R_h\rho_\sigma)^{l-1} \right) + R^b_o, \tag{C.1}
$$

$$
\begin{aligned}
M_t := R_o \rho_\sigma \Bigg[& \big(d_y \min\{\sqrt{m^\sharp}, \sqrt{d_y}\} + d_x \min\{\sqrt{m^\sharp}, \sqrt{d_x}\} \big) R_i R_x \\
& + \big(d_y \min\{\sqrt{m^\sharp}, \sqrt{d_y}\} + 1 \big) R^b_{hi} \Bigg] \left(\sum_{l=0}^{t-1} (R_h\rho_\sigma)^l \right) \\
& + (m^\sharp)^{\frac{3}{2}} R_h \rho_\sigma^2 R_o (R_i R_x + R^b_{hi}) \left(\sum_{l=1}^{t-1} \sum_{k=0}^{l-1} (R_h\rho_\sigma)^{t-1-l+k} \right) + d_y R^b_o.
\end{aligned} \tag{C.2}
$$

Proof. The generalization error of $f^\sharp_t \in \mathcal{F}^\sharp_t$ is decomposed into

$$
\Psi_j(f^\sharp) = \Psi_j(\widehat{f}) + \big(\widehat{\Psi}_j(f^\sharp) - \widehat{\Psi}_j(\widehat{f}) \big) + \big(\Psi_j(f^\sharp) - \widehat{\Psi}_j(f^\sharp) \big),
$$

where the second term $\widehat{\Psi}_j(f^\sharp) - \widehat{\Psi}_j(\widehat{f})$ is called the approximation error and the third term $\Psi_j(f^\sharp) - \widehat{\Psi}_j(f^\sharp)$ is called the estimation error. Since the loss function ψ is ρ_ψ-Lipschitz continuous, the approximation error is evaluated as

$$
\begin{aligned}
\big| \widehat{\Psi}_j(f^\sharp) - \widehat{\Psi}_j(\widehat{f}) \big| &\leq \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \big| \psi(y^i_{t,j}, f^\sharp_t(X^i_t)_j) - \psi(y^i_t, \widehat{f}_t(X^i_t)_j) \big| \\
&\leq \frac{\rho_\psi}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \big| f^\sharp(X^i_t)_j - \widehat{f}_t(X^i_t)_j \big| \\
&\leq \frac{\rho_\psi}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \big\| f^\sharp_t(X^i_t) - \widehat{f}_t(X^i_t) \big\|_2 \\
&\leq \rho_\psi \sqrt{\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \big\| f^\sharp_t(X^i_t) - \widehat{f}_t(X^i_t) \big\|_2^2} = \rho_\psi \big\| f^\sharp - \widehat{f} \big\|_{n,T}.
\end{aligned}
$$

The term $\|f^\sharp - \widehat{f}\|_{n,T}$ is evaluated by Proposition 4.2 (see also Proposition B.1). In the rest of the proof, let us concentrate on the estimation error bound.

First, we define the following function space

$$
\mathcal{G}^\sharp_{T,j} := \left\{ g_j \,\middle|\, g_j(Y_T, X_T) = \frac{1}{T} \sum_{t=1}^{T} \psi(y_{t,j}, f_t(X_t)_j) \text{ for } (X_T, Y_T) \in \operatorname{supp}(P_T),\ f \in \mathcal{F}^\sharp_T, \right\}
$$

for $j = 1, \dots, d_y$. For $g_j \in \mathcal{G}^{\sharp}_{T,j}$, we have

$$\begin{aligned}\left|g_j(Y_T, X_T)\right| &\leq \frac{1}{T}\sum_{t=1}^{T}\left\{|\psi(y_{t,j}, f_t(X_t)_j) - \psi(y_{t,j}, 0)| + |\psi(y_{t,j}, 0)|\right\} \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\rho_\psi |f_t(X_t)_j| + R_y\right) \leq \frac{\rho_\psi}{T}\sum_{t=1}^{T}\|f_t(X_t)\|_2 + R_y.\end{aligned}$$

The quantity $\|f_t(X_t)\|_2$ is evaluated by

$$\|f_t(X_t)\|_2 \leq R_o\rho_\sigma\left\|W^{\sharp h}\phi^{\sharp}(x_{t-1}, h^{\sharp i}_{t-2}) + W^{\sharp i}x_t + b^{\sharp hi}\right\|_2 + R^b_o.$$

The recurrent structure (B.2) and (B.3) give

$$\begin{aligned}&\left\|W^{\sharp h}\phi^{\sharp}\left(x_{t-1}, h^{\sharp i}_{t-2}\right) + W^{\sharp i}x_t + b^{\sharp hi}\right\|_2 \\ &\qquad \leq R_h\rho_\sigma\left\|W^{\sharp h}\phi^{\sharp}(x_{t-2}, h^{\sharp i}_{t-3}) + W^{\sharp i}x_{t-1} + b^{\sharp hi}\right\|_2 + R_iR_x + R^b_{hi},\end{aligned}$$

as this is repeated

$$\left\|W^{\sharp h}\phi^{\sharp}(x_{t-1}, h^{\sharp i}_{t-2}) + W^{\sharp i}x_t + b^{\sharp hi}\right\|_2 \leq (R_iR_x + R^b_{hi})\left(\sum_{l=1}^{t}(R_h\rho_\sigma)^{l-1}\right).$$

Hence, we see from (C.1) that

$$\|f_t(X_t)\|_2 \leq R_o\rho_\sigma(R_iR_x + R^b_{hi})\left(\sum_{l=1}^{t}(R_h\rho_\sigma)^{l-1}\right) + R^b_o = R_{\infty,t},$$

which implies that

$$\begin{aligned}\left|g_j(Y_T, X_T)\right| &\leq \rho_\psi R_o\rho_\sigma(R_iR_x + R^b_{hi})\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{t}(R_h\rho_\sigma)^{l-1}\right\} + R^b_o + R_y \\ &\leq \rho_\psi R_o\rho_\sigma(R_iR_x + R^b_{hi})\left(\sum_{t=1}^{T}(R_h\rho_\sigma)^{t-1}\right) + R^b_o + R_y \\ &= \rho_\psi R_{\infty,T} + R_y.\end{aligned}$$

By Theorem 3.4.5 in (Giné and Nickl, 2015), for any $\delta > \log 2$, we have the following inequality with probability grater than $1 - 2e^{-\delta}$:

$$\begin{aligned}\left|\Psi_j(f^{\sharp}) - \widehat{\Psi}_j(f^{\sharp})\right| &\leq \sup_{g_j \in \mathcal{G}^{\sharp}_{T,j}}\left|\frac{1}{n}\sum_{i=1}^{n}g_j(Y^i_T, X^i_T) - E_{P_T}[g_j(Y_T, X_T)]\right| \\ &\leq 2E_\epsilon\left[\sup_{g_j \in \mathcal{G}^{\sharp}_{T,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g_j(Y^i_T, X^i_T)\right|\right] + 3(\rho_\psi R_{\infty,T} + R_y)\sqrt{\frac{2\delta}{n}},\end{aligned}$$

where $(\epsilon_i)_{i=1}^n$ is the i.i.d. Rademacher sequence (see, e.g., Definition 3.1.19 in (Giné and Nickl, 2015)). The first term of right-hand side in the above inequality, called the Rademacher complexity, is estimated by using Theorem 4.12 in (Ledoux and Talagrand, 2013), and Lemma A.5 in (Bartlett et al., 2017) (or Lemma 9 in (Chen et al., 2019)) as follows:

$$\begin{aligned}E_\epsilon\left[\sup_{g_j \in \mathcal{G}^{\sharp}_{T,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g_j(Y^i_T, X^i_T)\right|\right] &\leq \frac{1}{T}\sum_{t=1}^{T}E_\epsilon\left[\sup_{f_{t,j} \in \mathcal{F}^{\sharp}_{t,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_j(y^i_{t,j}, f_t(X^i_t)_j)\right|\right] \\ &\leq \frac{2\rho_\psi}{T}\sum_{t=1}^{T}E_\epsilon\left[\sup_{f_{t,j} \in \mathcal{F}^{\sharp}_{t,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_t(X^i_t)_j\right|\right] \\ &\leq \frac{2\rho_\psi}{T}\sum_{t=1}^{T}\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_{\alpha}^{2R_{\infty,t}\sqrt{n}}\sqrt{\log N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S)}\,d\epsilon\right),\end{aligned}$$

where $\mathcal{F}^{\sharp}_{t,j}$ and $\|\cdot\|_S$ are defined by

$$\mathcal{F}^{\sharp}_{t,j} := \left\{ f_{t,j} \,\middle|\, f_{t,j}(X_t) = f_t(X_t)_j \text{ for } X_t \in \mathrm{supp}(P_{X_t}),\ f \in \mathcal{F}^{\sharp}_T \right\},$$

$$\|f_{t,j}\|_S := \left(\sum_{i=1}^{n} |f_t(X_t^i)_j|^2 \right)^{1/2}.$$

Here, we denote by $N(F, \epsilon, \|\cdot\|)$ the covering number of F which means the minimal cardinality of a subset $C \subset F$ that covers F at scale ϵ with respect to the norm $\|\cdot\|$. By using Lemma D.1 in Appendix D, for any $\delta > \log 2$, we conclude the following estimation error bound:

$$\begin{aligned}
&\left|\Psi_j(f^{\sharp}) - \widehat{\Psi}_j(f^{\sharp})\right| \\
&\le \frac{16\rho_\psi \alpha}{\sqrt{n}} + \frac{48\rho_\psi}{nT} \sum_{t=1}^{T} \int_{\alpha}^{2R_{\infty,t}\sqrt{n}} \sqrt{\log N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S)}\, d\epsilon + 3(\rho_\psi R_{\infty,T} + R_y)\sqrt{\frac{2\delta}{n}} \\
&\le \frac{48\rho_\psi}{nT} \sqrt{10 m^{\sharp}} n^{1/4} \sum_{t=1}^{T} M_t^{1/2} \int_{\alpha}^{2R_{\infty,t}\sqrt{n}} \frac{d\epsilon}{\sqrt{\epsilon}} + 3(\rho_\psi R_{\infty,T} + R_y)\sqrt{\frac{2\delta}{n}} + O(\alpha) \\
&= \frac{\widehat{c}\rho_\psi \sqrt{m^{\sharp}}}{\sqrt{n}T} \left(\sum_{t=1}^{T} M_t^{1/2} R_{\infty,t}^{1/2} \right) + 3(\rho_\psi R_{\infty,T} + R_y)\sqrt{\frac{2\delta}{n}} + O(\alpha),
\end{aligned}$$

for all $\alpha > 0$ with probability grater than $1 - 2e^{-\delta}$, where $\widehat{c} := 192\sqrt{5}$, and M_t is defined by (C.2). The proof of Theorem C.1 is complete. □

D Upper Bound of the Covering Number

Lemma D.1. *Under the same assumptions as in Theorem C.1, the covering number $N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S)$ has the following bound:*

$$\log N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S) \le \frac{10 m^{\sharp} n^{1/2} M_t}{\epsilon},$$

for any $\epsilon > 0$, where M_t is given by (C.2).

Proof. The proof is based on the argument of the proof of Lemma 3 in (Chen et al., 2019). For $f^{\sharp}_{t,j}, \widetilde{f}^{\sharp}_{t,j} \in \mathcal{F}^{\sharp}_{t,j}$, we estimate

$$\begin{aligned}
|f^{\sharp}_t(X_t)_j - \widetilde{f}^{\sharp}_t(X_t)_j| &\le \left\| f^{\sharp}_t(X_t) - \widetilde{f}^{\sharp}_t(X_t) \right\|_2 \\
&\le \left\| W^{\sharp o} - \widetilde{W}^{\sharp o} \right\|_{op} \left\| \phi^{\sharp}(x_t, h^{\sharp}_{t-1}) \right\|_2 \\
&\quad + \left\| \widetilde{W}^{\sharp o} \widetilde{\phi}^{\sharp}(x_t, \widetilde{h}^{\sharp}_{t-1}) - \widetilde{W}^{\sharp o} \phi^{\sharp}(x_t, h^{\sharp}_{t-1}) \right\|_2 + \left\| b^{\sharp o} - \widetilde{b}^{\sharp o} \right\|_2.
\end{aligned}$$

The second term of right-hand side is estimated as

$$\begin{aligned}
&\left\| \widetilde{W}^{\sharp o} \widetilde{\phi}^{\sharp}(x_t, \widetilde{h}^{\sharp}_{t-1}) - \widetilde{W}^{\sharp o} \phi^{\sharp}(x_t, h^{\sharp}_{t-1}) \right\|_2 \\
&\qquad \le \left\| \widetilde{W}^{\sharp o} \right\|_{op} \rho_\sigma \Big(\left\| \widetilde{W}^{\sharp h} \widetilde{\phi}^{\sharp}(x_{t-1}, \widetilde{h}^{\sharp}_{t-2}) - W^{\sharp h} \phi^{\sharp}(x_{t-1}, h^{\sharp}_{t-2}) \right\|_2 \\
&\qquad\qquad + \left\| W^{\sharp i} - \widetilde{W}^{\sharp i} \right\|_{op} \left\| x_t \right\|_2 + \left\| b^{\sharp hi} - \widetilde{b}^{hi} \right\|_2 \Big).
\end{aligned}$$

We estimate the first term of right-hand side in the above inequality as

$$\begin{aligned}
&\left\| \widetilde{W}^{\sharp h} \widetilde{\phi}^{\sharp}(x_{t-1}, \widetilde{h}^{\sharp}_{t-2}) - W^{\sharp h} \phi^{\sharp}(x_{t-1}, h^{\sharp}_{t-2}) \right\|_2 \\
&\le \left\| \widetilde{W}^{\sharp h} \right\|_{op} \left\| \widetilde{\phi}^{\sharp}(x_{t-1}, \widetilde{h}^{\sharp}_{t-2}) - \phi^{\sharp}(x_{t-1}, h^{\sharp}_{t-2}) \right\|_2 + \left\| W^{\sharp h} - \widetilde{W}^{\sharp h} \right\|_{op} \left\| \phi^{\sharp}(x_{t-1}, h^{\sharp}_{t-2}) \right\|_2 \\
&\le \left\| \widetilde{W}^{\sharp h} \right\|_{op} \rho_\sigma \Big(\left\| \widetilde{W}^{\sharp h} \widetilde{\phi}^{\sharp}(x_{t-2}, \widetilde{h}^{\sharp}_{t-3}) - W^{\sharp h} \phi^{\sharp}(x_{t-2}, h^{\sharp}_{t-3}) \right\|_2 \\
&\quad + \left\| W^{\sharp i} - \widetilde{W}^{\sharp i} \right\|_{op} \left\| x_{t-1} \right\|_2 + \left\| \widetilde{b}^{hi} - b^{\sharp hi}) \right\|_2 \Big) + \left\| W^{\sharp h} - \widetilde{W}^{\sharp h} \right\|_{op} \left\| \phi^{\sharp}(x_{t-1}, h^{\sharp}_{t-2}) \right\|_2,
\end{aligned}$$

and as this is repeated, we eventually obtain

$$
\begin{aligned}
&\left\|\widetilde{W}^{\sharp o}\widetilde{\phi}^{\sharp}(x_t,\widetilde{h}^{\sharp}_{t-1})-\widetilde{W}^{\sharp o}\phi^{\sharp}(x_t,h^{\sharp}_{t-1})\right\|_2\\
&\qquad\le R_o\rho_\sigma\Bigg\{\sum_{l=0}^{t-1}(R_h\rho_\sigma)^l\Big(\left\|W^{\sharp i}-\widetilde{W}^{\sharp i}\right\|_{op}\left\|x_{t-l}\right\|_2+\left\|b^{\sharp hi}-\widetilde{b}^{\sharp hi}\right\|_2\Big)\\
&\qquad\qquad+\sum_{l=1}^{t-1}(R_h\rho_\sigma)^{t-1-l}\left\|\phi^{\sharp}(x_l,h^{\sharp}_{l-1})\right\|_2\left\|W^{\sharp h}-\widetilde{W}^{\sharp h}\right\|_{op}\Bigg\}.
\end{aligned}
$$

Summarizing the above, we have

$$
\begin{aligned}
|f^{\sharp}_t(X_t)_j-\widetilde{f}^{\sharp}_t(X_t)_j|&\le\left\|W^{\sharp o}-\widetilde{W}^{\sharp o}\right\|_{op}\left\|\phi^{\sharp}(x_t,h^{\sharp}_{t-1})\right\|_2\\
&\quad+R_o\rho_\sigma\Bigg\{\sum_{l=0}^{t-1}(R_h\rho_\sigma)^l\Big(\left\|W^{\sharp i}-\widetilde{W}^{\sharp i}\right\|_{op}\left\|x_{t-l}\right\|_2+\left\|b^{\sharp hi}-\widetilde{b}^{\sharp hi}\right\|_2\Big)\\
&\quad+\sum_{l=1}^{t-1}(R_h\rho_\sigma)^{t-1-l}\left\|\phi^{\sharp}(x_l,h^{\sharp}_{l-1})\right\|_2\left\|W^{\sharp h}-\widetilde{W}^{\sharp h}\right\|_{op}\Bigg\}+\left\|b^{\sharp o}-\widetilde{b}^{\sharp o}\right\|_2.
\end{aligned}
$$

Since

$$
\begin{aligned}
\left\|\phi^{\sharp}_t(x_t,h^{\sharp}_{t-1})\right\|_2&\le\rho_\sigma\big(\left\|W^{\sharp h}\right\|_{op}\left\|\phi^{\sharp}(x_{t-1},h^{\sharp}_{t-2})\right\|_2+\left\|W^{\sharp i}\right\|_{op}\left\|x_t\right\|_2+\left\|b^{\sharp hi}\right\|_2\big)\\
&\le\rho_\sigma(R_iR_x+R^b_{hi})\sum_{l=0}^{t-1}(R_h\rho_\sigma)^l,
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{l=1}^{t-1}(R_h\rho_\sigma)^{t-1-l}\left\|\phi^{\sharp}_t(x_l,h^{\sharp}_{l-1})\right\|_2&\le\rho_\sigma(R_iR_x+R^b_{hi})\sum_{l=1}^{t-1}\sum_{k=0}^{l-1}(R_h\rho_\sigma)^{t-1-l}(R_h\rho_\sigma)^k\\
&=\rho_\sigma(R_iR_x+R^b_{hi})\sum_{l=1}^{t-1}\sum_{k=0}^{l-1}(R_h\rho_\sigma)^{t-1-l+k},
\end{aligned}
$$

we see that

$$
\begin{aligned}
|f^{\sharp}_t(X_t)_j-\widetilde{f}^{\sharp}_t(X_t)_j|&\le\underbrace{\rho_\sigma(R_iR_x+R^b_{hi})\Bigg(\sum_{l=0}^{t-1}(R_h\rho_\sigma)^l\Bigg)}_{=:L_{o,t}}\left\|W^{\sharp o}-\widetilde{W}^{\sharp o}\right\|_{op}\\
&+\underbrace{\rho_\sigma R_oR_x\Bigg(\sum_{l=0}^{t-1}(R_h\rho_\sigma)^l\Bigg)}_{=:L_{i,t}}\left\|W^{\sharp i}-\widetilde{W}^{\sharp i}\right\|_{op}+\underbrace{\rho_\sigma R_o\Bigg(\sum_{l=0}^{t-1}(R_h\rho_\sigma)^l\Bigg)}_{=:L_{b,t}}\left\|b^{\sharp hi}-\widetilde{b}^{\sharp hi}\right\|_2\\
&+\underbrace{\rho_\sigma^2R_o(R_iR_x+R^b_{hi})\Bigg(\sum_{l=1}^{t-1}\sum_{k=0}^{l-1}(R_h\rho_\sigma)^{t-1-l+k}\Bigg)}_{=:L_{h,t}}\left\|W^{\sharp h}-\widetilde{W}^{\sharp h}\right\|_{op}+\left\|b^{\sharp o}-\widetilde{b}^{\sharp o}\right\|_2.
\end{aligned}
\tag{D.1}
$$

Since the right-hand side of (D.1) is independent of the training data X^i_t, we estimate

$$
\begin{aligned}
\left\|f^{\sharp}_t(X_t)_j-\widetilde{f}^{\sharp}_t(X_t)_j\right\|_S&=\Bigg(\sum_{i=1}^{n}\left|f^{\sharp}_t(X^i_t)_j-\widetilde{f}^{\sharp}_t(X^i_t)_j\right|^2\Bigg)^{1/2}\\
&\le\sqrt{n}\Big(L_{o,t}\left\|W^{\sharp o}-\widetilde{W}^{\sharp o}\right\|_{op}+L_{i,t}\left\|W^{\sharp i}-\widetilde{W}^{\sharp i}\right\|_{op}\\
&\quad+L_{b,t}\left\|b^{\sharp hi}-\widetilde{b}^{\sharp hi}\right\|_2+L_{h,t}\left\|W^{\sharp h}-\widetilde{W}^{\sharp h}\right\|_{op}+\left\|b^{\sharp o}-\widetilde{b}^{\sharp o}\right\|_2\Big).
\end{aligned}
$$

Then, the covering number $N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S)$ is bounded as follows

$$
\begin{aligned}
N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S) \leq\, & N\Big(\mathcal{H}_{W^{\sharp o}, R_o}, \frac{\epsilon}{5\sqrt{n}L_{o,t}}, \|\cdot\|_F\Big) N\Big(\mathcal{H}_{W^{\sharp i}, R_i}, \frac{\epsilon}{5\sqrt{n}L_{i,t}}, \|\cdot\|_F\Big) \\
& \times N\Big(\mathcal{H}_{b^{\sharp hi}, R^b_{hi}}, \frac{\epsilon}{5\sqrt{n}L_{b,t}}, \|\cdot\|_F\Big) N\Big(\mathcal{H}_{W^{\sharp h}, R_h}, \frac{\epsilon}{5\sqrt{n}L_{h,t}}, \|\cdot\|_F\Big) N\Big(\mathcal{H}_{b^{\sharp o}, R^b_o}, \frac{\epsilon}{5\sqrt{n}}, \|\cdot\|_F\Big),
\end{aligned}
$$

where we used the notation

$$
\mathcal{H}_{A,R} := \big\{A \in \mathbb{R}^{d_1 \times d_2} \mid \|A\|_F \leq R\big\}.
$$

By Lemma 8 in (Chen et al., 2019), the above five covering numbers are bounded as

$$
N\Big(\mathcal{H}_{W^{\sharp o}, R_o}, \frac{\epsilon}{5\sqrt{n}L_{o,t}}, \|\cdot\|_F\Big) \leq \Big(1 + \frac{10\min\{\sqrt{m^\sharp}, \sqrt{d_y}\}R_o L_{o,t}\sqrt{n}}{\epsilon}\Big)^{m^\sharp d_y},
$$

$$
N\Big(\mathcal{H}_{W^{\sharp i}, R_i}, \frac{\epsilon}{5\sqrt{n}L_{i,t}}, \|\cdot\|_F\Big) \leq \Big(1 + \frac{10\min\{\sqrt{m^\sharp}, \sqrt{d_x}\}R_i L_{i,t}\sqrt{n}}{\epsilon}\Big)^{m^\sharp d_x},
$$

$$
N\Big(\mathcal{H}_{b^{\sharp hi}, R^b_{hi}}, \frac{\epsilon}{5\sqrt{n}L_{b,t}}, \|\cdot\|_F\Big) \leq \Big(1 + \frac{10 R^b_{hi} L_{b,t}\sqrt{n}}{\epsilon}\Big)^{m^\sharp},
$$

$$
N\Big(\mathcal{H}_{W^{\sharp h}, R_h}, \frac{\epsilon}{5\sqrt{n}L_{h,t}}, \|\cdot\|_F\Big) \leq \Big(1 + \frac{10\sqrt{m^\sharp} R_h L_{h,t}\sqrt{n}}{\epsilon}\Big)^{(m^\sharp)^2},
$$

$$
N\Big(\mathcal{H}_{b^{\sharp o}, R^b_o}, \frac{\epsilon}{5\sqrt{n}}, \|\cdot\|_F\Big) \leq \Big(1 + \frac{10 R^b_o \sqrt{n}}{\epsilon}\Big)^{d_y}.
$$

Therefore, by using $\log(1+x) \leq x$ for $x \geq 0$, we conclude that

$$
\begin{aligned}
& \log N(\mathcal{F}^{\sharp}_{t,j}, \epsilon, \|\cdot\|_S) \\
& \leq \frac{10 m^\sharp d_y \min\{\sqrt{m^\sharp}, \sqrt{d_y}\} R_o L_{o,t}\sqrt{n}}{\epsilon} + \frac{10 m^\sharp d_x \min\{\sqrt{m^\sharp}, \sqrt{d_x}\} R_i L_{i,t}\sqrt{n}}{\epsilon} \\
& \quad + \frac{10 m^\sharp R^b_{hi} L_{b,t}\sqrt{n}}{\epsilon} + \frac{10 (m^\sharp)^{\frac{5}{2}} R_h L_{h,t}\sqrt{n}}{\epsilon} + \frac{10 m^\sharp d_y R^b_o \sqrt{n}}{\epsilon} \\
& = \frac{10 m^\sharp \sqrt{n} M_t}{\epsilon},
\end{aligned}
$$

where M_t is the constant given by (C.2). The proof of Lemma D.1 is finished. □

E Proof of Proposition 4.6

We review the following proposition (see Proposition 1 in (Suzuki et al., 2020) and Proposition 1 in (Bach, 2017)).

Proposition E.1. *Let $v_1, \ldots, v_{m^\sharp}$ be i.i.d. sampled from the distribution q in (4.6), and $J = \{v_1, \ldots, v_{m^\sharp}\}$. Then, for any $\widetilde{\delta} \in (0, 1/2)$ and $\lambda > 0$, if $m^\sharp \geq 5\widehat{N}(\lambda)\log(16\widehat{N}(\lambda)/\widetilde{\delta})$, then we have the following inequality with probability greater than $1 - \widetilde{\delta}$:*

$$
\inf_{\alpha \in \mathbb{R}^{m^\sharp}} \big\{\big\|z^T\phi - \alpha^T\phi_J\big\|^2_{n,T} + \lambda m^\sharp \big\|\alpha^T\big\|^2_{\tau'}\big\} \leq 4\lambda z^T \widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1} z, \tag{E.1}
$$

for all $z \in \mathbb{R}^m$.

Proof. Let e_j be an indicator vector which has 1 at the j-th component and 0 in other components for $j = 1, \ldots, m$.

Applying Proposition E.1 with $z = e_j$ and taking the summation over $j = 1, \dots, m$, we obtain

$$\begin{aligned}
L_\tau^{(A)}(J) &= \big\|\phi - \widehat{A}_J \phi_J\big\|_{n,T}^2 + \lambda m^\sharp \big\|\widehat{A}_J\big\|_{\tau'}^2 \\
&\le \sum_{j=1}^m \big\{\big\|e_j^T \phi - e_j^T \widehat{A}_J \phi_J\big\|_{n,T}^2 + \lambda m^\sharp \big\|e_j^T \widehat{A}_J\big\|_{\tau'}^2\big\} \\
&= \sum_{j=1}^m \inf_{\alpha \in \mathbb{R}^{m^\sharp}} \big\{\big\|e_j^T \phi - \alpha^T \phi_J\big\|_{n,T}^2 + \lambda m^\sharp \big\|\alpha^T\big\|_{\tau'}^2\big\} \\
&\le 4\lambda \sum_{j=1}^m e_j^T \widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1} e_j \le 4\lambda.
\end{aligned}$$

□

F Proof of Theorem 4.8

We restate Theorem 4.8 in an exact form as follows:

Theorem F.1. *Suppose that Assumptions 4.1, 4.4 and 4.7 hold. Let $\{(X_T^i, Y_T^i)\}_{i=1}^n$ and $\{v_j\}_{j=1}^{m^\sharp}$ be sampled i.i.d. from the distributions P_T and q in (4.6), respectively. Let $J = \{v_1, \dots, v_{m^\sharp}\}$. Then, for any $\delta \ge \log 2$ and $\widetilde{\delta} \in (0, 1/2)$, we have the following inequality with probability greater than $(1 - 2e^{-\delta})\widetilde{\delta}$:*

$$\begin{aligned}
\Psi_j(f_J^\sharp) &\le \widehat{\Psi}_j(\widehat{f}) \\
&\quad + \sqrt{3}\rho_\psi \left\{ 2\widehat{R}_o + 4\widehat{R}_o \rho_\sigma \sqrt{\frac{m}{1 - 2\widetilde{\delta}}} \max\left\{1, \left(2\rho_\sigma \widehat{R}_h \sqrt{\frac{m}{1 - 2\widetilde{\delta}}}\right)^{T-2}\right\} T\widehat{R}_h \right\} \sqrt{\lambda} \\
&\quad + \frac{1}{\sqrt{n}} \left\{ \frac{\widehat{c}\rho_\psi \sqrt{m^\sharp}}{T} \left(\sum_{t=1}^T \widehat{M}_t^{1/2} \widehat{R}_{\infty,t}^{1/2} \right) + 3\sqrt{2\delta}(\rho_\psi \widehat{R}_{\infty,T} + R_y) \right\} \\
&\lesssim \widehat{\Psi}_j(\widehat{f}) + \sqrt{\lambda} + \frac{1}{\sqrt{n}} (m^\sharp)^{\frac{5}{4}} \widehat{R}_{\infty,T}^{1/2},
\end{aligned} \tag{F.1}$$

for $j = 1, \dots, d_y$ and for all $\lambda > 0$ satisfying (4.4), where $\widehat{R}_{\infty,t}$ and $\widehat{M}_t$ are defined by

$$\widehat{R}_{\infty,t} := 2\rho_\sigma \widehat{R}_o \sqrt{\frac{m}{1 - 2\widetilde{\delta}}} (\widehat{R}_i R_x + \widehat{R}_{hi}^b) \left\{ \sum_{l=1}^t \left(2\rho_\sigma \widehat{R}_h \sqrt{\frac{m}{1 - 2\widetilde{\delta}}}\right)^{l-1} \right\} + \widehat{R}_o^b,$$

$$\begin{aligned}
\widehat{M}_t := 2\rho_\sigma \widehat{R}_o \sqrt{\frac{m}{1 - 2\widetilde{\delta}}} &\left\{ \left(d_y \min\{\sqrt{m^\sharp}, \sqrt{d_y}\} + d_x \min\{\sqrt{m^\sharp}, \sqrt{d_x}\}\right) \widehat{R}_i R_x \right. \\
&\left. + \left(d_y \min\{\sqrt{m^\sharp}, \sqrt{d_y}\} + 1\right) \widehat{R}_{hi}^b \right\} \left\{ \sum_{l=0}^{t-1} \left(2\rho_\sigma \widehat{R}_h \sqrt{\frac{m}{1 - 2\widetilde{\delta}}}\right)^l \right\} \\
&+ 4(m^\sharp)^{3/2} \widehat{R}_h \widehat{R}_o \frac{m}{1 - 2\widetilde{\delta}} \rho_\sigma^2 (\widehat{R}_i R_x + \widehat{R}_{hi}^b) \left\{ \sum_{l=1}^{t-1} \sum_{k=0}^{l-1} \left(2\rho_\sigma \widehat{R}_h \sqrt{\frac{m}{1 - 2\widetilde{\delta}}}\right)^{t-1-l+k} \right\} + d_y \widehat{R}_o^b.
\end{aligned}$$

Proof. Let $\tilde{\delta} \in (0, 1/2)$, and let $f_J^\sharp$ be the compressed RNN with parameters

$$W_J^{\sharp o} := \widehat{W}^o \widehat{A}_J, \quad W_J^{\sharp h} := \widehat{W}_{J,[m]}^h \widehat{A}_J, \quad W_J^{\sharp i} := \widehat{W}_{J,[d_x]}^i, \quad b_J^{\sharp hi} := \widehat{b}_J^{hi}, \quad \text{and} \quad b_J^{\sharp o} := \widehat{b}^o.$$

Once we can prove that

$$f_J^\sharp \in \mathcal{F}_T^\sharp\left(2\widehat{R}_o \sqrt{\frac{m}{1 - 2\widetilde{\delta}}}, 2\widehat{R}_h \sqrt{\frac{m}{1 - 2\widetilde{\delta}}}, \widehat{R}_i, \widehat{R}_o^b, \widehat{R}_{hi}^b\right), \tag{F.2}$$

we can apply Theorem C.1 with $f^\sharp = f^\sharp_J$ to obtain, for any $\delta \ge \log 2$, the following inequality with probability greater than $1 - 2e^{-\delta}$:

$$\begin{aligned}\Psi_j(f^\sharp_J) \le \widehat{\Psi}_j(\widehat{f}) &+ \sqrt{3}\rho_\psi\Bigg\{\big\|\widehat{W}^o\phi - W^{\sharp o}\phi_J\big\|_{n,T} \\ &+ 2\widehat{R}_o\sqrt{\frac{m}{1-2\widetilde{\delta}}}\rho_\sigma \max\left\{1, \left(2\widehat{R}_h\sqrt{\frac{m}{1-2\widetilde{\delta}}}\rho_\sigma\right)^{T-2}\right\}T\big\|\widehat{W}^h_{J,[m]}\phi - W^{\sharp h}\phi_J\big\|_{n,T}\Bigg\} \\ &+ \frac{1}{\sqrt{n}}\left\{\frac{\widehat{c}\rho_\psi\sqrt{m^\sharp}}{T}\left(\sum_{t=1}^{T}\widehat{M}_t^{1/2}\widehat{R}_{\infty,t}^{1/2}\right) + 3\sqrt{2\delta}(\rho_\psi\widehat{R}_{\infty,T} + R_y)\right\},\end{aligned} \tag{F.3}$$

for $j = 1, \dots, d_y$. Moreover, by using Proposition E.1, we have

$$\begin{aligned}\big\|\widehat{W}^o\phi - W^{\sharp o}_J\phi_J\big\|^2_{n,T} &= \big\|\widehat{W}^o\phi - \widehat{W}^o\widehat{A}_J\phi_J\big\|^2_{n,T} \\ &\le \sum_{j=1}^{d_y}\left(\big\|\widehat{W}^o_{j,:}\phi - \widehat{W}^o_{j,:}\widehat{A}_J\phi_J\big\|^2_{n,T} + \lambda m^\sharp\big\|\widehat{W}^o_{j,:}\widehat{A}_J\big\|^2_{\tau'}\right) \\ &= \sum_{j=1}^{d_y}\inf_{\alpha\in\mathbb{R}^{m^\sharp}}\left(\big\|\widehat{W}^o_{j,:}\phi - \alpha^T\phi_J\big\|^2_{n,T} + \lambda m^\sharp\big\|\alpha^T\big\|^2_{\tau'}\right) \\ &\le 4\lambda\sum_{j=1}^{d_y}\widehat{W}^o_{j,:}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}(\widehat{W}^o_{j,:})^T \\ &\le 4\lambda\big\|\widehat{W}^o\big\|^2_F \le 4\lambda(\widehat{R}_o)^2,\end{aligned} \tag{F.4}$$

and

$$\begin{aligned}\big\|\widehat{W}^h_{J,[m]}\phi - W^{\sharp h}_J\phi_J\big\|^2_{n,T} &= \big\|\widehat{W}^h_{J,[m]}\phi - \widehat{W}^h_{J,[m]}\widehat{A}_J\phi_J\big\|^2_{n,T} \\ &\le \sum_{j\in J}\left(\big\|\widehat{W}^h_{j,:}\phi - \widehat{W}^h_{j,:}\widehat{A}_J\phi_J\big\|^2_{n,T} + \lambda m^\sharp\big\|\widehat{W}^h_{j,:}\widehat{A}_J\big\|^2_{\tau'}\right) \\ &= \sum_{j\in J}\inf_{\alpha\in\mathbb{R}^{m^\sharp}}\left(\big\|\widehat{W}^h_{j,:}\phi - \alpha^T\phi_J\big\|^2_{n,T} + \lambda m^\sharp\big\|\alpha^T\big\|^2_{\tau'}\right) \\ &\le 4\lambda\sum_{j\in J}\widehat{W}^h_{j,:}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}(\widehat{W}^h_{j,:})^T \\ &\le 4\lambda\big\|\widehat{W}^h\big\|^2_F \le 4\lambda(\widehat{R}_h)^2.\end{aligned} \tag{F.5}$$

Therefore, by combining (F.3), (F.4) and (F.5), we conclude the inequality (F.1). It remains to prove (F.2). Finally, we prove that (F.2) holds with probability greater than $\widetilde{\delta}$.

Let us recall the definition (4.5) of the leverage score $\tau' = (\tau'_j)_{j\in J} \in \mathbb{R}^{m^\sharp}$, i.e.,

$$\tau'_j := \frac{1}{\widehat{N}(\lambda)}\big[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\big]_{j,j}, \quad j = 1, \cdots, m.$$

By Markov's inequality, we have

$$\begin{aligned}P\left[\sum_{j\in J}(\tau'_j)^{-1} < \frac{mm^\sharp}{1-2\widetilde{\delta}}\right] &= 1 - P\left[\sum_{j\in J}(\tau'_j)^{-1} \ge \frac{mm^\sharp}{1-2\widetilde{\delta}}\right] \\ &\ge 1 - \frac{E\big[\sum_{j\in J}(\tau'_j)^{-1}\big]}{\frac{mm^\sharp}{1-2\widetilde{\delta}}} = 2\widetilde{\delta},\end{aligned} \tag{F.6}$$

because $E\big[\sum_{j\in J}(\tau'_j)^{-1}\big] = mm^\sharp$ (see the proof of Lemma 1 in (Suzuki et al., 2020)). Therefore, the probability of two events (E.1) and

$$\sum_{j\in J}(\tau'_j)^{-1} < \frac{mm^\sharp}{1-2\widetilde{\delta}}, \tag{F.7}$$

happening simultaneously is greater than $(1-\widetilde{\delta}) + 2\widetilde{\delta} - 1 = \widetilde{\delta}$. By the same argument as in (F.4) and (F.5), and by using (F.6), we have

$$\begin{aligned}
\left\|W_J^{\sharp o}\right\|_F^2 &= \frac{\lambda m^\sharp}{\lambda m^\sharp}\left\|\widehat{W}^o \widehat{A}_J\right\|_F^2 \\
&\le \frac{(\sum_{j\in J}(\tau_j')^{-1})}{\lambda m^\sharp}\sum_{j=1}^{d_y}\left(\left\|\widehat{W}^o_{j,:}\phi - \widehat{W}^o_{j,:}\widehat{A}_J\phi_J\right\|_{n,T}^2 + \lambda m^\sharp\left\|\widehat{W}^o_{j,:}\widehat{A}_J\right\|_{\tau'}^2\right) \\
&\le 4(\widehat{R}_o)^2\frac{m}{1-2\widetilde{\delta}},
\end{aligned}$$

$$\begin{aligned}
\left\|W_J^{\sharp h}\right\|_F^2 &= \frac{\lambda m^\sharp}{\lambda m^\sharp}\left\|\widehat{W}^h_{J,[m]} \widehat{A}_J\right\|_F^2 \\
&\le \frac{(\sum_{j\in J}(\tau_j')^{-1})}{\lambda m^\sharp}\sum_{j\in J}\left(\left\|\widehat{W}^h_{j,:}\phi - \widehat{W}^h_{j,:}\widehat{A}_J\phi_J\right\|_{n,T}^2 + \lambda m^\sharp\left\|\widehat{W}^h_{j,:}\widehat{A}_J\right\|_{\tau'}^2\right) \\
&\le 4(\widehat{R}_h)^2\frac{m}{1-2\widetilde{\delta}},
\end{aligned}$$

and

$$\left\|W_J^{\sharp i}\right\|_F^2 \le \left\|\widehat{W}^i\right\|_F^2 \le (\widehat{R}_i)^2, \quad \left\|b_J^{\sharp o}\right\|_F^2 \le \left\|\widehat{b}^o\right\|_F^2 \le (\widehat{R}_o^b)^2, \quad \left\|b_J^{\sharp hi}\right\|_F^2 \le \left\|\widehat{b}^{hi}\right\|_F^2 \le (\widehat{R}_{hi}^b)^2.$$

Hence, (F.2) holds with probability greater than $\widetilde{\delta}$. Thus, we conclude Theorem F.1. □

G Remarks for Theorems 4.8 and F.1

Remark G.1. We remark that the index J in Theorem 4.8 is a random variable with a distribution q. If the deterministic J satisfying (E.1) and (F.7) is considered, the inequality (4.9) holds with a probability greater than $1 - 2e^{-\delta}$, which is the same probability obtained with the inequality in Theorem 2 of (Suzuki et al., 2020). The index J in Theorem 2 of (Suzuki et al., 2020) is chosen deterministically by minimizing the information losses (2) with the additional constraint $\sum_{j\in J}(\tau_j')^{-1} < \frac{5}{3}mm^\sharp$. This constraint can be interpreted as the leverage score τ_J' corresponding to J becomes larger, which implies that important nodes are selected from the spectral information of the covariance matrix $\widehat{\Sigma}$.

Remark G.2. In the case of $m > m_{\rm nzr}$, we can obtain a sharper error bound than (4.9) in Theorem 4.8. More precisely, the constant omitted in (4.9), which depends on the size m of $\widehat{f}$, can be improved to the constant depending on $m_{\rm nzr}$, not on m. In fact, when $m > m_{\rm nzr}$, let $\widehat{f}_{\rm nzr}$ be the network obtained by deleting the nodes corresponding to the non-zero rows of the covariance matrix $\widehat{\Sigma}$. By the same argument, replacing $\widehat{\Psi}_j(\widehat{f})$ with $\widehat{\Psi}_j(\widehat{f}_{\rm nzr})$ in the proof of Theorem 4.8, we can obtain Theorem 4.8 by replacing m by $m_{\rm nzr}$, which means that a sharper error bound can be obtained.

H Detailed configurations for training, pruning and fine-tuning

Employed architecture for the Pixel-MNIST classification task consists of a single IRNN layer and an output layer, while that for the PTB word level language modeling consists of an embedding layer, a single RNN layer and an output layer, where we can merge an embedding weight matrix and an RNN input weight matrix into an single weight matrix. The loss function is the cross entropy function following the soft-max function for both tasks. Each training and fine-tuning is optimized by Adam, and hyper-parameters obtained by grid search are summarized in Table 3, where "FT" means the parameter used in fine-tuning and "bptt" means the step size for back-propagation through time. As regards regularization techniques for the PTB task, we adopt the dropout, whose ratio is 0.1, in any case and the weight tying (Inan et al., 2016) in effective case.

We sample five models for each baseline in section 5. Furthermore, pruning methods including randomness are applied five times for each baseline model. Other detailed configurations for each method are the following:

- Baseline (128)
  - train:

Table 3: Hyper-parameters for learning.

Task	epochs (FT)	batch size	learning rate (FT)	LR decay (step)	gradient clip	bptt
Pixel-MNIST	500 (250)	120	10^{-4} (5^{-5})	0.95 (10)	1.0	784
PTB	200 (200)	20	5.0 (2.5)	0.95 (1)	0.01	35

        - hidden size: 128
        - weight tying: True

- Baseline (42)
    - train:
        - hidden size: 42
        - weight tying: True
    - prune:
        - None
    - finetune: (only PTB case)
        - hidden size: 42 (stay)
        - weight tying: False
- Spectral w/ rec. or w/o rec.
    - train:
        - Use Baseline (128)
    - prune:
        - size of hidden-to-hidden weight matrix: $16384 (= 128 \times 128) \to 1764 (= 42 \times 42)$
        - size of input-to-hidden weight matrix: $128 (= 1 \times 128) \to 42 (= 1 \times 42)$ (Pixel-MNIST) or $1270016 (= 9922 \times 128) \to 416724 (= 9922 \times 42)$ (PTB)
        - size of hidden-to-output weight matrix: $1280 (= 128 \times 10) \to 420 (= 42 \times 10)$ (Pixel-MNIST) or $1270016 (= 9922 \times 128) \to 416724 (= 9922 \times 42)$ (PTB)
        - Reduce the RNN weight matrices based on our proposed method with or without the reconstruction matrix
    - finetune:
        - hidden size: 42 (reduced from 128)
        - weight tying: False
- Random w/ rec. or w/o rec.
    - Same as "Spectral" except for reducing the RNN weight matrices randomly in pruning phase
- Column Sparsification
    - train:
        - hidden size: 128
        - weight tying: True
        - Mask the lowest $86 (= 128 - 42)$ columns of the hidden-to-hidden weight matrix by L^2-norm for each iteration (add noise to the weight matrix before masking when applied to the IRNN)
    - prune:
        - Fix the mask
    - finetune:
        - None
- Low Rank Factorization
    - train:
        - Use Baseline (128)

  - prune:
    - intrinsic parameters of hidden-to-hidden weight matrix: $16384 (= 128 \times 128) \to 10752 (= 128 \times 42 + 42 \times 128)$
    - Decompose hidden-to-hidden weight matrix based on SVD: $W = USV^\top \to W' = U[:, : 42]S[: 42]V^\top[: 42, :]$
    - Entry of S, which is singular values, are in descending order
  - finetune:
    - None

- Magnitude-based Weight
  - train:
    - Use Baseline (128)
  - prune:
    - parameters of hidden-to-hidden weight matrix: $16384 (= 128 \times 128) \to 1764 (= 42 \times 42)$
    - Remove the lowest $14620 (= 128 \times 128 - 42 \times 42)$ parameters by L^1-norm
  - finetune:
    - hidden size: 128 (stay but have sparse weight matrix)

- Random Weight
  - Same as "Magnitude-based Weight" except for removing parameters randomly in pruning phase