

---

# On Generalization Bounds of a Family of Recurrent Neural Networks

---

Minshuo Chen\* Xingguo Li† Tuo Zhao\*

\* Georgia Tech † Princeton University

## Abstract

Recurrent Neural Networks (RNNs) have been widely applied to sequential data analysis. Due to their complicated modeling structures, however, the theory behind is still largely missing. To connect theory and practice, we study the generalization properties of vanilla RNNs as well as their variants, including Minimal Gated Unit (MGU), Long Short Term Memory (LSTM), and Convolutional (Conv) RNNs. Specifically, our theory is established under the PAC-Learning framework. The generalization bound is presented in terms of the spectral norms of the weight matrices and the total number of parameters. We also establish refined generalization bounds with additional norm assumptions, and draw a comparison among these bounds. We remark: (1) Our generalization bound for vanilla RNNs is significantly tighter than the best of existing results; (2) We are not aware of any other generalization bounds for MGU and LSTM RNNs in the existing literature; (3) We demonstrate the advantages of these variants in generalization.

## 1 Introduction

Recurrent Neural Networks (RNNs) have successfully revolutionized sequential data analysis, and been widely applied to many real world problems, such as natural language processing (Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014), speech recognition (Graves et al., 2006; Mikolov et al., 2010; Graves, 2012; Graves et al., 2013), computer vision (Gregor et al., 2015; Xu et al., 2015; Donahue et al., 2015; Karpathy and Fei-Fei, 2015), healthcare (Lipton et al., 2015; Choi et al., 2016a,b), and robot control

(Lee and Teng, 2000; Yoo et al., 2006). Quite a few of these applications can be approached easily in our daily life, such as Google Translate and Apple Siri.

The sequential modeling nature of RNNs is significantly different from feedforward neural networks, though they both have neurons as the basic components. RNNs exploit the internal state (also known as hidden unit) to process the sequence of inputs, which naturally captures the dependence of the sequence. Besides the vanilla version, RNNs have many other variants. A large class of variants incorporate the so-called “gated” units to trim RNNs for different tasks. Typical examples include Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber (1997)), Gated Recurrent Unit (GRU, Jozefowicz et al. (2015)) and Minimal Gated Unit (MGU, Zhou et al. (2016)).

The success of RNNs owes not only to their special network structures and the ability to fit data, but also to their good generalization property: They provide accurate predictions on unseen data. For example, Graves et al. (2013) report that after training with merely 462 speech samples, deep LSTM RNNs achieve a test set error of 17.7% on TIMIT phoneme recognition benchmark, which is the best recorded score. Despite of the popularity of RNNs in applications, their theory is less studied than other feedforward neural networks (Haussler, 1992; Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2017; Li et al., 2018). There are still several long lasting fundamental questions regarding the approximation, trainability, and generalization of RNNs.

In this paper, we propose to understand the generalization ability of RNNs and their variants. We aim to answer two questions from a theoretical perspective:

**Q.1) Do RNNs suffer from significant curse of dimensionality?**

**Q.2) What are the advantages of MGU and LSTM over vanilla RNNs?**

The investigation of generalization properties of RNNs has a long history. Many early works are based on over-simplified assumptions. For example, Dasgupta and Sontag (1996) and Koiran (1998) adopt a VC-

dimension argument to show complexity bounds of RNNs that are polynomial in the size of the network. They, however, either consider linear RNNs for binary classification tasks, or assume RNNs take the first coordinate of their hidden states as outputs. More recently, Bartlett et al. (2017) propose a new technique for developing generalization bounds for feedforward neural networks based on empirical Rademacher complexity under the PAC-Learning framework. Neyshabur et al. (2017) further adapt the technique to establish their generalization bound using the PAC-Bayes approach. The follow-up work Zhang et al. (2018) use the PAC-Bayes approach to establish a generalization bound for vanilla RNNs.

We exploit the compositional nature of RNNs, and decouples the spectral norms of weight matrices and the number of weight parameters. This makes our analysis conceptually much simpler (e.g. avoid layer wise analysis), and also yields better generalization bound than Zhang et al. (2018).

Consider vanilla RNNs, we observe  $m$  sequences of data points  $(x_{i,t}, z_{i,t})_{t=1}^T$ , where  $x_{i,t} \in \mathbb{R}^{d_x}$  and the response  $z_{i,t} \in \mathcal{Z}$  for all  $t = 1, \dots, T$  and  $i = 1, \dots, m$ . Each sequence is drawn independently from some underlying distribution over  $\mathbb{R}^{d_x \times T} \times \mathcal{Z}$ . Extensions to dependent sequences are discussed in Section 7, however, note that data points  $(x_{i,t}, z_{i,t})$  can be dependent within a sequence, i.e., for a fixed  $i \in \{1, \dots, m\}$ . The vanilla RNNs compute  $h_{i,t}$  and  $y_{i,t}$  iteratively as follows,

$$h_{i,t} = \sigma_h(Uh_{i,t-1} + Wx_{i,t}), \quad \text{and} \quad y_{i,t} = \sigma_y(Vh_{i,t}),$$

where  $\sigma_y$  and  $\sigma_h$  are activation operators,  $h_{i,t} \in \mathbb{R}^{d_h}$  is the hidden state with  $h_{i,0} = 0$ , and  $U \in \mathbb{R}^{d_h \times d_h}$ ,  $V \in \mathbb{R}^{d_y \times d_h}$ , and  $W \in \mathbb{R}^{d_h \times d_x}$  are weight matrices. The activation operators  $\sigma_h$  and  $\sigma_y$  are entrywise, i.e.,  $\sigma_h([v_1, \dots, v_d]^\top) = [\sigma_h(v_1), \dots, \sigma_h(v_d)]^\top$ , and Lipschitz with parameters  $\rho_h$  and  $\rho_y$  respectively. We assume  $\sigma_h(\cdot) = \tanh(\cdot)$ ,  $\sigma_y(0) = 0$ , and  $\rho_y = 1$ . Extensions to general activations are given in Section 2.

**Our Contribution.** To establish the generalization bound, we need to define the ‘‘model complexity’’ of vanilla RNNs. In this paper, we adopt the empirical Rademacher complexity (ERC, see more details in Section 2), which has been widely used in the existing literature on PAC-Learning. For many nonparametric function classes, we often need complicated argument to upper bound their ERC. Our analysis, however, shows that we can upper bound the ERC of vanilla RNNs in a very simple manner by exploiting their Lipschitz continuity with respect to the model parameters, since they are essentially in parametric forms. More specifically, denote  $\mathcal{F}_t = \{f_t : \{x_1, \dots, x_t\} \mapsto y_t\}$

as the class of mappings from the first  $t$  inputs to the  $t$ -th output computed by vanilla RNNs. For a matrix  $A$ ,  $\|A\|_2$  denotes the spectral norm, and for a vector  $v$ ,  $\|v\|_2$  denotes the Euclidean norm. Define  $\frac{a^t - 1}{a - 1} = t$  for  $a = 1$ . Then, informally speaking, the ‘‘model complexity’’ of vanilla RNNs satisfies

$$\text{Complexity} = O\left(d\|V\|_2 \min\{\sqrt{d}, \|W\|_2 \frac{\|U\|_2^t - 1}{\|U\|_2 - 1}\} \times \sqrt{\log\left(t\sqrt{d} \frac{\|U\|_2^t - 1}{\|U\|_2 - 1}\right)}\right),$$

where  $d = \sqrt{d_x d_h + d_h^2 + d_h d_y}$ .

We then consider a new testing sequence  $(x_t, z_t)_{t=1}^T$ . The response sequence is computed by  $\tilde{z}_t = \phi(y_t)$ , for all  $t = 1, \dots, T$ , where  $\phi$  is a function mapping the output of vanilla RNNs to the response of interest. In practice, the function  $\phi$  varies across different data analysis tasks. For example, in sequence classification, we take  $\phi(y_t) = \text{argmax}_j [y_t]_j$ ; in regression, we take  $\phi(y_t) = y_t$ ; in density estimation, we can take  $\phi(y_t) = \text{softmax}(y_t)$ .

We further define a risk function that can unify different data analysis tasks. Specifically, let  $\mathcal{L}(\mathcal{A}(y, z))$  be a loss function, where  $\mathcal{A}(y, z)$  is a function taking the output  $y_t$  and the observed response  $z_t$  as inputs, and  $\mathcal{L}$  is chosen according to different tasks. Then we define the population risk for the  $t$ -th output as  $\mathcal{R}(f_t) = \mathbb{E}[\mathcal{L}(\mathcal{A}(y_t, z_t))]$ . Its empirical counterpart is similarly defined as  $\widehat{\mathcal{R}}(f_t) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\mathcal{A}(y_{i,t}, z_{i,t}))$ . Training RNNs is essentially minimizing the empirical risk  $\widehat{\mathcal{R}}(f_t)$ . Many applications can be formulated into this framework. For example, in classification, we take  $\mathcal{A} = -\mathcal{M}$  as the functional margin operator and  $\mathcal{L} = \mathcal{L}_\gamma$  as the ramp loss with  $\gamma$  being the margin value (see detailed definitions in Section 2); in regression, we take  $\mathcal{A}(y_t, z_t) = y_t - z_t$  and  $\mathcal{L}$  as the  $\ell_p$  loss for  $p \in \mathbb{Z}_+$ . We then give the generalization bound in the following statement.

**Theorem 1** (informal). Assume the input data space is bounded, i.e.,  $\|x\|_2 \leq 1$  and  $z \in \mathcal{Z}$  bounded. Suppose the mapping  $\mathcal{A}(y, z)$  is Lipschitz in  $y$ , and the loss function  $\mathcal{L}$  satisfies  $|\mathcal{L}(\mathcal{A}(y, z))| \leq B$  and is  $L$ -Lipschitz for any  $y$  computed by RNNs and  $z \in \mathcal{Z}$ . Given a collection of samples  $S = \{(x_{i,t}, z_{i,t})_{t=1}^T, i = 1, \dots, m\}$  and a new testing sequence  $(x_t, z_t)_{t=1}^T$ , with probability at least  $1 - \delta$  over  $S$ , for any  $f_t \in \mathcal{F}_t$  with integer  $t \leq T$ , we have,

$$\mathcal{R}(f_t) \leq \widehat{\mathcal{R}}(f_t) + \tilde{O}\left(\frac{L \times \text{Complexity}}{\sqrt{m}} + B\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

Please refer to Section 2 for a complete statement. Most of the aforementioned commonly used  $\mathcal{A}$  and  $\mathcal{L}$

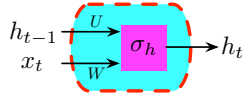


Figure 1: A building block of vanilla RNNs.

satisfy the assumptions in Theorem 1. For example, in classification, the functional margin operator  $\mathcal{M}(y, z)$  is 2-Lipschitz in  $\tilde{z}$ . The ramp loss  $\mathcal{L}_\gamma$  is uniformly bounded by 1 and  $\frac{1}{\gamma}$ -Lipschitz. In regression,  $\mathcal{A}(y, z)$  is 1-Lipschitz in  $y$  and bounded since the input data are bounded. Then the  $\ell_p$  loss becomes bounded and Lipschitz due to its bounded input.

**Comparison with Existing Results.** To better understand the obtained generalization bound and draw a comparison among existing literature, we instantiate Theorem 1 for sequence to sequence classification using vanilla RNNs. Recall that for classification tasks, we have  $L = 1/\gamma$ ,  $B = 1$  and  $\mathcal{M}(y, z)$  is 2-Lipschitz in  $y$ . We list the corresponding generalization bounds in Table 1 according to the magnitude of  $\|U\|_2$ .

Table 1: Generalization bounds for vanilla RNNs in classification tasks (we only list the order of the gap  $\mathcal{R}(f_t) - \widehat{\mathcal{R}}(f_t)$ ). The third column lists the result obtained in Zhang et al. (2018).

	Theorem 1	Zhang et al. (2018)
(I) $\ U\ _2 < 1$	$\tilde{O}(d/\sqrt{m\gamma})$	$\tilde{O}(dt^2/\sqrt{m\gamma})$
(II) $\ U\ _2 = 1$	$\tilde{O}(dt/\sqrt{m\gamma})$	$\tilde{O}(dt^2/\sqrt{m\gamma})$
(III) $\ U\ _2 > 1$	$\tilde{O}(\sqrt{d^3 t}/\sqrt{m\gamma})$	$\tilde{O}(dt^2\ U\ _2^t/\sqrt{m\gamma})$

As can be seen, the obtained generalization bound only has a polynomial dependence on the size of vanilla RNNs, i.e., width  $d$  and sequence length  $t$ . Thus, we theoretically justify that the complexity of vanilla RNNs do not suffer from significant curse of dimensionality. Because they compute outputs  $y_t$  recursively using the same weight matrices, and their hidden states  $h_t$  are entrywise bounded.

We compare Theorem 1 with the generalization bound obtained in Zhang et al. (2018), which is of the order  $\tilde{O}(dt^2\|W\|_2\|V\|_2\max\{1, \|U\|_2^t\}/\sqrt{m\gamma})$ , and we distinguish the same three different scenarios as listed in Table 1. Our bound is tighter by a factor of  $t^2$  for case (I), a factor of  $t$  for case (II). Additionally, Zhang et al. (2018) fail to incorporate the boundedness condition of hidden state into their analysis, thus the generalization bound is exponential in  $t$  for case (III). Our generalization bound, however, is still polynomial in  $d$  and  $t$  for case (III).

Moreover, (II) is closely related to a few recent results on imposing orthogonal constraints on weight matrices to stabilize the training of RNNs (Saxe et al., 2013; Le et al., 2015; Arjovsky et al., 2016; Vorontsov et al., 2017; Zhang et al., 2018). We remark that from a learning theory perspective, (II) implies that orthogonal constraints can potentially help generalization.

We also present refined generalization bounds with additional matrix norm assumptions. These assumptions allow us to derive norm-based generalization bounds.

We draw a comparison among these bounds and highlight their advantage under different scenarios.

Our theory can be further extended to several variants, including MGU and LSTM RNNs. Specifically, we show that the gated units in MGU and LSTM RNNs can introduce extra decaying factors to further reduce the dependence on  $d$  and  $t$  in generalization. Such an advantage in generalization makes these RNNs do not suffer from significant curse of dimensionality. To the best of our knowledge, these are the first results on generalization guarantees for these neural networks.

**Notations:** Given a vector  $v \in \mathbb{R}^d$ , we denote its Euclidean norm by  $\|v\|_2^2 = \sum_{i=1}^d |v_i|^2$ , and the infinity norm by  $\|v\|_\infty = \max_j |v_j|$ . Given a matrix  $M \in \mathbb{R}^{m \times n}$ , we denote the spectral norm by  $\|M\|_2$  as the largest singular value of  $M$ , the Frobenius norm by  $\|M\|_F^2 = \text{trace}(MM^\top)$ , and the (2, 1) norm by  $\|M\|_{2,1} = \sum_{i=1}^n \|M_{:,i}\|_2$ . Given a function  $f$ , we denote the function infinity norm by  $\|f\|_\infty = \sup |f|$ . We use  $\tilde{O}(\cdot)$  to denote  $O(\cdot)$  with hidden log factors. We denote  $a \wedge b = \min\{a, b\}$ .

## 2 Generalization of Vanilla RNNs

To establish the generalization bound, we start with imposing some mild assumptions.

**Assumption 1.** Input data are bounded, i.e.,  $\|x_{i,t}\|_2 \leq B_x$  for all  $i = 1, \dots, m$  and  $t = 1, \dots, T$ .

**Assumption 2.** The spectral norms of weight matrices are bounded respectively, i.e.,  $\|U\|_2 \leq B_U$ ,  $\|V\|_2 \leq B_V$ , and  $\|W\|_2 \leq B_W$ .

**Assumption 3.** Activation operators  $\sigma_h$  and  $\sigma_y$  are Lipschitz with parameters  $\rho_h$  and  $\rho_y$  respectively, and  $\sigma_h(0) = \sigma_y(0) = 0$ . Additionally,  $\sigma_h$  is entrywise bounded by  $b$ .

Assumptions 1 and 2 are moderate assumptions. Moreover, Assumption 3 holds for most commonly used activation operators, such as  $\sigma_h(\cdot) = \tanh(\cdot)$  and  $\sigma_y(\cdot) = \text{ReLU}(\cdot) = \max\{\cdot, 0\}$  (1-Lipschitz).

Recall vanilla RNNs compute  $h_{i,t}$  and  $y_{i,t}$  as follows,  $h_{i,t} = \sigma_h(Uh_{i,t-1} + Wx_{i,t})$  and  $y_{i,t} = \sigma_y(Vh_{i,t})$ , where  $U \in \mathbb{R}^{d_h \times d_h}$ ,  $V \in \mathbb{R}^{d_y \times d_h}$ , and  $W \in \mathbb{R}^{d_h \times d_x}$ . We consider multiclass classification tasks with the label  $z \in \mathcal{Z} = \{1, \dots, K\}$ . Given a sequence  $(x_t, z_t)_{t=1}^T$ , we define  $X_t \in \mathbb{R}^{d_x \times t}$  by concatenating  $x_1, \dots, x_t$  as columns of  $X_t$ . Recall that we denote  $\mathcal{F}_t = \{f_t : X_t \mapsto y_t\}$  as the class of mappings from the first  $t$  inputs to the  $t$ -th output computed by vanilla RNNs.

As previously mentioned, we define the functional margin for the  $t$ -th output in vanilla RNNs as

$$\mathcal{M}(f_t(X_t), z_t) = [f_t(X_t)]_{z_t} - \max_{j \neq z_t} [f_t(X_t)]_j.$$

We further define a ramp loss  $\mathcal{L}_\gamma(-\mathcal{M}(f_t(X_t), z_t)) : \mathbb{R} \mapsto \mathbb{R}^+$  to each margin, where  $\mathcal{L}_\gamma$  is a piecewise linear function defined as

$$\mathcal{L}_\gamma(a) = \mathbb{1}\{a > 0\} + (1 + a/\gamma)\mathbb{1}\{-\gamma \leq a \leq 0\},$$

where  $\mathbb{1}\{A\}$  denotes the indicator function of a set  $A$ . Accordingly, the ramp risk is defined as  $\mathcal{R}_\gamma(f_t) = \mathbb{E}[\mathcal{L}_\gamma(-\mathcal{M}(f_t(X_t), z_t))]$ , and its empirical counterpart is defined as  $\widehat{\mathcal{R}}_\gamma(f_t) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}_\gamma(-\mathcal{M}(f_t(X_{i,t}), z_{i,t}))$ . We then present the formal statement of Theorem 1.

**Theorem 2.** Let activation operators  $\sigma_h$  and  $\sigma_y$  be given, and Assumptions 1–3 hold. Then for  $(x_t, z_t)_{t=1}^T$  and  $S = \{(x_{i,t}, z_{i,t})_{t=1}^T, i = 1, \dots, m\}$  drawn i.i.d. from any underlying distribution over  $\mathbb{R}^{d_x \times T} \times \{1, \dots, K\}$ , with probability at least  $1 - \delta$  over  $S$ , for every margin value  $\gamma > 0$  and every  $f_t \in \mathcal{F}_t$  for integer  $t \leq T$ , we have

$$\mathbb{P}(\tilde{z}_t \neq z_t) \leq \widehat{\mathcal{R}}_\gamma(f_t) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + O\left(\frac{d\rho_y B_V \lambda_t \sqrt{\log(t\sqrt{dm} \frac{\beta^t - 1}{\beta - 1})}}{\sqrt{m}\gamma}\right), \quad (1)$$

where  $d = \sqrt{d_x d_h + d_h^2 + d_h d_y}$ ,  $\beta = \rho_h B_U$ , and  $\lambda_t = \min\{b\sqrt{d}, \rho_h B_x B_W \frac{\beta^t - 1}{\beta - 1}\}$ .

**Remark 1.** To ease the presentation, we only provide the generalization bound for the classification task. Extensions to general tasks are straightforward by replacing functions  $\mathcal{A}$  and  $\mathcal{L}$  and substituting suitable values of  $L$  and  $B$ .

The generalization bound depends on the total number of weights, and the range of  $\rho_h B_U$  in three cases as indicated in Section 1. More precisely, if  $\rho_h B_U \lesssim (1 + \frac{1}{\alpha})$  for constant  $\alpha > 0$  bounded away from zero, the generalization bound is of the order  $\tilde{O}\left(\frac{dt^\alpha}{\sqrt{m}\gamma}\right)$ , which has a polynomial dependence on  $d$  and  $t$ . As can be seen, with proper normalization on model parameters, the model complexity of vanilla RNNs do not suffer from significant curse of dimensionality.

We also highlight a tradeoff between generalization and representation of vanilla RNNs. As can be seen, when  $\rho_h B_U$  is strictly smaller than 1, the generalization bound is nearly independent on  $t$ . The hidden state, however, only has limited representation ability, since its magnitude diminishes as  $t$  grows large. On the contrary, when  $\rho_h B_U$  is strictly greater than 1, the representation ability is amplified but the generalization becomes worse. As a consequence, recent empirical results show that imposing extra constraints or regularization, such as  $U^\top U = I$  or  $\|U\|_2 \leq 1$  (Saxe et al., 2013; Le et al., 2015; Arjovsky et al., 2016; Vorontsov et al., 2017; Zhang et al., 2018), helps balance the generalization and representation of RNNs.

### 3 Proof of Main Results

Our analysis is based on the PAC-learning framework. Due to space limit, we only present an outline of our proof. More technical details are deferred to Appendix A. Before we proceed, we first define the empirical Rademacher complexity as follows.

**Definition 1** (Empirical Rademacher Complexity). Let  $\mathcal{H}$  be a function class and  $S = \{s_1, \dots, s_m\}$  be a collection of samples. The empirical Rademacher complexity of  $\mathcal{H}$  given  $S$  is defined as

$$\mathfrak{R}_S(\mathcal{H}) = \mathbb{E}_\epsilon \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \epsilon_i h(s_i) \right],$$

where  $\epsilon_i$ 's are i.i.d. Rademacher random variables, i.e.,  $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 0.5$ .

We then proceed with our analysis. Recall that Mohri et al. (2012) give an empirical Rademacher complexity (ERC)-based generalization bound, which is restated in the following lemma with  $\mathcal{F}_{\gamma,t} = \{(X_t, z_t) \mapsto \ell_\gamma(-\mathcal{M}(f_t(X_t), z_t)) : f_t \in \mathcal{F}_t\}$ .

**Lemma 1.** Given a testing sequence  $(x_t, z_t)_{t=1}^T$ , with probability at least  $1 - \delta$  over samples  $S = \{(x_{i,t}, z_{i,t})_{t=1}^T, i = 1, \dots, m\}$ , for every margin value  $\gamma > 0$  and any  $f_t \in \mathcal{F}_t$ , we have

$$\mathcal{R}_\gamma(f_t) \leq \widehat{\mathcal{R}}_\gamma(f_t) + 2\mathfrak{R}_S(\mathcal{F}_{\gamma,t}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Note that Lemma 1 adapts the original version (Theorem 3.1, Chapter 3.1, Mohri et al. (2012)) for the multiclass ramp loss, and we have  $\mathbb{P}(\tilde{z}_t \neq z_t) \leq \mathcal{R}_\gamma(f_t)$  by definition.

Now we only need to bound the ERC  $\mathfrak{R}_S(\mathcal{F}_{\gamma,t})$ . Our analysis consists of three steps. First, we characterize the Lipschitz continuity of vanilla RNNs w.r.t model parameters. Next, we bound the covering number of function class  $\mathcal{F}_t$ . At last, we derive an upper bound on  $\mathfrak{R}_S(\mathcal{F}_{\gamma,t})$  via the standard machinery in the PAC-learning framework. Specifically, consider two different sets of weight matrices  $(U, V, W)$  and  $(U', V', W')$ . Given the same activation operators and input data, denote the  $t$ -th output as  $y_t$  and  $y'_t$  respectively. We characterize the Lipschitz property of  $\|y_t\|_2$  w.r.t model parameters in the following lemma.

**Lemma 2.** Under Assumptions 1–3, given input  $(x_t)_{t=1}^T$  and for any integer  $t \leq T$ ,  $\|y_t\|_2$  is Lipschitz in  $U, V$  and  $W$ , i.e.,

$$\|y_t - y'_t\|_2 \leq L_{U,t} \|U - U'\|_F + L_{V,t} \|V - V'\|_F + L_{W,t} \|W - W'\|_F,$$

where  $L_{U,t} = \rho_h B_V B_W t a_t$ ,  $L_{V,t} = B_W a_t$ , and  $L_{W,t} = B_V a_t$  with  $a_t = \rho_y \rho_h B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}$ .



The detailed proof is provided in Appendix A.2. We give a simple example to illustrate the proof technique. Specifically, we consider a single layer network that outputs  $y = \sigma(Wx)$ , where  $x$  is the input,  $\sigma$  is an activation operator with Lipschitz parameter  $\rho$ , and  $W$  is a weight matrix. Such a network is Lipschitz in both  $x$  and  $W$  as follows. Given weight matrices  $W$  and  $W'$ , we have

$$\|y - y'\|_2 = \|\sigma(Wx) - \sigma(W'x)\|_2 \leq \rho \|x\|_2 \|W - W'\|_F.$$

Additionally, given inputs  $x$  and  $x'$ , we have

$$\|y - y'\|_2 = \|\sigma(Wx) - \sigma(Wx')\|_2 \leq \rho \|W\|_2 \|x - x'\|_2.$$

Since vanilla RNNs are multilayer networks, Lemma 2 can be obtained by telescoping.

We remark that Lemma 2 is the key to the proof of our generalization bound, which separates the spectral norms of weight matrices and the total number of parameters.

Next, we bound the covering number of  $\mathcal{F}_t$ . Denote by  $\mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot))$  the minimal cardinality of a subset  $\mathcal{C} \subset \mathcal{F}_t$  that covers  $\mathcal{F}_t$  at scale  $\epsilon$  w.r.t the metric  $\text{dist}(\cdot, \cdot)$ , such that for any  $f_t \in \mathcal{F}_t$ , there exists  $\hat{f}_t \in \mathcal{C}$  satisfying  $\text{dist}(f_t, \hat{f}_t) = \sup_{X_t} \|f_t(X_t) - \hat{f}_t(X_t)\|_2 \leq \epsilon$ . The following lemma gives an upper bound on  $\mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot))$ .

**Lemma 3.** Under Assumptions 1–3, given any  $\epsilon > 0$ , the covering number of  $\mathcal{F}_t$  satisfies

$$\mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot)) \leq \left(1 + \frac{6c\sqrt{dt}((\rho_h B_U)^t - 1)}{\epsilon(\rho_h B_U - 1)}\right)^{3d^2},$$

where  $c = \rho_y \rho_h B_V B_W B_x \max\{1, \rho_h B_U\}$ .

The detailed proof is provided in Appendix A.3. We briefly explain the proof technique. Given activation operators, since vanilla RNNs are in parametric forms,  $f_t$  has a one-to-one correspondence to its weight matrices  $U, V$ , and  $W$ . Lemma 2 implies that  $\text{dist}(\cdot, \cdot)$  is controlled by the Frobenius norms of the differences of weight matrices. Thus, it suffices to bound the covering numbers of three weight matrices. The product of covering numbers of three weight matrices gives us Lemma 3.

Lastly, we give an upper bound on  $\mathfrak{R}_S(\mathcal{F}_{\gamma,t})$  in the following lemma.

**Lemma 4.** Under Assumptions 1–3, given activation operators and samples  $S = \{(x_{i,t}, z_{i,t})_{i=1}^T, i = 1, \dots, m\}$ , the empirical Rademacher complexity  $\mathfrak{R}_S(\mathcal{F}_{\gamma,t})$  satisfies

$$\mathfrak{R}_S(\mathcal{F}_{\gamma,t}) = O\left(d \min\left\{b\sqrt{d}, \rho_h B_x B_W \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}\right\} \times \frac{\rho_y B_V \sqrt{\log(t\sqrt{dm} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1})}}{\sqrt{m}\gamma}\right).$$

The detailed proof is provided in Appendix A.4. Our proof exploits the Lipschitz continuity of  $\mathcal{M}$  and  $\ell_\gamma$ , and uses Dudley’s entropy integral as the standard machinery to establish Lemma 4. Combining Lemma 1 and Lemma 4, we complete the proof.

## 4 Refined Generalization Bounds

When additional norm constraints on weight matrices  $U, V$  and  $W$  are available, we can further refine generalization bounds. Specifically, we consider assumptions as follows.

**Assumption 4.** The weight matrices satisfy  $\|U\|_{2,1} \leq M_U$ ,  $\|V\|_{2,1} \leq M_V$ , and  $\|W\|_{2,1} \leq M_W$ .

**Assumption 5.** The weight matrices satisfy  $\|U\|_F \leq B_{U,F}$ ,  $\|V\|_F \leq B_{V,F}$ , and  $\|W\|_F \leq B_{W,F}$ .

Note that Assumption 4 appears in Bartlett et al. (2017) and Assumption 5 appears in Neyshabur et al. (2017). We have an equivalent relation between matrix norms, i.e.,  $\|\cdot\|_2 \leq \|\cdot\|_{2,1} \leq \sqrt{d}\|\cdot\|_F \leq d\|\cdot\|_2$ . Comparing to Assumption 2, Assumptions 4 and 5 further restrict the model class. We then establish refined empirical Rademacher complexities for vanilla RNNs, the corresponding generalization bounds follows immediately.

**Theorem 3.** Let activation operators  $\sigma_h$  and  $\sigma_y$  be given, and Assumptions 1–3 hold. Then for  $(x_t, z_t)_{t=1}^T$  and  $S = \{(x_{i,t}, z_{i,t})_{i=1}^T, i = 1, \dots, m\}$  drawn i.i.d. from any underlying distribution over  $\mathbb{R}^{d_x \times T} \times \{1, \dots, K\}$ , with probability at least  $1 - \delta$  over  $S$ , for every margin value  $\gamma > 0$  and every  $f_t \in \mathcal{F}_t$  for integer  $t \leq T$ , the following two bounds hold:

- Suppose Assumption 4 also holds. We have

$$\mathfrak{R}_S(\mathcal{F}_{\gamma,t}) = O\left(\frac{t\alpha S_{2,1} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \sqrt{\log d} \log(\sqrt{dm})}{\sqrt{m}\gamma}\right), \quad (2)$$

where  $\alpha = \rho_h^2 \rho_y B_V B_W B_x$ ,  $S_{2,1} = M_U + M_V + M_W$ , and  $d = \sqrt{d_x d_h + d_h^2} + d_h d_y$ .

- Suppose Assumption 5 also holds. We have

$$\mathfrak{R}_S(\mathcal{F}_{\gamma,t}) = O\left(\frac{\alpha' B_U \lambda_t S_F \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \sqrt{d} \log d}{\sqrt{m}\gamma}\right), \quad (3)$$

where  $\alpha' = \rho_h \rho_y B_W B_x$ ,  $S_F = B_{U,F} + B_{W,F} + B_{V,F}$ ,  $\lambda_t = \min\{b\sqrt{d}, \rho_h B_x B_W \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}\}$ , and  $d = \sqrt{d_x d_h + d_h^2} + d_h d_y$ .

The detailed proof is provided in Appendix B.1. The first bound (2) adapts the matrix covering lemma in Bartlett et al. (2017). The second bound (3) adapts the PAC-Bayes approach (Neyshabur et al., 2017) by

analyzing the divergence when imposing small perturbations on the weight matrices.

We highlight the improvements of the obtained refined generalization bounds: When the weight matrices are approximately low rank, that is,  $\|\cdot\|_{2,1} \ll d\|\cdot\|_2$  and  $\|\cdot\|_F \ll \sqrt{d}\|\cdot\|_2$ , for  $\beta \leq 1$ , bound (3) improves bound (1) by reducing dependence on  $d$ . Additionally, if  $t(M_U + M_V + M_W) < d$ , bound (2) also tightens bound (1). Note that  $t(M_U + M_V + M_W) < d$  implies that the input sequence is relatively short.

## 5 Extensions to MGU, LSTM, and Conv RNNs

We extend our analysis to Minimal Gated Unit (MRU), Long Short-Term Memory (LSTM) RNNs, and convolutional RNNs (Conv RNNs).

The MGU RNNs compute

$$\begin{aligned} r_t &= \sigma(W_r x_t + U_r h_{t-1}), \\ \tilde{h}_t &= \sigma_h(W_h x_t + U_h(r_t \odot h_{t-1})), \\ h_t &= (1 - r_t) \odot h_{t-1} + r_t \odot \tilde{h}_t, \end{aligned}$$

where  $W_r, W_h \in \mathbb{R}^{d_h \times d_x}$ ,  $U_r, U_h \in \mathbb{R}^{d_h \times d_h}$ ,  $V \in \mathbb{R}^{d_y \times d_h}$ , and  $r_t \in \mathbb{R}^{d_h}$ . The notation  $\odot$  denotes the Hadamard product (entrywise product) of vectors. Denote by  $\mathcal{F}_{g,t}$  the class of mappings from the first  $t$  inputs to the  $t$ -th output computed by gated (MGU or LSTM) RNNs. For simplicity, we consider  $\sigma$  being the sigmoid function, i.e.,  $\sigma(x) = (1 + \exp(-x))^{-1}$ ,  $\sigma_h(\cdot) = \tanh(\cdot)$ , and  $\sigma_y$  being  $\rho_y$ -Lipschitz with  $\sigma_y(0) = 0$ . Extensions to general Lipschitz activation operators as in Assumption 3 are straightforward. Suppose we have  $h_0 = 0$  and the following assumption.

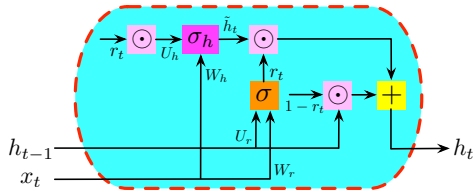


Figure 2: A building block of MGU RNNs.

**Assumption 6.** All the weight matrices have bounded spectral norms respectively, i.e.  $\|W_r\|_2 \leq B_{W_r}$ ,  $\|W_h\|_2 \leq B_{W_h}$ ,  $\|U_r\|_2 \leq B_{U_r}$ ,  $\|U_h\|_2 \leq B_{U_h}$ , and  $\|V\|_2 \leq B_V$ .

A similar argument for vanilla RNNs yields a generalization bound of MGU RNNs as follows.

**Theorem 4.** Let the activation operator  $\sigma_y$  be given and Assumptions 1 and 6 hold. Then for  $(x_t, z_t)_{t=1}^T$  and  $S = \{(x_{i,t}, z_{i,t})_{t=1}^T, i = 1, \dots, m\}$  drawn i.i.d. from any underlying distribution over  $\mathbb{R}^{d_x \times T} \times \{1, \dots, K\}$ , with probability at least  $1 - \delta$  over  $S$ , for

every margin value  $\gamma > 0$  and every  $f_t \in \mathcal{F}_{g,t}$  for integer  $t \leq T$ , we have

$$\begin{aligned} \mathbb{P}(\tilde{z}_t \neq z_t) &\leq \hat{\mathcal{R}}_\gamma(f_t) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \\ &+ O\left(\frac{d\rho_y B_V(\sqrt{d} \wedge B_{W_h} B_x \frac{\beta^t - 1}{\beta - 1})\sqrt{\log(\frac{\theta^t - 1}{\theta - 1} d\sqrt{m})}}{\sqrt{m}\gamma}\right), \end{aligned}$$

where  $\beta = \max_j \{\|1 - r_j\|_\infty + B_{U_h} \|r_j\|_\infty^2\}$ ,  $\theta = \beta + 2B_{U_r} + B_{U_r} B_{U_h}$ ,  $d = \max\{d_x, d_y, d_h\}$ .

The detailed proof is provided in Appendix C.1. As can be seen,  $r_t$  shrinks the magnitude of hidden state to reduce the dependence on  $d$  and  $t$  in generalization. As a result, with proper normalization of weight matrices, the generalization bound of MGU RNNs is less dependent on  $d, t$ .

The LSTM RNNs are more complicated than MGU RNNs, which introduce more gates to control the information flow in RNNs. LSTM RNNs have two hidden states, and compute them as,

$$\begin{aligned} g_t &= \sigma(W_g x_t + U_g h_{t-1}), & r_t &= \sigma(W_r x_t + U_r h_{t-1}), \\ o_t &= \sigma(W_o x_t + U_o h_{t-1}), & \tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1}), \\ c_t &= g_t \odot c_{t-1} + r_t \odot \tilde{c}_t, & h_t &= o_t \odot \tanh(c_t), \end{aligned}$$

where  $W_g, W_r, W_o, W_c \in \mathbb{R}^{d_h \times d_x}$ ,  $U_g, U_r, U_o, U_c \in \mathbb{R}^{d_h \times d_h}$ , and  $g_t, r_t, o_t \in \mathbb{R}^{d_h}$ . For simplicity, we also consider  $\sigma$  being the sigmoid function, and  $\sigma_c(\cdot) = \tanh(\cdot)$ . The  $t$ -th output is  $y_t = \sigma_y(V h_t)$ , where  $V \in \mathbb{R}^{d_y \times d_h}$ , and  $\sigma_y$  is  $\rho_y$ -Lipschitz with  $\sigma_y(0) = 0$ . Suppose we have  $h_0 = c_0 = 0$  and the following assumption.

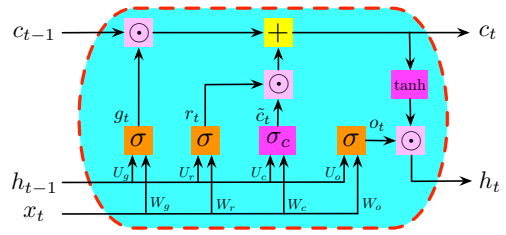


Figure 3: A building block of LSTM RNNs.

**Assumption 7.** The spectral norms of weight matrices are bounded respectively, i.e.  $\|W_g\|_2 \leq B_{W_g}$ ,  $\|W_r\|_2 \leq B_{W_r}$ ,  $\|W_o\|_2 \leq B_{W_o}$ ,  $\|W_c\|_2 \leq B_{W_c}$ ,  $\|U_g\|_2 \leq B_{U_g}$ ,  $\|U_r\|_2 \leq B_{U_r}$ ,  $\|U_o\|_2 \leq B_{U_o}$ ,  $\|U_h\|_2 \leq B_{U_h}$ , and  $\|V\|_2 \leq B_V$ .

For properly normalized weight matrices  $W_o$  and  $U_o$ , the generalization bound of LSTM RNNs is given in the following theorem.

**Theorem 5.** Let the activation operator  $\sigma_y$  be given and Assumptions 1 and 7 hold. Then for  $(x_t, z_t)_{t=1}^T$  and  $S = \{(x_{i,t}, z_{i,t})_{t=1}^T, i = 1, \dots, m\}$  drawn i.i.d.

from any underlying distribution over  $\mathbb{R}^{d_x \times T} \times \{1, \dots, K\}$ , with probability at least  $1 - \delta$  over  $S$ , for every margin value  $\gamma > 0$  and every  $f_t \in \mathcal{F}_{g,t}$  for integer  $t \leq T$ , we have

$$\mathbb{P}(\tilde{z}_t \neq z_t) \leq \widehat{\mathcal{R}}_\gamma(f_t) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + O\left(\frac{d\rho_y B_V (\sqrt{d} \wedge B_{W_c} B_x \frac{\beta^t - 1}{\beta - 1}) \sqrt{\log(\frac{\theta^t - 1}{\theta - 1} d\sqrt{m})}}{\sqrt{m}\gamma}\right),$$

where  $\beta = \max\{\|g_j\|_\infty + B_{U_c} \|r_j\|_\infty \|o_j\|_\infty\}$ ,  $\theta = \beta + B_{U_g} + B_{U_r} + B_{U_o}$ , and  $d = \max\{d_x, d_y, d_h\}$ .

The detailed proof is provided in Appendix C.2. Similar to MGU RNNs, LSTM RNNs also introduce extra decaying factors to reduce the dependence on  $d$  and  $t$  in generalization. However, LSTM RNNs are more complicated, but more flexible than MGU RNNs, since three factors,  $r_t$ ,  $o_t$  and  $g_t$  are used to jointly control the spectrum of  $U_c$ . We further remark that LSTM RNNs need spectral norms of weight matrices,  $W_g, W_r, W_o, U_g, U_r$  and  $U_o$ , to be properly controlled for obtaining better generalization bounds.

We further extend our analysis to Convolutional RNNs (Conv RNNs). Conv RNNs integrate convolutional filters and recurrent neural networks. Specifically, we consider input  $x \in \mathbb{R}^d$  and  $k$ -channel  $k$ -dimensional convolutional filters  $\mathcal{I}_1, \dots, \mathcal{I}_k \in \mathbb{R}^k$  followed by an average pooling layer over the  $k$  channels for reducing dimensionality. Extensions to convolution with strides and other kinds of average pooling layers (e.g., block-wise pooling) are straightforward.

Here we denote the circulant-like matrix generated by  $\mathcal{I}_i$  as

$$C_i = \begin{bmatrix} \mathcal{I}_i^\top & \underbrace{0 \dots \dots 0}_{d-k} \\ 0 & \mathcal{I}_i^\top & \underbrace{0 \dots \dots 0}_{d-k-1} \\ & & \ddots \\ \underbrace{0 \dots \dots 0}_{d-k} & & & \mathcal{I}_i^\top \end{bmatrix} \in \mathbb{R}^{(d-k+1) \times d},$$

and write  $W_{\mathcal{I}} = [C_1^\top, \dots, C_k^\top]^\top$ . We further denote  $P = \frac{1}{k} [\underbrace{I_{d-k+1} \ I_{d-k+1} \ \dots \ I_{d-k+1}}_{\text{totally } k \text{ identity matrices}}]$ , where  $I_d$  de-

notes the  $d$ -dimensional identity matrix. Define  $\mathcal{I} = [\mathcal{I}_1, \dots, \mathcal{I}_k]$ , and  $\mathcal{I} * x = P W_{\mathcal{I}} x$ . Given a sample  $(x_t, z_t)_{t=1}^T$ , the Conv RNNs compute  $h_t$  and  $y_t$  as follows,

$h_t = \sigma_h(\mathcal{U} * h_{t-1} + \mathcal{W} * x_t)$ , and  $y_t = \sigma_y(\mathcal{V} * h_t)$ , where  $h_t, x_t \in \mathbb{R}^d$ , and  $\mathcal{U}, \mathcal{V}, \mathcal{W} \in \mathbb{R}^{k \times k}$  are matrices with column vectors being  $k$ -dimensional convolutional filters. We use zero-padding to ensure the output dimension of convolutional filters matches the input (Krizhevsky et al., 2012). To get  $y_t$ , we convolve  $h_t$

with  $\mathcal{V}$  followed by an average pooling to reduce the dimension to  $K$ . Since we aim to show that Conv RNNs reduce the dependence on  $d$  in generalization through parameter sharing, we simplify the notations to assume  $h_0 = 0$ , and impose the following assumption. Extensions to general settings are straightforward.

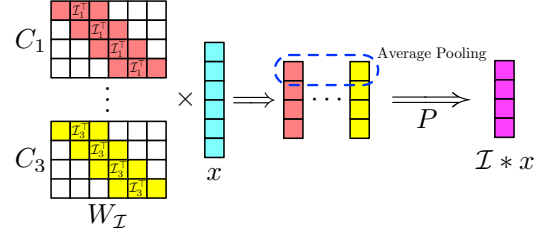


Figure 4: Illustration of input  $x \in \mathbb{R}^6$  convolving with 3-channel 3-dimensional convolutional filters  $\mathcal{I}_1, \mathcal{I}_2$ , and  $\mathcal{I}_3$ , followed by an average pooling.

**Assumption 8.** The activation operators  $\sigma_h$  and  $\sigma_y$  are 1-Lipschitz with  $\sigma_h(0) = \sigma_y(0) = 0$ .  $\sigma_h$  is entrywise bounded by 1. The convolutional filters  $\mathcal{U}, \mathcal{V}$ , and  $\mathcal{W}$  are orthogonal with normalized columns, i.e.,  $\mathcal{U}^\top \mathcal{U} = \mathcal{U} \mathcal{U}^\top = \frac{1}{k} I_k$ ,  $\mathcal{V}^\top \mathcal{V} = \mathcal{V} \mathcal{V}^\top = \frac{1}{k} I_k$ , and  $\mathcal{W}^\top \mathcal{W} = \mathcal{W} \mathcal{W}^\top = \frac{1}{k} I_k$ .

We remark that the orthogonality constraints enhance the diversity among convolutional filters (Xie et al., 2017; Huang et al., 2017). Additionally, the normalization factor  $\frac{1}{k}$  is to control the spectral norms of  $W_{\mathcal{U}}, W_{\mathcal{V}}$ , and  $W_{\mathcal{W}}$ , which prevents the blowup of hidden state. Denote by  $\mathcal{F}_{c,t}$  the class of mappings from the first  $t$  inputs to the  $t$ -th output computed by Conv RNNs. Then the generalization bound is given in the following theorem.

**Theorem 6.** Let activation operators  $\sigma_h$  and  $\sigma_y$  be given, and Assumptions 1 and 8 hold. Then for  $(x_t, z_t)_{t=1}^T$  and  $S = \{(x_{i,t}, z_{i,t})_{t=1}^T, i = 1, \dots, m\}$  drawn i.i.d. from any underlying distribution over  $\mathbb{R}^{d \times T} \times \{1, \dots, K\}$ , with probability at least  $1 - \delta$  over  $S$ , for every margin value  $\gamma > 0$  and every  $f_t \in \mathcal{F}_{c,t}$  for integer  $t \leq T$ , we have

$$\mathbb{P}(\tilde{z}_t \neq z_t) \leq \widehat{\mathcal{R}}_\gamma(f_t) + O\left(\frac{kt\sqrt{\log(dt m)}}{\sqrt{m}\gamma} + \sqrt{\frac{\log \frac{1}{\delta}}{m}}\right).$$

The detailed proof is provided in C.3. Similar to the analysis of vanilla RNNs, our proof is based on the Lipschitz continuity of Conv RNNs with respect to its model parameters in the convolutional filters. Specifically, by Assumption 8, the spectral norms of  $W_{\mathcal{U}}, W_{\mathcal{V}}$ , and  $W_{\mathcal{W}}$  are all bounded by 1. Combining with the inequality,  $\|W_{\mathcal{U}}\|_F \leq \sqrt{d}\|\mathcal{U}\|_F$ , we have  $\|y_t - y'_t\|_2 \leq L_{V,t}\|\mathcal{V} - \mathcal{V}'\|_F + L_{U,t}\|\mathcal{U} - \mathcal{U}'\|_F + L_{W,t}\|\mathcal{W} - \mathcal{W}'\|_F$ , where  $L_{U,t}, L_{V,t}$ , and  $L_{W,t}$  are polynomials in  $d$  and  $t$ . Additionally, observe that the total number of parameters in a Conv RNN is at most  $3k^2$ , which is independent of input dimension  $d$ . As a consequence, the

generalization bound of Conv RNNs only has a linear dependence on  $k$  and  $t$ .

## 6 Numerical Evaluation

We demonstrate a comparison among our obtained generalization bound with Bartlett et al. (2017), Neyshabur et al. (2017), and Zhang et al. (2018). Specifically, we train<sup>1</sup> a vanilla RNN on the wikitext language modeling dataset (Merity et al., 2016). We take  $\sigma_h = \tanh$  and set the hidden state  $h \in \mathbb{R}^{128}$  and the input  $x \in \mathbb{R}^{14}$  with  $\|x\|_2 \leq 1$ . Accordingly, we have  $d = 128$  and take the sequence length  $t = 56$ . We list the complexity bounds for vanilla RNNs in Theorem 2 (Ours), Zhang et al. (2018) (Bound 1), (2) of Theorem 3 (Bound 2), and (3) of Theorem 3 (Bound 3) neglecting common log factors in  $d$  and  $t$ :

- Ours:  $dB_V \min \left\{ \sqrt{d}, B_W \frac{B_U^t - 1}{B_U - 1} \right\} \sqrt{\log \left( \frac{B_U^t - 1}{B_U - 1} \right)}$ ;
- Bound 1:  $dt^2 B_V B_W \max \{1, B_U^t\}$ ;
- Bound 2:  $B_V B_W (M_U + M_V + M_W) t \frac{B_U^t - 1}{B_U - 1}$ ;
- Bound 3:  $\left( \min \left\{ \sqrt{d}, B_W \frac{B_U^t - 1}{B_U - 1} \right\} B_U + B_W \right) \frac{B_U^t - 1}{B_U - 1} \times \sqrt{d(B_{U,F}^2 + B_{W,F}^2 + B_{V,F}^2)}$ .

The corresponding complexity bounds are shown in Figure 5. As can be seen, our complexity bound in Theorem 2 is much smaller than Bounds 1-3. In more detail, the trained vanilla RNN has  $B_U = 2.6801 > 1$ . As discussed earlier, for  $B_U > 1$ , only our bound in Theorem 2 is polynomial in the size of the network, while Bounds 1-3 are all exponential in  $t$ . The resulting complexity bounds corroborate such a conclusion.

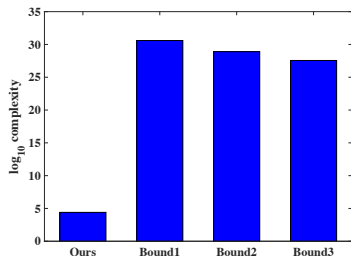


Figure 5: Complexity bounds on wikitext dataset.

We also observe that Bound 3 is smaller than Bound 2. The reason behind is that the weight matrices in the trained vanilla RNN have relatively small Frobenius norms but large  $(2, 1)$  norms. Taking matrix  $U$  as an example, we have  $B_{U,F} = 13.6823$  and  $M_U = 154.5439$ . Then, we can calculate the stable rank  $\frac{B_{U,F}}{B_U} = 5.1 < \sqrt{d}/2$ , and the ratio  $\frac{M_U}{B_{U,F}} = 11.3 \approx \sqrt{d}$ . This implies that the singular values of  $U$  are not evenly distributed, while the norms of row vectors in  $U$  are approximately equal.

<sup>1</sup>We adopt the code: [https://github.com/pytorch/examples/tree/master/word\\_language\\_model](https://github.com/pytorch/examples/tree/master/word_language_model).

## 7 Discussions and Open Problems

**(I) Tighter bounds:** Our obtained generalization bounds depend on the spectral norms of weight matrices and the network size. Can we exploit other modeling structures to further reduce the dependence on the network size? Or can we find better choices of norms of weight matrices that yield better bounds?

**(II) Margin value:** Our generalization bounds depend on the margin value of the predictors. As can be seen, a larger margin value yields a better generalization bound. However, establishing a sharp characterization of the margin value is technically very challenging, because of its complicated dependence on the underlying data distribution and the training algorithm.

**(III) Implicit bias of SGD:** Numerous empirical evidences have already shown that RNNs trained by stochastic gradient descent (SGD) algorithms have superior generalization performance. There have been a few theoretical results showing that SGD tends to yield low complexity models, which can generalize (Neyshabur et al., 2014, 2015; Zhang et al., 2016; Soudry et al., 2017). Can we extend this argument to RNNs? For example, can SGD always yield weight matrices with well controlled spectra? This is crucial to the generalization of MGU and LSTM RNNs.

**(IV) Adaptivity to the underlying distribution:** The current PAC-Learning framework focuses on the worst case. Taking classification as an example, the theoretical analysis holds even when the input features and labels are completely independent. Therefore, this often yields very pessimistic results. For many real applications, however, data are not obtained adversarially. Some recent empirical evidences suggest that the generalization of neural networks seems very adaptive to the underlying distribution: Easier tasks lead to low complexity neural networks, while harder ones lead to highly complex neural networks. Unfortunately, none of the existing analysis can take the underlying distribution into consideration.

**(V) Sequentially dependent data:** To extend the analysis to scenarios where input sequences are dependent is quite challenging and largely open. Rakhlin et al. (2015) propose a so-called “Sequential Rademacher Complexity” to quantify the model complexity with dependent data. Their bound however, is exponential in the depth of a neural network, even with proper normalization on the weight matrices. Kuznetsov and Mohri (2017) also derive generalization bounds for dependent data under mixing conditions. They assume block independence for a sub-sample selection trick. The extension to fully dependent data is beyond the scope of this paper. We leave it for future investigation.



## References

- ARJOVSKY, M., SHAH, A. and BENGIO, Y. (2016). Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*.
- BAHDANAU, D., CHO, K. and BENGIO, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- BARTLETT, P. L., FOSTER, D. J. and TELGARSKY, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*.
- CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. and BENGIO, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- CHOI, E., BAHADORI, M. T., SCHUETZ, A., STEWART, W. F. and SUN, J. (2016a). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*.
- CHOI, E., SCHUETZ, A., STEWART, W. F. and SUN, J. (2016b). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, **24** 361–370.
- DASGUPTA, B. and SONTAG, E. D. (1996). Sample complexity for learning recurrent perceptron mappings. In *Advances in Neural Information Processing Systems*.
- DONAHUE, J., ANNE HENDRICKS, L., GUADARRAMA, S., ROHRBACH, M., VENUGOPALAN, S., SAENKO, K. and DARRELL, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- GOLOWICH, N., RAKHLIN, A. and SHAMIR, O. (2017). Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*.
- GRAVES, A. (2012). Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.
- GRAVES, A., FERNÁNDEZ, S., GOMEZ, F. and SCHMIDHUBER, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. ACM.
- GRAVES, A., MOHAMED, A.-R. and HINTON, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*. IEEE.
- GREGOR, K., DANIHELKA, I., GRAVES, A., REZENDE, D. J. and WIERSTRA, D. (2015). Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.
- HAUSSLER, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, **100** 78–150.
- HOCHREITER, S. and SCHMIDHUBER, J. (1997). Long short-term memory. *Neural computation*, **9** 1735–1780.
- HUANG, L., LIU, X., LANG, B., YU, A. W. and LI, B. (2017). Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. *arXiv preprint arXiv:1709.06079*.
- JOZEFOWICZ, R., ZAREMBA, W. and SUTSKEVER, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*.
- KARPATY, A. and FEI-FEI, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- KOIRAN, P. (1998). Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, **86** 63–79.
- KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.
- KUZNETSOV, V. and MOHRI, M. (2017). Generalization bounds for non-stationary mixing processes. *Machine Learning*, **106** 93–117.
- LE, Q. V., JAITLY, N. and HINTON, G. E. (2015). A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.
- LEE, C.-H. and TENG, C.-C. (2000). Identification and control of dynamic systems using recurrent fuzzy neural networks. *IEEE Transactions on fuzzy systems*, **8** 349–366.
- LI, X., LU, J., WANG, Z., HAUPT, J. and ZHAO, T. (2018). On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159*.
- LIPTON, Z. C., KALE, D. C., ELKAN, C. and WETZEL, R. (2015). Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*.
- MERITY, S., XIONG, C., BRADBURY, J. and SOCHER, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

- MIKOLOV, T., KARAFIÁT, M., BURGET, L., ČERNOCKÝ, J. and KHUDANPUR, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.
- MOHRI, M., ROSTAMIZADEH, A. and TALWALKAR, A. (2012). *Foundations of machine learning*. MIT press.
- NEYSHABUR, B., BHOJANAPALLI, S., MCALESTER, D. and SREBRO, N. (2017). A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*.
- NEYSHABUR, B., SALAKHUTDINOV, R. R. and SREBRO, N. (2015). Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*.
- NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.
- RAKHLIN, A., SRIDHARAN, K. and TEWARI, A. (2015). Online learning via sequential complexities. *The Journal of Machine Learning Research*, **16** 155–186.
- SAXE, A. M., MCCLELLAND, J. L. and GANGULI, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.
- SOUDRY, D., HOFFER, E. and SREBRO, N. (2017). The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*.
- SUTSKEVER, I., VINYALS, O. and LE, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.
- VORONTSOV, E., TRABELSI, C., KADOURY, S. and PAL, C. (2017). On orthogonality and learning recurrent networks with long term dependencies. *arXiv preprint arXiv:1702.00071*.
- XIE, D., XIONG, J. and PU, S. (2017). All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. *arXiv preprint arXiv:1703.01827*.
- XU, K., BA, J., KIROS, R., CHO, K., COURVILLE, A., SALAKHUTDINOV, R., ZEMEL, R. and BENGIO, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.
- YOO, S. J., PARK, J. B. and CHOI, Y. H. (2006). Adaptive dynamic surface control of flexible-joint robots using self-recurrent wavelet neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **36** 1342–1355.
- ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.
- ZHANG, J., LEI, Q. and DHILLON, I. S. (2018). Stabilizing gradients for deep neural networks via efficient svd parameterization. *arXiv preprint arXiv:1803.09327*.
- ZHOU, G.-B., WU, J., ZHANG, C.-L. and ZHOU, Z.-H. (2016). Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, **13** 226–234.

## A Proofs in Section 2

### A.1 Lipschitz Continuity of $\mathcal{M}$ and $\ell_\gamma$

We show the Lipschitz continuity of the margin operator  $\mathcal{M}$  and the loss function  $\ell_\gamma$  in the following lemma.

**Lemma 5.** The margin operator  $\mathcal{M}$  is 2-Lipschitz in its first argument with respect to vector Euclidean norm, and  $\ell_\gamma$  is  $\frac{1}{\gamma}$ -Lipschitz.

*Proof.* Let  $y, y'$  and  $z$  be given, then

$$\begin{aligned} \left| \mathcal{M}(y, z) - \mathcal{M}(y', z) \right| &= \left| y_z - y'_z + \left( \max_{j \neq z} y'_j - \max_{j \neq z} y_j \right) \right| \\ &\leq \left| y_z - y'_z \right| + \left| \max_{j \neq z} y'_j - y_j \right| \\ &\leq 2 \|y - y'\|_\infty \leq \|y - y'\|_2. \end{aligned}$$

For function  $\ell_\gamma$ , it is a piecewise linear function. Thus, it is straightforward to see that  $\ell_\gamma$  is  $\frac{1}{\gamma}$ -Lipschitz.  $\square$

### A.2 Proof of Lemma 2

*Proof.* The Lemma is stated with matrix Frobenius norms. However, we can show a tighter bound only involving the spectral norms of weight matrices. Given weight matrices  $U, V, W$  and  $U', V', W'$ , consider the  $t$ -th outputs  $y_t$  and  $y'_t$  of vanilla RNNs,

$$\begin{aligned} \|y_t - y'_t\|_2 &= \|\sigma_y(Vh_t) - \sigma_y(V'h'_t)\|_2 \\ &\leq \rho_y \|Vh_t - V'h'_t + V'h'_t - V'h_t\|_2 \\ &\leq \rho_y (\|(V - V')h_t\|_2 + \|V'(h_t - h'_t)\|_2) \\ &\leq \rho_y (\|h_t\|_2 \|V - V'\|_2 + B_V \|h_t - h'_t\|_2). \end{aligned} \quad (4)$$

We have to bound the norm of  $h_t$  as in the following lemma.

**Lemma 6.** Under Assumptions 1 to 3, for  $t \geq 0$ , the norm of  $h_t$  is bounded by

$$\|h_t\|_2 \leq \min \left\{ b\sqrt{d}, \rho_h B_W B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \right\}. \quad (5)$$

*Proof.* We prove by induction. Observe that for  $t \geq 1$ , we have

$$\begin{aligned} \|h_t\|_2 &= \|\sigma_h(Wx_t + Uh_{t-1})\|_2 \\ &\leq \rho_h \|Wx_t + Uh_{t-1}\|_2 \\ &\leq \rho_h (\|Wx_t\|_2 + \|Uh_{t-1}\|_2) \\ &\leq \rho_h (B_W B_x + B_U \|h_{t-1}\|_2). \end{aligned} \quad (6)$$

Applying equation (6) recursively with  $h_0 = 0$ , we arrive at,

$$\|h_t\|_2 \leq \rho_h B_W B_x \sum_{j=0}^{t-1} (\rho_h B_U)^j = \rho_h B_W B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1},$$

We also have  $\|h_t\|_\infty \leq b$ . Thus, combining with the above upper bound, we get  $\|h_t\|_2 \leq \min \left\{ b\sqrt{d}, \rho_h B_W B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \right\}$ . Clearly,  $\|h_0\|_2 = 0$  satisfies the upper bound.  $\square$

When  $\rho_h B_U = 1$ , the ratio is defined, by L'Hospital's rule, to be the limit,

$$\lim_{\rho_h B_U \rightarrow 1} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} = t.$$

With Lemma 6 in hand, we plug the bound (5) into equation (4) and end up with

$$\|y_t - y'_t\|_2 \leq \rho_y \rho_h B_W B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \|V - V'\|_2 + \rho_y B_V \|h_t - h'_t\|_2. \quad (7)$$

The remaining task is to bound  $\|h_t - h'_t\|_2$  in terms of the spectral norms of the difference of weight matrices,  $\|W - W'\|_2$  and  $\|U - U'\|_2$ .

**Lemma 7.** Under Assumptions 1 to 3, for  $t \geq 1$ , the difference of hidden states  $h_t$  and  $h'_t$  satisfies

$$\|h_t - h'_t\|_2 \leq L_{W,t} \|W - W'\|_2 + L_{U,t} \|U - U'\|_2,$$

where  $L_{W,t} = \rho_h B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}$  and  $L_{U,t} = \rho_h^2 B_W B_x t \frac{(\rho_h B_U)^t - 1}{(\rho_h B_U) - 1}$ .

*Proof.* Similar to the proof of Lemma 6, we use induction.

$$\begin{aligned} \|h_t - h'_t\|_2 &= \|\sigma_h(Wx_t + Uh_{t-1}) - \sigma_h(W'x_t + U'h'_{t-1})\|_2 \\ &\leq \rho_h \|(W - W')x_t + Uh_{t-1} - U'h'_{t-1}\|_2 \\ &\leq \rho_h (\|(W - W')x_t\|_2 + \|Uh_{t-1} - U'h'_{t-1}\|_2) \\ &\leq \rho_h (B_x \|W - W'\|_2 + \|Uh_{t-1} - U'h'_{t-1} + U'h'_{t-1} - U'h'_{t-1}\|_2) \\ &\leq \rho_h B_x \|W - W'\|_2 + \rho_h (\|h_{t-1}\|_2 \|U - U'\|_2 + B_U \|h_{t-1} - h'_{t-1}\|_2). \end{aligned}$$

Repeat this derivation recursively, we have

$$\begin{aligned} \|h_t - h'_t\|_2 &\leq \rho_h B_x \|W - W'\|_2 + \rho_h \|h_{t-1}\|_2 \|U - U'\|_2 + \rho_h B_U \|h_{t-1} - h'_{t-1}\|_2 \\ &\leq \rho_h B_x (1 + \rho_h B_U) \|W - W'\|_2 + \rho_h (\|h_{t-1}\|_2 + \rho_h B_U \|h_{t-2}\|_2) \|U - U'\|_2 \\ &\quad + (\rho_h B_U)^2 \|h_{t-2} - h'_{t-2}\|_2 \\ &\leq \dots \\ &\leq \rho_h B_x \sum_{j=0}^{t-1} (\rho_h B_U)^j \|W - W'\|_2 + \rho_h \sum_{j=0}^{t-1} ((\rho_h B_U)^{t-1-j} \|h_j\|_2) \|U - U'\|_2 \\ &\quad + (\rho_h B_U)^t \|h_0 - h'_0\|_2 \\ &\leq \rho_h B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \|W - W'\|_2 + \rho_h \sum_{j=0}^{t-1} ((\rho_h B_U)^{t-1-j} \|h_j\|_2) \|U - U'\|_2. \end{aligned} \quad (8)$$

We now plug in the upper bound (5) to calculate the summation involving the Euclidean norms of the hidden state  $h_t$ .

$$\begin{aligned} \sum_{j=0}^{t-1} (\rho_h B_U)^{t-1-j} \|h_j\|_2 &\leq \sum_{j=0}^{t-1} (j+1) (\rho_h B_U)^j \rho_h B_W B_x \leq t \sum_{j=0}^{t-1} (\rho_h B_U)^j \rho_h B_W B_x \\ &\leq \rho_h B_W B_x t \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}. \end{aligned}$$

Plugging back into equation (8), we have as desired,

$$\|h_t - h'_t\|_2 \leq \rho_h B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \|W - W'\|_2 + \rho_h^2 B_W B_x t \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \|U - U'\|_2. \quad \square$$

Combining equation (7) and Lemma 7, and  $\|W\|_F \geq \|W\|_2$ , we immediately get Lemma 2.  $\square$

### A.3 Proof of Lemma 3

*Proof.* Our goal is to construct a covering  $\mathcal{C}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot))$ , i.e., for any  $f_t \in \mathcal{F}_t$ , there exists  $\hat{f}_t \in \mathcal{F}_t$ , for any input data  $(x_t)_{t=1}^T$ , satisfying

$$\sup_{X_t} \|f_t(X_t) - \hat{f}_t(X_t)\|_2 \leq \epsilon.$$

Note that  $f$  is determined by weight matrices  $U, V$  and  $W$ . By Lemma 2, we have

$$\sup_{X_t} \|f(X_t) - \hat{f}_t(X_t)\|_2 \leq L_{V,t} \|V - \hat{V}\|_F + L_{W,t} \|W - \hat{W}\|_F + L_{U,t} \|U - \hat{U}\|_F.$$

Then it is enough to construct three matrix coverings,  $\mathcal{C}\left(U, \frac{\epsilon}{3L_{U,t}}, \|\cdot\|_F\right)$ ,  $\mathcal{C}\left(V, \frac{\epsilon}{3L_{V,t}}, \|\cdot\|_F\right)$  and  $\mathcal{C}\left(W, \frac{\epsilon}{3L_{W,t}}, \|\cdot\|_F\right)$ . Their Cartesian product gives us the covering  $\mathcal{C}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot))$ . The following lemma gives an upper bound on the covering number of matrices with a bounded Frobenius norm.



**Lemma 8.** Let  $\mathcal{G} = \{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_2 \leq \lambda\}$  be the set of matrices with bounded spectral norm and  $\epsilon > 0$  be given. The covering number  $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F)$  is upper bounded by

$$\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F) \leq \left(1 + 2 \frac{\min\{\sqrt{d_1}, \sqrt{d_2}\} \lambda}{\epsilon}\right)^{d_1 d_2}.$$

*Proof.* For any matrix  $A \in \mathcal{G}$ , we define a mapping  $\phi : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}^{d_1 d_2}$ , such that  $\phi(A) = [A_{:,1}^\top, A_{:,2}^\top, \dots, A_{:,h}^\top]^\top$ , where  $A_{:,i}$  denotes the  $i$ -th column of matrix  $A$ . Denote the vector space induced by the mapping  $\phi$  by  $\mathcal{V}(\mathcal{G}) = \{\phi(A) : A \in \mathcal{G}\}$ . Note that we have  $\|A\|_F^2 = \sum_{i=1}^h A_{:,i}^\top A_{:,i} = \|\phi(A)\|_2^2$  and the mapping  $\phi$  is one-to-one and onto. By definition, the square of Frobenius norm equals the square of sum of singular values and the spectral norm is the largest singular value. Hence, the equivalence of Frobenius norm and spectral norm is given by the following inequalities,

$$\|A\|_2 \leq \|A\|_F \leq \min\{\sqrt{d_1}, \sqrt{d_2}\} \|A\|_2.$$

Now, we see that if we construct a covering  $\mathcal{C}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2)$ , then  $\phi^{-1}\mathcal{C}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2) = \{\phi^{-1}(v) : v \in \mathcal{C}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2)\}$  is a covering of  $\mathcal{G}$  at scale  $\epsilon$  with respect to the matrix Frobenius norm. Therefore, we get  $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F) \leq \mathcal{N}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2)$ . As a consequence, it suffices to upper bound the covering number of  $\mathcal{V}(\mathcal{G})$ . In order to do so, we need another closely related concept, packing number.

**Definition 2 (Packing).** Let  $\mathcal{G}$  be an arbitrary set and  $\epsilon > 0$  be given. We say  $\mathcal{P}(\mathcal{G}, \epsilon, \|\cdot\|)$  is a packing of  $\mathcal{G}$  at scale  $\epsilon$  with respect to the norm  $\|\cdot\|$ , if for any two elements  $A, B \in \mathcal{P}$ , we have

$$\|A - B\| > \epsilon.$$

Denote by  $\mathcal{M}(\mathcal{G}, \epsilon, \|\cdot\|)$  the maximal cardinality of  $\mathcal{P}(\mathcal{G}, \epsilon, \|\cdot\|)$ .

By the maximality, we can check that  $\mathcal{N}(\mathcal{C}, \epsilon, \|\cdot\|) \leq \mathcal{M}(\mathcal{C}, \epsilon, \|\cdot\|)$ . Indeed, let  $\mathcal{P}^*(\mathcal{G}, \epsilon, \|\cdot\|)$  be a maximal packing. Suppose there exists  $A \in \mathcal{G}$  such that for any  $B \in \mathcal{P}^*(\mathcal{G}, \epsilon, \|\cdot\|)$ , the inequality  $\|A - B\| > \epsilon$  holds. Then we can add  $A$  to  $\mathcal{P}^*(\mathcal{G}, \epsilon, \|\cdot\|)$ , while still keeping it being a packing, which contradicts the maximality of  $\mathcal{P}^*(\mathcal{G}, \epsilon, \|\cdot\|)$ . Thus, we have  $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|) \leq \mathcal{M}(\mathcal{G}, \epsilon, \|\cdot\|)$ .

Observe that  $\mathcal{V}(\mathcal{G})$  is contained in an Euclidean ball  $\mathcal{B}(0; R) \in \mathbb{R}^{d_1 d_2}$  of radius at most

$$R = \max_{A \in \mathcal{G}} \|\phi(A)\|_2 \leq \min\{\sqrt{d_1}, \sqrt{d_2}\} \|A\|_2 \leq \min\{\sqrt{d_1}, \sqrt{d_2}\} \lambda.$$

Additionally, the union of Euclidean balls  $\mathcal{B}(v; \epsilon/2) \subset \mathbb{R}^{d_1 d_2}$  with radius  $\epsilon/2$  and center  $v \in \mathcal{P}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2)$  is further contained in an Euclidean ball  $\mathcal{B}(0; R_\epsilon)$  of slightly enlarged radius  $R_\epsilon = \min\{\sqrt{d_1}, \sqrt{d_2}\} \lambda + \epsilon/2$ . Those balls  $\mathcal{B}(v; \epsilon/2)$  are disjoint by the definition of packing, thus we have

$$\begin{aligned} \mathcal{N}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2) &\leq \mathcal{P}(\mathcal{V}(\mathcal{G}), \epsilon, \|\cdot\|_2) \leq \frac{\text{vol}(\mathcal{B}(0, R_\epsilon))}{\text{vol}(\mathcal{B}(v; \epsilon/2))} = \left(\frac{R_\epsilon}{\epsilon/2}\right)^{d_1 d_2} \\ &= \left(1 + 2 \frac{\min\{\sqrt{d_1}, \sqrt{d_2}\} \lambda}{\epsilon}\right)^{d_1 d_2}, \end{aligned}$$

where  $\text{vol}(\cdot)$  denotes the volume. □

By Lemma 8, we can directly write out the upper bounds on the covering numbers of weight matrices,

$$\begin{aligned} \mathcal{N}\left(U, \frac{\epsilon}{3L_{U,t}}, \|\cdot\|_F\right) &\leq \left(1 + 6 \frac{\sqrt{d_h} B_U L_{U,t}}{\epsilon}\right)^{d_h^2}, \\ \mathcal{N}\left(V, \frac{\epsilon}{3L_{V,t}}, \|\cdot\|_F\right) &\leq \left(1 + 6 \frac{\min\{\sqrt{d_y}, \sqrt{d_h}\} B_V L_{V,t}}{\epsilon}\right)^{d_y d_h}, \quad \text{and} \\ \mathcal{N}\left(W, \frac{\epsilon}{3L_{W,t}}, \|\cdot\|_F\right) &\leq \left(1 + 6 \frac{\min\{\sqrt{d_x}, \sqrt{d_h}\} B_W L_{W,t}}{\epsilon}\right)^{d_x d_h}. \end{aligned}$$

Then we immediately have,

$$\begin{aligned} \mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot)) &\leq \mathcal{N}\left(U, \frac{\epsilon}{3L_{U,t}}, \|\cdot\|_F\right) \times \mathcal{N}\left(V, \frac{\epsilon}{3L_{V,t}}, \|\cdot\|_F\right) \times \mathcal{N}\left(W, \frac{\epsilon}{3L_{W,t}}, \|\cdot\|_F\right) \\ &\leq \left(1 + \frac{6\sqrt{d_h}B_U L_{U,t}}{\epsilon}\right)^{d_h^2} \left(1 + \frac{\min\{6\sqrt{d_y}, \sqrt{d_h}\}B_V L_{V,t}}{\epsilon}\right)^{d_y d_h} \\ &\quad \times \left(1 + \frac{6\min\{\sqrt{d_x}, \sqrt{d_h}\}B_W L_{W,t}}{\epsilon}\right)^{d_x d_h}. \end{aligned}$$

Substituting the coefficients  $L_{U,t}$ ,  $L_{V,t}$  and  $L_{W,t}$  from Lemma 2, we get

$$\begin{aligned} \mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot)) &\leq \left(1 + \frac{6\sqrt{d}\rho_y\rho_h B_V B_W B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\epsilon}\right)^{2d^2} \left(1 + \frac{6\sqrt{d}\rho_y\rho_h^2 B_U B_V B_W B_x t \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\epsilon}\right)^{d^2} \\ &\leq \left(1 + \frac{6c\sqrt{dt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\epsilon}\right)^{3d^2}, \end{aligned}$$

where  $c = \rho_y\rho_h B_V B_W B_x \max\{1, \rho_h B_U\}$ . For future usage, we also write down for small  $\epsilon > 0$ , such that  $\frac{6c\sqrt{dt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\epsilon} > 1$ , the logarithm of covering number satisfies,

$$\log \mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot)) \leq 3d^2 \log \left( \frac{12c\sqrt{dt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\epsilon} \right).$$

□

#### A.4 Proof of Lemma 4

*Proof.* Define  $\mathcal{F}_{\mathcal{M},t} = \{(X_t, z_t) \mapsto \mathcal{M}(f_t(X_t), z_t) : f_t \in \mathcal{F}_t\}$ . By Lemma 5, we see that  $\mathcal{M}$  is 2-Lipschitz in its first argument. In order to cover  $\mathcal{F}_{\mathcal{M},t}$  at scale  $\epsilon$ , it suffices to cover  $\mathcal{F}_t$  at scale  $\frac{\epsilon}{2}$ . This immediately gives us the covering number  $\mathcal{N}(\mathcal{F}_{\mathcal{M},t}, \epsilon, \|\cdot\|_\infty) \leq \mathcal{N}(\mathcal{F}_t, \epsilon/2, \text{dist}(\cdot, \cdot))$ .

We then give the statement of Dudley's entropy integral.

**Lemma 9.** Let  $\mathcal{H}$  be a real-valued function class taking values in  $[-r, r]$  for some constant  $r$ , and assume that  $0 \in \mathcal{H}$ . Let  $S = (s_1, \dots, s_m)$  be given points, then

$$\mathfrak{R}_S(\mathcal{H}) \leq \inf_{\alpha > 0} \left( \frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_\alpha^{2r\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{H}, \epsilon, \|\cdot\|)} d\epsilon \right).$$

The proof can be found in Bartlett et al. (2017). Taking  $\mathcal{H} = \mathcal{F}_{\mathcal{M},t}$ , we can easily verify that  $\mathcal{F}_{\mathcal{M},t}$  takes values in  $[-r, r]$  with  $r = \rho_y B_V \|h_t\|_2 \leq \rho_y B_V \min\left\{b\sqrt{d}, \rho_h B_W B_x \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}\right\}$  and  $0 \in \mathcal{F}_{\mathcal{M},t}$ . Thus, directly applying Lemma 9 yields the following bound,

$$\mathfrak{R}_S(\mathcal{F}_{\mathcal{M},t}) \leq \inf_{\alpha > 0} \left( \frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_\alpha^{2r\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{\mathcal{M},t}, \epsilon, \|\cdot\|_\infty)} d\epsilon \right).$$

We bound the integral as follows,

$$\begin{aligned} \int_\alpha^{2r\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_{\mathcal{M},t}, \epsilon, \|\cdot\|_\infty)} d\epsilon &\leq \int_\alpha^{2r\sqrt{m}} \sqrt{3d^2 \log \left( \frac{24c\sqrt{dt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\epsilon} \right)} d\epsilon \\ &\leq 2r\sqrt{m} \sqrt{3d^2 \log \left( \frac{24c\sqrt{dt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}}{\alpha} \right)}. \end{aligned}$$

Picking  $\alpha = \frac{1}{\sqrt{m}}$  is enough to give us an upper bound on  $\mathfrak{R}_S(\mathcal{F}_{\mathcal{M},t})$ ,

$$\mathfrak{R}_S(\mathcal{F}_{\mathcal{M}}) \leq \frac{4}{m} + \frac{24}{\sqrt{m}} \sqrt{3d^2 r^2 \log \left( 24c\sqrt{dmt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \right)}.$$

Finally, by Talagrand's lemma (Mohri et al., 2012) and  $\mathcal{L}_\gamma$  being  $\frac{1}{\gamma}$ -Lipschitz, we have

$$\mathfrak{R}_S(\mathcal{F}_{\gamma,t}) \leq \frac{1}{\gamma} \mathfrak{R}_S(\mathcal{F}_{\mathcal{M},t}) \leq \frac{4}{m\gamma} + \frac{24}{\sqrt{m\gamma}} \sqrt{3d^2 r^2 \log \left( 24c\sqrt{dmt} \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \right)}.$$

□

## B Proof in Section 4

### B.1 Proof of Theorem 3

*Proof.* Under additional Assumption 4, we only need to show that, with the additional matrix induced norm bound, we have a refined upper bound on the matrix covering number. The proof relies on the following lemma adapted from Bartlett et al. (2017) Lemma 3.2.

**Lemma 10.** Let  $\mathcal{G} = \{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_{2,1} \leq \lambda\}$ . We have the following matrix covering upper bound

$$\log \mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_2) \leq \frac{\lambda^2}{\epsilon^2} \log(2d_1 d_2).$$

The above Lemma is a direct consequence of Lemma 3.2 in Bartlett et al. (2017) with  $X$  being identity,  $a = \lambda$ ,  $b = 1$ , and  $m = d_1, d = d_2$ . We apply the same trick to split the overall covering accuracy  $\epsilon$  into 3 parts,  $\frac{\epsilon}{3L_{U,t}}$ ,  $\frac{\epsilon}{3L_{V,t}}$ , and  $\frac{\epsilon}{3L_{W,t}}$ , corresponding to  $U, V, W$  respectively. Then we derive a refined bound on the covering number of  $\mathcal{F}_t$ :

$$\log \mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot, \cdot)) \leq \frac{9(M_U L_{U,t}^2 + M_V L_{V,t}^2 + M_W L_{W,t}^2)}{\epsilon^2} \log(2d^2), \quad (9)$$

where  $d = \max\{d_x, d_y, d_h\}$ . Substituting (9) into the Dudley integral as in the proof of Lemma 4 yields

$$\mathfrak{R}_S(\mathcal{F}_{\mathcal{M},t}) \leq \inf_{\alpha > 0} \left( \frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_{\alpha}^{2r\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_t, \epsilon/2, \|\cdot\|_{\infty})} d\epsilon \right).$$

We bound the integral as follows,

$$\begin{aligned} \int_{\alpha}^{2r\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_t, \epsilon/2, \|\cdot\|_{\infty})} d\epsilon &\leq \int_{\alpha}^{2r\sqrt{m}} 36 \frac{\sqrt{M_U L_{U,t}^2 + M_V L_{V,t}^2 + M_W L_{W,t}^2}}{\epsilon} \sqrt{\log(2d^2)} d\epsilon \\ &= 36 \sqrt{M_U L_{U,t}^2 + M_V L_{V,t}^2 + M_W L_{W,t}^2} \sqrt{\log(2d^2)} \log \frac{2r\sqrt{m}}{\alpha}. \end{aligned}$$

Choosing  $\alpha = \frac{1}{\sqrt{m}}$  yields

$$\mathfrak{R}_S(\mathcal{F}_{\mathcal{M}}) \leq \frac{4}{m} + \frac{432}{\sqrt{m}} \sqrt{M_U L_{U,t}^2 + M_V L_{V,t}^2 + M_W L_{W,t}^2} \sqrt{\log(2d^2)} \log(2m\sqrt{d}).$$

Finally, substituting the Lipschitz constant  $L_{U,t}, L_{V,t}, L_{W,t}$  into the expression, we have

$$\begin{aligned} \mathfrak{R}_S(\mathcal{F}_{\gamma,t}) &\leq \frac{1}{\gamma} \mathfrak{R}_S(\mathcal{F}_{\mathcal{M},t}) \leq \frac{4}{m\gamma} + \frac{432}{\gamma\sqrt{m}} \sqrt{M_U L_{U,t}^2 + M_V L_{V,t}^2 + M_W L_{W,t}^2} \sqrt{\log(2d^2)} \log(2m\sqrt{d}) \\ &\leq O \left( \frac{\alpha \max\{M_U, M_V, M_W\} t^{\frac{\beta t - 1}{\beta - 1}}}{\gamma\sqrt{m}} \sqrt{\log d} \log(m\sqrt{d}) \right). \end{aligned}$$

Combining with Lemma 1 completes the proof.

Under additional Assumption 5, our proof is based on the following result from Lemma 1 in Neyshabur et al. (2017).

**Lemma 11.** Let  $f_\alpha(x) : \mathcal{X} \rightarrow \mathbb{R}^d$  be any predictor with parameter  $\alpha$ , and  $\mathcal{P}$  be any distribution on the parameter that is independent of training data. Then, for any  $\gamma, \delta > 0$ , with probability at least  $1 - \delta$  over the training set of size  $m$ , for any  $\alpha$  and any random perturbation  $\beta$  s.t.  $\mathbb{P}_\beta [\max_{x \in \mathcal{X}} |f_{\alpha+\beta}(x) - f_\alpha(x)|_\infty < \frac{\gamma}{4}] \geq \frac{1}{2}$ , we have

$$\mathcal{R}_0(f_\alpha) - \widehat{\mathcal{R}}_\gamma(f_\alpha) \leq 4\sqrt{\frac{\text{KL}(\alpha + \beta \| \mathcal{P}) + \log\left(\frac{6m}{\delta}\right)}{m-1}},$$

where  $\text{KL}(\alpha + \beta \| \mathcal{P})$  is KL divergence of distributions  $\alpha + \beta$  and  $\mathcal{P}$ .

For convenience, we omit the superscript for sample index. Denote  $h_t(\alpha)$  and  $h_t(\alpha + \beta)$  as the hidden variables with parameters  $\alpha$  and  $\alpha + \beta$  respectively. Then we provide an upper bound of the gap of hidden layers before and after the perturbation. Denote the parameters  $\alpha = \text{vec}(\{W, U, V\})$  and the perturbation  $\beta = \text{vec}(\{\delta W, \delta U, \delta V\})$ .

For any  $t \in \{1, 2, \dots, T\}$ , we have

$$\begin{aligned} & \|h_t(\alpha + \beta) - h_t(\alpha)\|_2 \\ & \stackrel{(i)}{\leq} \rho_h \|(U + \delta U)h_{t-1}(\alpha + \beta) + (W + \delta W)x_t - Uh_{t-1}(\alpha) - Wx_t\|_2 \\ & \stackrel{(ii)}{\leq} \rho_h B_U \|h_{t-1}(\alpha + \beta) - h_{t-1}(\alpha)\|_2 + \delta \rho_h B_U \|h_{t-1}(\alpha + \beta)\|_2 + \delta \rho_h B_x B_W \\ & \leq (\rho_h B_U)^t \|h_0(\alpha + \beta) - h_0(\alpha)\|_2 + \delta \sum_{i=1}^t (\rho_h B_U)^i \|h_{t-i}(\alpha + \beta)\|_2 + \delta \rho_h B_x B_W \sum_{i=0}^{t-1} (\rho_h B_U)^i, \end{aligned} \quad (10)$$

By Lemma 6, we have that for any  $t \leq T$ ,

$$\|h_t(\alpha)\|_2 \leq \min \left\{ b\sqrt{p}, \rho_h B_x B_W \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} \right\} = \lambda_t. \quad (11)$$

Combining (10), (11), and  $h_0 = 0$ , we have

$$\begin{aligned} \|h_t(\alpha + \beta) - h_t(\alpha)\|_2 & \leq \delta \lambda_t \sum_{i=1}^t (\rho_h B_U)^i + \delta \rho_h B_x B_W \sum_{i=0}^{t-1} (\rho_h B_U)^i \\ & \leq \delta (\lambda_t \rho_h B_U + \rho_h B_x B_W) \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1}. \end{aligned} \quad (12)$$

Denote  $y_t(\alpha)$  and  $y_t(\alpha + \beta)$  as the out with parameters  $\alpha$  and  $\alpha + \beta$  respectively. Then we have

$$\begin{aligned} \|y_t(\alpha + \beta) - y_t(\alpha)\|_2 & \stackrel{(i)}{\leq} \rho_y \|(1 + \delta)Vh_t(\alpha + \beta) - Vh_t(\alpha)\|_2 \\ & \leq \rho_y B_V \|h_t(\alpha + \beta) - h_t(\alpha)\|_2 + \delta \rho_y B_V \|h_t(\alpha + \beta)\|_2 \\ & \stackrel{(ii)}{\leq} \delta \rho_y B_V (\lambda_t \rho_h B_U + \rho_h B_x B_W) \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} + \delta \rho_y B_V \lambda_t, \end{aligned} \quad (13)$$

where (i) is from Lipschitz continuity of  $\sigma_y$  and (ii) is from (11) and (12).

Then choosing the prior distribution and the perturbation distribution as  $\mathcal{N}(0, \sigma^2 I)$ , and from the concentration result for the spectral norm bounds, we have

$$\mathbb{P}_{A \sim \mathcal{N}(0, \sigma^2 I_{d \times d})} [\|A\|_2 > \xi] \leq 2p \exp\left(\frac{-\xi^2}{2d\sigma^2}\right).$$

This implies with probability at least  $1/2$ , we have  $\max\{\delta B_U, \delta B_W, \delta B_V\} \leq \sigma \sqrt{2d \ln(12d)}$ . Taking  $\sigma = \left(\gamma/4\rho_y \left((\lambda_t \rho_h B_U + \rho_h B_x B_W) \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} + \lambda_t\right) \sqrt{2d \ln(12d)}\right)$  and combining with (13), with probability at least  $1/2$ , we have

$$\begin{aligned} & \max_{x \in \mathcal{X}_m} \|y_t(\alpha + \beta) - y_t(\alpha)\|_2 \\ & \leq \left( (\lambda_t \rho_h B_U + \rho_h B_x B_W) \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} + \lambda_t \right) \cdot \sigma \sqrt{2d \ln(12d)} \leq \frac{\gamma}{4}. \end{aligned}$$



Finally, we calculate the KL divergence of  $\mathcal{P}$  and  $\alpha + \beta$  with respect to this choice of  $\sigma$ ,

$$\begin{aligned} \text{KL}(\alpha + \beta \| \mathcal{P}) &\leq \frac{\|\alpha\|_2^2}{2\sigma^2} \\ &= O\left(\frac{\rho_y^2}{\gamma^2} \cdot \left((\lambda_t \rho_h B_U + \rho_h B_x B_W) \frac{(\rho_h B_U)^t - 1}{\rho_h B_U - 1} + \lambda_t\right)^2 d \ln(d) (B_{U,F}^2 + B_{W,F}^2 + B_{V,F}^2)\right) \\ &= O\left(\frac{\rho_y^2 (\lambda_t \rho_h B_U + \rho_h B_x B_W)^2 (\beta^t - 1)^2 p \ln(p) (B_{U,F}^2 + B_{W,F}^2 + B_{V,F}^2)}{\gamma^2 (\beta - 1)^2}\right). \end{aligned}$$

We complete the proof by applying Lemma 11.  $\square$

## C Proofs in Sections 5 and ??

### C.1 Proof of Theorem 4

*Proof.* We use the same argument from the analysis of vanilla RNNs to investigate the Lipschitz continuity of MGU RNNs. Consider  $h_t$  and  $h'_t$  computed by different sets of weight matrices.

$$\begin{aligned} \|h_t - h'_t\|_2 &= \left\| (1 - r_t) \odot h_{t-1} + r_t \odot \tilde{h}_t - (1 - r'_t) \odot h'_{t-1} - r'_t \odot \tilde{h}'_t \right\|_2 \\ &\leq \left\| (r'_t - r_t) \odot h'_{t-1} \right\|_2 + \left\| (1 - r_t) \odot (h_{t-1} - h'_{t-1}) \right\|_2 + \left\| (r_t - r'_t) \odot \tilde{h}'_t \right\|_2 + \left\| r_t \odot (\tilde{h}_t - \tilde{h}'_t) \right\|_2 \\ &\leq \|r'_t - r_t\|_2 \|h'_{t-1}\|_\infty + \|1 - r_t\|_\infty \|h_{t-1} - h'_{t-1}\|_2 + \|r_t - r'_t\|_2 \|\tilde{h}'_t\|_\infty + \|r_t\|_\infty \|\tilde{h}_t - \tilde{h}'_t\|_2 \end{aligned}$$

Expand the expression of  $\tilde{h}_t$ . Note that  $r_t$  is nonnegative, and  $\|r_t\|_\infty \leq 1$ . Then we have  $\|h_t\|_\infty \leq 1$ . Additionally  $\tanh(\cdot)$  is 1-Lipschitz. Thus we get

$$\begin{aligned} \|\tilde{h}_t - \tilde{h}'_t\|_2 &\leq \|U_h(h_{t-1} \odot r_t) + W_h x_t - U'_h(h'_{t-1} \odot r'_t) - W'_h x_t\|_2 \\ &\leq \|U_h(h_{t-1} \odot r_t) - U'_h(h'_{t-1} \odot r'_t)\|_2 + B_x \|W_h - W'_h\|_2 \\ &\leq \|U_h - U'_h\|_2 \|h_{t-1} \odot r_t\|_2 + B_{U_h} \|r_t\|_\infty \|h_{t-1} - h'_{t-1}\|_2 + B_{U_h} \|h'_{t-1}\|_\infty \|r_t - r'_t\|_2 \\ &\quad + B_x \|W_h - W'_h\|_2 \\ &\leq \|h_t\|_2 \|U_h - U'_h\|_2 + B_{U_h} \|r_t\|_\infty \|h_{t-1} - h'_{t-1}\|_2 + B_{U_h} \|r_t - r'_t\|_2 + B_x \|W_h - W'_h\|_2. \end{aligned}$$

We have to expand  $r_t - r'_t$  as follows,

$$\begin{aligned} \|r_t - r'_t\|_2 &= \|W_r x_t + U_r h_{t-1} - W'_r x_t - U'_r h'_{t-1}\|_2 \\ &\leq B_x \|W_r - W'_r\|_2 + B_{U_r} \|h_{t-1} - h'_{t-1}\|_2 + \|h'_{t-1}\|_2 \|U_r - U'_r\|_2. \end{aligned}$$

We also need to bound  $\|h_t\|_2$ ,

$$\begin{aligned} \|h_t\|_2 &\leq \|1 - r_t\|_\infty \|h_{t-1}\|_2 + \|r_t\|_\infty \|\tilde{h}_t\|_2 \\ &\leq \|1 - r_t\|_\infty \|h_{t-1}\|_2 + \|r_t\|_\infty (B_{W_h} B_x + B_{U_h} \|r_t\|_\infty \|h_{t-1}\|_2) \\ &= \left( \|1 - r_t\|_\infty + B_{U_h} \|r_t\|_\infty^2 \right) \|h_{t-1}\|_2 + B_{W_h} B_x, \\ &\leq \max_{j \leq t} \left\{ \|1 - r_j\|_\infty + B_{U_h} \|r_j\|_\infty^2 \right\} \|h_{t-1}\|_2 + B_{W_h} B_x. \end{aligned}$$

Applying the above inequality recursively and remember  $\|h_t\|_\infty \leq 1$ , we get  $\|h_t\|_2 \leq \min \left\{ \sqrt{d}, \frac{\beta^t - 1}{\beta - 1} B_{W_h} B_x \right\}$  with  $\beta = \max_{j \leq t} \left\{ \|1 - r_j\|_\infty + B_{U_h} \|r_j\|_\infty^2 \right\}$ . Put all the above ingredients together, we have

$$\begin{aligned} \|h_t - h'_t\|_2 &\leq (\beta + 2B_{U_r} + B_{U_r} B_{U_h}) \|h_{t-1} - h'_{t-1}\|_2 \\ &\quad + \sqrt{d} \|U_h - U'_h\|_2 + B_x \|W_h - W'_h\|_2 \\ &\quad + (2 + B_{U_h}) \sqrt{d} \|U_r - U'_r\|_2 + (2B_x + B_{U_h} B_x) \|W_r - W'_r\|_2. \end{aligned}$$

Apply the above inequality recursively, denote by  $\theta = \beta + 2B_{U_r} + B_{U_r}B_{U_h}$ , we have

$$\begin{aligned} \|h_t - h'_t\|_2 &\leq \sqrt{d} \sum_{j=1}^t \theta^j \|U_h - U'_h\|_2 + B_x \sum_{j=1}^t \theta^j \|W_h - W'_h\|_2 \\ &\quad + \left(2\sqrt{d} + B_{U_h}\sqrt{d}\right) \sum_{j=1}^t \theta^j \|U_r - U'_r\|_2 + (2B_x + B_{U_h}B_x) \sum_{j=1}^t \theta^j \|W_r - W'_r\|_2. \end{aligned}$$

We then derive the Lipschitz continuity of  $\|y_t\|_2$ ,

$$\begin{aligned} \|y_t - y'_t\|_2 &\leq \rho_y B_V \|h_t - h'_t\|_2 + \rho_y \sqrt{d} \|V - V'\|_2 \\ &\leq \rho_y B_V \sqrt{d} \frac{\theta^t - 1}{\theta - 1} \|U_h - U'_h\|_2 + \rho_y B_V B_x \frac{\theta^t - 1}{\theta - 1} \|W_h - W'_h\|_2 + \rho_y \sqrt{d} \|V - V'\|_2 \\ &\quad + \rho_y B_V \left(2\sqrt{d} + B_{U_h}\sqrt{d}\right) \frac{\theta^t - 1}{\theta - 1} \|U_r - U'_r\|_2 + \rho_y B_V (2B_x + B_{U_h}B_x) \frac{\theta^t - 1}{\theta - 1} \|W_r - W'_r\|_2. \end{aligned}$$

Following the same argument for proving the generalization bound of vanilla RNNs, we can get the generalization bound for MGU RNNs as

$$\mathbb{P}(\tilde{z}_t \neq z_t) \leq \widehat{\mathcal{R}}_\gamma(f_t) + O\left(\frac{d\rho_y B_V \min\left\{\sqrt{d}, B_{W_h} B_x \frac{\beta^t - 1}{\beta - 1}\right\} \sqrt{\log\left(d\sqrt{m} \frac{\theta^t - 1}{\theta - 1}\right)}}{\sqrt{m}\gamma} + \sqrt{\frac{\log \frac{1}{\delta}}{m}}\right).$$

□

## C.2 Proof of Theorem 5

*Proof.* We first bound the norm of  $h_t$  as follows,

$$\begin{aligned} \|h_t\|_2 &\leq \|o_t\|_\infty \|\tanh(c_t)\|_2 \leq \|o_t\|_\infty \|c_t\|_2 \\ &\leq \|g_t\|_\infty \|c_{t-1}\|_2 + \|r_t\|_\infty \|\tilde{c}_t\|_2 \\ &\leq \|g_t\|_\infty \|c_{t-1}\|_2 + \|r_t\|_\infty (B_{W_c} B_x + B_{U_c} \|h_{t-1}\|_2) \\ &\leq \|g_t\|_\infty \|c_{t-1}\|_2 + \|r_t\|_\infty (B_{W_c} B_x + B_{U_c} \|o_t\|_\infty \|c_{t-1}\|_2) \\ &\leq (\|g_t\|_\infty + \|r_t\|_\infty \|o_t\|_\infty B_{U_c}) \|c_{t-1}\|_2 + B_{W_c} B_x. \end{aligned}$$

By applying the above inequality recursively, we have  $\|h_t\|_2 \leq \|c_t\|_2 \leq B_{W_c} B_x \frac{\beta^t - 1}{\beta - 1}$ , where  $\beta = \max_{j \leq t} \{\|g_j\|_\infty + \|r_j\|_\infty \|o_j\|_\infty B_{U_c}\}$ . We also have  $\|h_t\|_2 \leq \sqrt{d}$ . Thus, put together, we have  $\|h_t\|_2 \leq \min\left\{\sqrt{d}, B_{W_c} B_x \frac{\beta^t - 1}{\beta - 1}\right\}$ . Next, we investigate the Lipschitz continuity of  $h_t$ .

$$\begin{aligned} \|h_t - h'_t\|_2 &\leq \|o_t \odot \tanh(c_t) - o'_t \odot \tanh(c'_t)\|_2 \\ &\leq \|o_t - o'_t\|_2 \|\tanh(c_t)\|_\infty + \|o'_t\|_\infty \|c_t - c'_t\|_2 \end{aligned}$$

We have to expand  $o_t - o'_t$ ,

$$\|o_t - o'_t\|_2 \leq B_x \|W_o - W'_o\|_2 + B_{U_o} \|h_{t-1} - h'_{t-1}\|_2 + \|h_{t-1}\|_2 \|U_o - U'_o\|_2.$$

Note that  $\|B_{U_o}\|_2$  is usually small,  $o_t$  and  $o'_t$  are close, and we have  $\|h_{t-1} - h'_{t-1}\|_2 \leq \|o_t\|_\infty \|c_{t-1} - c'_{t-1}\|_2 \leq \|c_{t-1} - c'_{t-1}\|_2$ . Thus, we can derive

$$\|o_t - o'_t\|_2 \leq B_x \|W_o - W'_o\|_2 + B_{U_o} \|c_{t-1} - c'_{t-1}\|_2 + \sqrt{d} \|U_o - U'_o\|_2.$$

We also expand  $c_t - c'_t$  to get,

$$\|c_t - c'_t\|_2 \leq \|c_{t-1}\|_\infty \|g_t - g'_t\|_2 + \|r'_t\|_\infty \|c_{t-1} - c'_{t-1}\|_2 + \|\tilde{c}_t\|_\infty \|r_t - r'_t\|_2 + \|r'_t\|_\infty \|\tilde{c}_t - \tilde{c}'_t\|_2.$$

We also have,

$$\begin{aligned} \|\tilde{c}_t - \tilde{c}'_t\|_2 &\leq B_{U_c} \|h_{t-1} - h'_{t-1}\|_2 + \|h_{t-1}\|_2 \|U_c - U'_c\|_2 + B_x \|W_c - W'_c\|_2, \\ \|g_t - g'_t\|_2 &\leq B_x \|W_g - W'_g\|_2 + B_{W_g} \|h_{t-1} - h'_{t-1}\|_2 + \sqrt{d} \|U_g - U'_g\|_2, \quad \text{and} \\ \|r_t - r'_t\|_2 &\leq B_x \|W_r - W'_r\|_2 + B_{W_r} \|h_{t-1} - h'_{t-1}\|_2 + \sqrt{d} \|U_r - U'_r\|_2. \end{aligned}$$

Putting together, we get

$$\begin{aligned}
 & \|c_t - c'_t\|_2 \\
 & \leq B_x (\|W_c - W'_c\|_2 + \|W_g - W'_g\|_2 + \|W_r - W'_r\|_2) \\
 & \quad + \sqrt{d} (\|U_c - U'_c\|_2 + \|U_g - U'_g\|_2 + \|U_r - U'_r\|_2) \\
 & \quad + \|g_t\|_\infty \|c_{t-1} - c'_{t-1}\|_2 + (\|r_t\|_\infty B_{U_c} + B_{U_g} + B_{U_r}) \|h_{t-1} - h'_{t-1}\|_2 \\
 & \leq B_x (\|W_c - W'_c\|_2 + \|W_g - W'_g\|_2 + \|W_r - W'_r\|_2 + (B_{U_c} + B_{U_g} + B_{U_r}) \|W_o - W'_o\|_2) \\
 & \quad + \sqrt{d} (\|U_c - U'_c\|_2 + \|U_g - U'_g\|_2 + \|U_r - U'_r\|_2 + (B_{U_c} + B_{U_g} + B_{U_r}) \|U_o - U'_o\|_2) \\
 & \quad + (\|o_t\|_\infty \|r_t\|_\infty B_{U_c} + B_{U_g} + B_{U_r} + B_{U_o}) \|c_{t-1} - c'_{t-1}\|_2.
 \end{aligned}$$

By induction, we have

$$\begin{aligned}
 & \|c_t - c'_t\|_2 \\
 & \leq B_x \frac{\theta^t - 1}{\theta - 1} (\|W_c - W'_c\|_2 + \|W_g - W'_g\|_2 + \|W_r - W'_r\|_2 + (B_{U_c} + B_{U_g} + B_{U_r}) \|W_o - W'_o\|_2) \\
 & \quad + \sqrt{d} \frac{\theta^t - 1}{\theta - 1} (\|U_c - U'_c\|_2 + \|U_g - U'_g\|_2 + \|U_r - U'_r\|_2 + (B_{U_c} + B_{U_g} + B_{U_r}) \|U_o - U'_o\|_2),
 \end{aligned}$$

where  $\theta = \beta + B_{U_g} + B_{U_r} + B_{U_o}$ . Now we immediately have

$$\begin{aligned}
 & \|h_t - h'_t\|_2 \\
 & \leq B_x \frac{\theta^t - 1}{\theta - 1} (\|W_c - W'_c\|_2 + \|W_g - W'_g\|_2 + \|W_r - W'_r\|_2 + (B_{U_c} + B_{U_g} + B_{U_r}) \|W_o - W'_o\|_2) \\
 & \quad + \sqrt{d} \frac{\theta^t - 1}{\theta - 1} (\|U_c - U'_c\|_2 + \|U_g - U'_g\|_2 + \|U_r - U'_r\|_2 + (B_{U_c} + B_{U_g} + B_{U_r}) \|U_o - U'_o\|_2).
 \end{aligned}$$

Then the Lipschitz continuity of  $y_t$  can be written as

$$\|y_t - y'_t\|_2 \leq \rho_y B_V \|h_t - h'_t\|_2 + \rho_y \sqrt{d} \|V - V'\|_2.$$

Following the same argument for proving the generalization bound of vanilla RNNs, we can get the generalization bound for LSTM RNNs as

$$\mathbb{P}(\tilde{z}_t \neq z_t) \leq \hat{\mathcal{R}}_\gamma(f_t) + O\left(\frac{d\rho_y B_V \min\{\sqrt{d}, B_{W_c} B_x \frac{\beta^t - 1}{\beta - 1}\} \sqrt{\log\left(d\sqrt{m} \frac{\theta^t - 1}{\theta - 1}\right)}}{\sqrt{m}\gamma} + \sqrt{\frac{\log \frac{1}{\delta}}{m}}\right).$$

□

### C.3 Proof of Theorem 6

*Proof.* We first characterize the Lipschitz continuity of  $\|y_t\|_2$  with respect to model parameters  $\mathcal{U}$ ,  $\mathcal{W}$  and  $\mathcal{V}$ . We have

$$\|y_t - y'_t\|_2 \leq \rho_y \|h_t\|_2 \|W_{\mathcal{V}} - W_{\mathcal{V}'}\|_2 + \rho_y \|W_{\mathcal{V}}\|_2 \|h_t - h'_t\|_2.$$

Since  $\|h_t\|_\infty \leq 1$ , we have  $\|h_t\|_2 \leq \sqrt{d}$ . Then we expand  $h_t - h'_t$ ,

$$\begin{aligned}
 \|h_t - h'_t\|_2 & \leq \rho_h \|\mathcal{U}' * h_{t-1} + \mathcal{W} * x_t - \mathcal{U}' * h'_{t-1} - \mathcal{W}' * x_t\|_2 \\
 & = \rho_h \|PW_{\mathcal{U}}h_{t-1} + PW_{\mathcal{W}}x_t - PW'_{\mathcal{U}}h'_{t-1} - PW'_{\mathcal{W}}x_t\|_2 \\
 & \leq \rho_h \|P\|_2 \|W_{\mathcal{U}}h_{t-1} + W_{\mathcal{W}}x_t - W'_{\mathcal{U}}h'_{t-1} - W'_{\mathcal{W}}x_t\|_2 \\
 & \leq \rho_h \|P\|_2 \left[ B_x \|W_{\mathcal{W}} - W'_{\mathcal{W}}\|_2 + \sqrt{d} \|W_{\mathcal{U}} - W'_{\mathcal{U}}\|_2 + \|W_{\mathcal{U}}\|_2 \|h_{t-1} - h'_{t-1}\|_2 \right].
 \end{aligned}$$

Observe that we have by the definition of circulant matrix,

$$\|W_{\mathcal{U}} - W'_{\mathcal{U}}\|_2^2 \leq \|W_{\mathcal{U}} - W'_{\mathcal{U}}\|_{\mathbb{F}}^2 = (d - k) \|\mathcal{U} - \mathcal{U}'\|_{\mathbb{F}}^2 \leq d \|\mathcal{U} - \mathcal{U}'\|_{\mathbb{F}}^2.$$

The same holds for  $W_{\mathcal{W}} - W'_{\mathcal{W}}$  and  $W_{\mathcal{V}} - W'_{\mathcal{V}}$ . We also have  $\|P\|_2 = 1$ . The remaining task is to bound the spectral norm of  $W_{\mathcal{U}}$  and  $W_{\mathcal{V}}$ . Consider the matrix product  $W_{\mathcal{U}}^\top W_{\mathcal{U}}$ . We claim that the diagonal elements of  $W_{\mathcal{U}}^\top W_{\mathcal{U}}$  is bounded by  $\sum_{i=1}^k \|\mathcal{U}_i\|_2^2$ , and the off-diagonal elements are zero. To see this, denote by  $C_{\mathcal{U}_i}$  the

circulant like matrix generated by  $\mathcal{U}_i$ . Then we have  $W_{\mathcal{U}} = [C_{\mathcal{U}_1}^\top, \dots, C_{\mathcal{U}_k}^\top]^\top$ . The diagonal elements of  $W_{\mathcal{U}}^\top W_{\mathcal{U}}$  are

$$(W_{\mathcal{U}}^\top W_{\mathcal{U}})_{ii} = \sum_{j=1}^k (C_{\mathcal{U}_j}^\top C_{\mathcal{U}_j})_{ii} \leq \sum_{i=1}^k \|\mathcal{U}_i\|_2^2.$$

By the orthogonality of  $\mathcal{U}$ , the off-diagonal elements are

$$(W_{\mathcal{U}}^\top W_{\mathcal{U}})_{pq} = \sum_{j=1}^k (C_{\mathcal{U}_j}^\top C_{\mathcal{U}_j})_{pq} = \sum_{j=1}^k (C_{\mathcal{U}_j})_{:p}^\top (C_{\mathcal{U}_j})_{:q} = 0.$$

Thus, the spectral norm  $\|W_{\mathcal{U}}\|_2 \leq \sqrt{\sum_{i=1}^k \|\mathcal{U}_i\|_2^2} \leq 1$ , and  $\|W_{\mathcal{V}}\|_2, \|W_{\mathcal{W}}\|_2 \leq 1$  also hold. Then we can derive

$$\|h_t - h'_t\|_2 \leq \rho_h B_x \sqrt{d} \|\mathcal{W} - \mathcal{W}'\|_{\text{F}} + \rho_h d \|\mathcal{U} - \mathcal{U}'\|_{\text{F}} + \rho_h \|h_{t-1} - h'_{t-1}\|_2.$$

Apply the above inequality recursively, we get

$$\begin{aligned} \|h_t - h'_t\|_2 &\leq \rho_h B_x \sqrt{d} \frac{\rho_h^t - 1}{\rho_h - 1} \|\mathcal{W} - \mathcal{W}'\|_{\text{F}} + \rho_h d \frac{\rho_h^t - 1}{\rho_h - 1} \|\mathcal{U} - \mathcal{U}'\|_{\text{F}} \\ &\leq B_x \sqrt{dt} \|\mathcal{W} - \mathcal{W}'\|_{\text{F}} + dt \|\mathcal{U} - \mathcal{U}'\|_{\text{F}}. \end{aligned}$$

Thus, we have the following Lipschitz continuity of  $\|y_t\|_2$ ,

$$\|y_t - y'_t\|_2 \leq d \|\mathcal{V} - \mathcal{V}'\|_{\text{F}} + B_x \sqrt{dt} \|\mathcal{W} - \mathcal{W}'\|_{\text{F}} + dt \|\mathcal{U} - \mathcal{U}'\|_{\text{F}}.$$

We also bound the norm of  $h_t$  by induction. Specifically, we have

$$\|h_t\|_2 \leq \rho_h \|PW_{\mathcal{U}}h_{t-1} + PW_{\mathcal{W}}x_t\|_2 \leq \rho_h \|W_{\mathcal{U}}h_{t-1}\|_2 + \rho_h \|W_{\mathcal{W}}x_t\|_2 \leq \|h_{t-1}\|_2 + B_x.$$

Applying the above expression recursively, we have  $\|h_t\|_2 \leq \min\{\sqrt{d}, B_x t\} \leq B_x t$ . Then following the same argument for proving the generalization bound of vanilla RNNs, we can get the generalization bound for Conv RNNs as

$$\mathbb{P}(\tilde{z}_t \neq z_t) \leq \widehat{\mathcal{R}}_\gamma(f_t) + O\left(\frac{B_x kt \sqrt{\log(dt\sqrt{m})}}{\sqrt{m}\gamma} + \sqrt{\frac{\log \frac{1}{\delta}}{m}}\right).$$

□