A. Proofs

A.1. Proof of Proposition 1

Proposition 1. *(Householder QR factorization) Let $B \in \mathbb{R}^{n\times n}$. There exists an upper triangular matrix R with positive diagonal elements, and vectors $\{u_i\}_{i=1}^n$ with $u_i \in \mathbb{R}^i$, such that $B = \mathcal{H}_n^n(u_n)...\mathcal{H}_1^n(u_1)R$. (Note that we allow $u_i = 0$, in which case, $H_i^n(u_i) = I_n$ as in (1))*

Proof of Proposition 1. For $n = 1$, note that $\mathcal{H}_1^1(u_1) = \pm 1$. By setting $u_1 = 0$ if $B_{1,1} > 0$ and $u_1 \neq 0$ otherwise, we have the factorization desired.

Assume that the result holds for $n = k$, then for $n = k+1$ set $u_{k+1} = B_1 - \|B_1\| e_1$. Here B_1 is the first column of B and $e_1 = (1, 0, ..., 0)^\top$. Thus we have

$$\mathcal{H}_{k+1}^{k+1}(u_{k+1})B = \begin{pmatrix} \|B_1\| & \hat{B}_{1,2:k+1} \\ 0 & \hat{B} \end{pmatrix},$$

where $\hat{B} \in \mathbb{R}^{k\times k}$. Note that $\mathcal{H}_{k+1}^{k+1}(u_{k+1}) = I_{k+1}$ when $u_{k+1} = 0$ and the above still holds. By assumption we have $\hat{B} = \mathcal{H}_k^k(u_k)...\mathcal{H}_1^k(u_1)\hat{R}$. Notice that $\mathcal{H}_i^{k+1}(u_i) = \begin{pmatrix} 1 & \\ & \mathcal{H}_i^k(u_i) \end{pmatrix}$, so we have that

$$\mathcal{H}_1^{k+1}(u_1)...\mathcal{H}_k^{k+1}(u_k)\mathcal{H}_{k+1}^{k+1}(u_{k+1})B = \begin{pmatrix} \|B_1\| & \tilde{B}_{1,2:k+1} \\ 0 & \hat{R} \end{pmatrix} = R$$

is an upper triangular matrix with positive diagonal elements. Thus the result holds for any n by the theory of mathematical induction. □

A.2. Proof of Theorem 1

Proof. Observe that the image of $\mathcal{M}_1$ is a subset of $\mathbf{O}(n)$, and we now show that the converse is also true. Given $A \in \mathbf{O}(n)$, by Proposition 1, there exists an upper triangular matrix R with positive diagonal elements, and an orthogonal matrix Q expressed as $Q = \mathcal{H}_n^n(u_n)...\mathcal{H}_1^n(u_1)$ for some set of Householder vectors $\{u_i\}_{i=1}^n$, such that $A = QR$. Since A is orthogonal, we have $A^\top A = AA^\top = I_n$, thus:

$$A^\top A = R^\top Q^\top QR = R^\top R = I_n;\ Q^\top AA^\top Q = Q^\top QRR^\top Q^\top Q = RR^\top = I_n$$

Thus R is orthogonal and upper triangular matrix with positive diagonal elements. So $R = I_n$ and $A = Q = \mathcal{H}_n^n(u_n)...\mathcal{H}_1^n(u_1)$. □

A.3. Proof of Theorem 2

Proof. It is easy to see that the image of $\mathcal{M}_{1,1}$ is a subset of $\mathbb{R}^{n\times n}$. For any $W \in \mathbb{R}^{n\times n}$, we have its SVD, $W = U\Sigma V^\top$, where $\Sigma = diag(\sigma)$. By Theorem 1, for any orthogonal matrix $U, V \in \mathbb{R}^{n\times n}$, there exists $\{u_i\}_{i=1}^n\{v_i\}_{i=1}^n$ such that $U = \mathcal{M}_1(u_1, ..., u_n)$ and $V = \mathcal{M}_1(v_1, ..., v_n)$, then we have:

$$\begin{aligned} W &= \mathcal{H}_n^n(u_n)...\mathcal{H}_1^n(u_1)\Sigma\mathcal{H}_1^n(v_1)...\mathcal{H}_n^n(v_n) \\ &= \mathcal{M}_{1,1}(u_1, ..., u_n, v_1, ..., v_n, \sigma) \end{aligned}$$

□

A.4. Proof of Theorem 3

Proof. Let $A \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. By Theorem 1, there exist $\{a_i\}_{i=1}^n$, such that $A = \mathcal{M}_1(a_1, ..., a_n)$. Since $A^\top$ is also orthogonal, for the same reason, there exist $\{b_i\}_{i=1}^n$, such that $A^\top = \mathcal{M}_1(b_1, ..., b_n)$. Thus we have:

$$A = \mathcal{H}_n(a_n)...\mathcal{H}_1(a_1) = \mathcal{H}_1(b_1)...\mathcal{H}_n(b_n)$$

Observe that one of $k_2 \geq k_1 - 1$ and $k_1 \geq k_2 - 1$ must be true. If $k_2 \geq k_1 - 1$, set

$$\begin{aligned} u_k &= a_k, k = n, n-1, ..., k_1, \\ v_{k_2+k_1-k-1} &= a_k, k = k_1 - 1, ..., 1, \\ v_t &= \mathbf{0}, t = k_2 + k_1 - 2, ..., n, \end{aligned} \tag{18}$$

and then we have:

$$\begin{aligned}\mathcal{M}_{k_1,k_2}(u_{k_1},...,u_n,v_{k_2},...,v_n,\mathbf{1}) &= \mathcal{H}_n(u_n)...\mathcal{H}_{k_1}(u_{k_1})I_n\mathcal{H}_{k_2}(v_{k_2})...\mathcal{H}_n(v_n)\\ &= \mathcal{H}_n(a_n)...\mathcal{H}_{k_1}(a_{k_1})I_n\mathcal{H}_{k_1-1}(a_{k_1-1})...\mathcal{H}_1(a_1)\\ &= A \end{aligned} \tag{19}$$

Else, assign:

$$\begin{aligned} v_k &= b_k, k = n, n-1, ..., k_2,\\ u_{k_2+k_1-k-1} &= b_k, k = k_2-1, ..., 1,\\ u_t &= \mathbf{0}, t = k_2+k_1-2, ..., n, \end{aligned} \tag{20}$$

and then we have:

$$\begin{aligned}\mathcal{M}_{k_1,k_2}(u_{k_1},...,u_n,v_{k_2},...,v_n,\mathbf{1}) &= \mathcal{H}_1(b_1)...\mathcal{H}_{k_2-1}(b_{k_2-1})I_n\mathcal{H}_{k_2}(b_{k_2})...\mathcal{H}_n(b_n)\\ &= A \end{aligned} \tag{21}$$

□

A.5. Proof of Theorem 4

Proof. It is easy to see that the image of $\mathcal{M}_{*,*}^{m,n}$ is a subset of $\mathbb{R}^{m\times n}$. For any $W \in \mathbb{R}^{m\times n}$, we have its SVD, $W = U\Sigma V^\top$, where Σ is an $m \times n$ diagonal matrix. By Theorem 1, for any orthogonal matrix $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$, there exists $\{u_i\}_{i=1}^m\{v_i\}_{i=1}^n$ such that $U = \mathcal{H}_m^m(u_m)...\mathcal{H}_1^m(u_1)$ and $V = \mathcal{H}_n^n(v_n)...\mathcal{H}_1^n(v_1)$. By Lemma 1, if $m < n$ we have:

$$\begin{aligned} W &= \mathcal{H}_n^m(u_n)...\mathcal{H}_1^m(u_1)\Sigma\mathcal{H}_1^n(v_1)...\mathcal{H}_n^n(v_n)\\ &= \mathcal{H}_n^m(u_n)...\mathcal{H}_1^m(u_1)\Sigma\mathcal{H}_{n-m+1}^n(v_{n-m+1})...\mathcal{H}_n^n(v_n). \end{aligned}$$

Similarly, for $n < m$, we have:

$$\begin{aligned} W &= \mathcal{H}_n^m(u_n)...\mathcal{H}_1^m(u_1)\Sigma\mathcal{H}_1^n(v_1)...\mathcal{H}_n^n(v_n)\\ &= \mathcal{H}_n^m(u_n)...\mathcal{H}_{m-n+1}^m(u_{m-n+1})\Sigma\mathcal{H}_1^n(v_1)...\mathcal{H}_n^n(v_n). \end{aligned}$$

□

A.6. Proof of Theorem 5

Notations: Recall from Definition 1 that L_0 is the expected error with margin $\gamma = 0$, and we write $\hat{L}_\gamma$ as the empirical error when margin equals γ with m samples, i.e.,

$$\hat{L}_\gamma(f_w) = \frac{1}{m}\sum_{i=1}^m \left[f_w(x_i)[y_i] \le \gamma + \max_{j\ne y_i} f_w(x_i)[j] \right].$$

We are looking at a recurrent neural network with T time steps:

$$\begin{aligned} &h^{(t)} = \phi(Wh^{(t-1)} + Mx^{(t)}), h^{(0)} = 0, t = 1, 2, \cdots T\\ &\hat{y}^{(t)} = Yh^{(t)}, \end{aligned}$$

where ϕ is the activation function. The dimensions are as follows: $x^{(t)} \in \mathbb{R}^{n_i}, \hat{y}^{(t)} \in \mathbb{R}^{n_y}$, and $h^{(t)} \in \mathbb{R}^n$. Therefore $W \in \mathbb{R}^{n\times n}, M \in \mathbb{R}^{n_i\times n}, Y \in \mathbb{R}^{n\times n_y}$. To incorporate the different parameters W, M, Y into the neural network, we write $w = \text{vec}(\{W, Y, M\})$ and use subscript w to denote dependence on the parameter w. For instance, $h_w^{(t)}$ denotes the activation that takes $w = \text{vec}(\{W, Y, M\})$ as parameters, and similar notation also holds for the output $\hat{y}_w^{(t)}$. We use $\|\cdot\|$ to denote l_2 norm for vectors and spectral norm for matrices when there is no ambiguity.

To get a generalization bound for RNN, we need to use the following lemma from (Neyshabur et al., 2017).

Lemma 2. *(Neyshabur et al., 2017) Let $f_w(x) : \mathcal{X} \to \mathbb{R}^k$ be any predictor (not necessarily a neural network) with parameters w, and P be any distribution on the parameters that is independent of the training data. Then, for any $\gamma, \delta > 0$, with probability $\geq 1 - \delta$ over the training set of size m, for any w, and any random perturbation u s.t. $\mathbb{P}_u[\max_{x\in\mathcal{X}} \|f_{w+u}(x) - f_w(x)\|_\infty < \frac{\gamma}{4}] \geq \frac{1}{2}$, we have:*

$$L_0(f_w) \leq \hat{L}_\gamma(f_w) + 4\sqrt{\frac{KL(w+u||P) + \ln\frac{6m}{\delta}}{m-1}}$$

Here $KL(P||Q)$ is the Kullback-Leibler divergence of two continuous random variables P and Q:

$$KL(P||Q) := \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx,$$

where p and q denote the density of P and Q. In order for the random variable u to satisfy the probability property in Lemma 2, we study the change in output with respect to perturbation u.

Lemma 3. *Write $w = vec(\{W, Y, M\})$, and perturbation $u = vec(\{\delta W, \delta Y, \delta M\})$ such that $\|\delta W\| \leq \frac{1}{T}\|W\|$, $\|\delta Y\| \leq \frac{1}{T}\|Y\|$, $\|\delta M\| \leq \frac{1}{T}\|M\|$. For a recurrent neural network (17) with T time steps that satisfies Assumption 1, the perturbation in the activation is bounded by*

$$\|h_{w+u}^{(T)} - h_w^{(T)}\| \leq BTe(T\|M\|\|\delta W\| + \|\delta M\|)\max\{\|W\|^{T-1}, 1\}, \tag{22}$$

while the perturbation in the output satisfies:

$$\|\hat{y}_{w+u}^{(T)} - \hat{y}_w^{(T)}\| \leq TB\max\{\|W\|^{T-1}, 1\} \cdot (\|Y\|\|\delta W\|\|M\|Te + \|Y\|\|\delta M\|e + \|\delta Y\|\|M\|).$$

Here e is the natural logarithm base.

Proof of Lemma 3. First we bound the norm of $h_w^{(t)}$,

$$\begin{aligned}
\|h_w^{(t)}\| &= \|\phi(Wh_w^{(t-1)} + Mx^{(t)})\| \\
&\leq \|Wh_w^{(t-1)} + Mx^{(t)}\| && \text{(by Assumption 1.2)} \\
&\leq \|W\|\|h_w^{(t-1)}\| + \|M\|\|x^{(t)}\| && \text{(by triangle inequality)} \qquad (23) \\
&\leq \|W\|\left(\|W\|\|h_w^{(t-2)}\| + \|M\|\|x^{(t-1)}\|\right) + \|M\|\|x^{(t)}\| \\
& && \text{(applying (23) to } \|h_w^{(t-1)}\|\text{)} \\
&\leq \cdots \\
&\leq \|W\|^t\|h_w^{(0)}\| + \|M\|\sum_{j=0}^{t-1}\|W\|^{t-1-j}\|x^{(j+1)}\| \\
&= \|M\|\sum_{j=0}^{t-1}\|W\|^{t-1-j}\|x^{(j+1)}\| && \text{(since } h_w^{(0)} = 0\text{)} \\
&\leq B\|M\|\sum_{j=0}^{t-1}\|W\|^{t-1-j} && \text{(by Assumption 1.1)} \\
\Longrightarrow \|h_w^{(t)}\| &\leq B\|M\|t\max\{\|W\|^{t-1}, 1\} && (24) \\
& && \text{(since } \sum_{i=0}^{t-1}\|W\|^i \leq t\max\{\|W\|^{t-1-i}, 1\}\text{)}
\end{aligned}$$

Denoting $\Delta_t = \|h_{w+u}^{(t)} - h_w^{(t)}\|$ for short, in order to prove (22), we now prove the following tighter result by induction,

$$\Delta_t \leq Bt(1 + \frac{1}{T})^{t-1}(\|\delta W\|\|M\|T + \|\delta M\|)\max\{\|W\|^{t-1}, 1\}, \forall t \leq T \tag{25}$$

Clearly $\Delta_0 = 0$ satisfies the inequality. Suppose Δ_{t-1} satisfies the assumption, then,

$$
\begin{aligned}
\Delta_t &= \|\phi\left((W+\delta W)h_{w+u}^{(t-1)} + (M+\delta M)x^{(t)}\right) - \phi\left(Wh_w^{(t-1)} + Mx^{(t)}\right)\| \\
&\le \|\left((W+\delta W)h_{w+u}^{(t-1)} + (M+\delta M)x^{(t)}\right) - \left(Wh_w^{(t-1)} + Mx^{(t)}\right)\| \\
&\qquad \text{(by Assumption 1.2)} \\
&= \|(W+\delta W)(h_{w+u}^{(t-1)} - h_w^{(t-1)}) + \delta W h_w^{(t-1)} + \delta M x^{(t)}\| \\
&\le (\|W\| + \|\delta W\|\Delta_{t-1} + \|\delta W\|\|h_w^{(t-1)}\| + \|\delta M\|\|x^{(t)}\| \qquad \text{(by triangle inequality)} \\
&\le (1+\frac{1}{T})\|W\|\Delta_{t-1} + \|\delta W\|\|h_w^{(t-1)}\| + \|\delta M\|B \\
&\qquad \text{(by Assumption 1.1 and requirement of } \|\delta W\|\text{)}
\end{aligned}
$$

Then by induction and the bound of the activations, we have:

$$
\begin{aligned}
\Delta_t \le &(1+\frac{1}{T})\|W\|\left(B(t-1)(1+\frac{1}{T})^{t-2}(\|\delta W\|\|M\|T + \|\delta M\|)\max\{\|W\|^{t-2},1\}\right) \qquad \text{(by induction)} \\
&+ \|\delta W\|\left(B(t-1)\|M\|\max\{\|W\|^{t-2},1\}\right) + B\|\delta M\| \qquad \text{(by activation bound (24))} \\
=&B(t-1)T(1+\frac{1}{T})^{t-1}\|\delta W\|\|M\|\|W\|\max\{\|W\|^{t-2},1\} + B(t-1)\|\delta W\|\|M\|\max\{\|W\|^{t-2},1\} \\
&+ (1+\frac{1}{T})^{t-1}\|W\|B(t-1)\max\{\|W\|^{t-2},1\}\|\delta M\| + B\|\delta M\| \\
=&B(t-1)\|\delta W\|\|M\|\max\{\|W\|^{t-2},1\}\left((1+\frac{1}{T})^{t-1}T\|W\| + 1\right) \\
&+ B\|\delta M\|\left((1+\frac{1}{T})^{t-1}\|W\|(t-1)\max\{\|W\|^{t-2},1\} + 1\right) \\
\le&B(t-1)\|\delta W\|\|M\|\max\{\|W\|^{t-2},1\}\left((1+\frac{1}{T})^{t-1}T + 1\right)\max\{\|W\|,1\} \\
&+ B\|\delta M\|\left((1+\frac{1}{T})^{t-1}(t-1)\max\{\|W\|^{t-2},1\} + 1\right)\max\{\|W\|,1\} \qquad \text{(both } 1, \|W\| \le \max\{\|W\|,1\}\text{)} \\
\le&B\|\delta W\|\|M\|tT(1+\frac{1}{T})^{t-1}\max\{\|W\|^{t-1},1\} \\
&+ B\|\delta M\|t(1+\frac{1}{T})^{t-1}\max\{\|W\|^{t-1},1\} \qquad \text{(since } (t-1)a+1 \le ta \text{ for } a \ge 1\text{)} \\
=&Bt(1+\frac{1}{T})^{t-1}(T\|\delta W\|\|M\| + \|\delta M\|)\max\{\|W\|^{t-1},1\}
\end{aligned}
$$

Since $(1+\frac{1}{T})^{T-1} \le e$, therefore $\Delta_T \le BTe(T\|M\|\|\delta W\| + \|\delta M\|)\max\{\|W\|^{T-1},1\}$. Meanwhile for the perturbation of output $\hat{y}$,

$$
\begin{aligned}
&\|\hat{y}_{w+u}^{(T)} - \hat{y}_w^{(T)}\| \\
=&\|(Y+\delta Y)h_{w+u}^{(T)} - Yh_w^{(T)}\| \\
=&\|(Y+\delta Y)(h_{w+u}^{(T)} - h_w^{(T)}) + (Y+\delta Y)h_w^{(T)} - Yh_w^{(T)}\| \\
\le&\|(Y+\delta Y)\|\Delta_T + \|\delta Y h_w^{(T)}\| \qquad \text{(by triangle inequality)} \\
\le&\|Y\|(1+\frac{1}{T})BT(1+\frac{1}{T})^{T-1}(T\|\delta W\|\|M\| + \|\delta M\|)\max\{\|W\|^{T-1},1\} \\
&\qquad \text{(by perturbation bound (25))} \\
&+ \|\delta Y\|TB\|M\|\max\{\|W\|^{T-1},1\} \qquad \text{(by activation bound (24))} \\
\le&TB\max\{\|W\|^{T-1},1\}(\|Y\|\|\delta W\|\|M\|Te + \|Y\|\|\delta M\|e + \|\delta Y\|\|M\|) \\
&\qquad \text{(since } (1+\frac{1}{T})^T \le e\text{)}
\end{aligned}
$$

□

Finally we are able to prove Theorem 5:

Proof of Theorem 5. In order to finish the proof, we first calculate the maximum allowed perturbation u that satisfies the requirement in Lemma 2, and we define the prior P and calculate the KL divergence of P and $w + u$.

Let $\beta = \max\{\|W\|_2^{T-1}, 1\} \max\{\|Y\|_2, 1\} \max\{\|M\|_2, 1\}$. We choose the distribution of the prior $P = \mathcal{N}(0, \sigma^2 I)$ and consider the random perturbation $u = \text{vec}(\{\delta W, \delta Y, \delta M\})$ with the same zero mean Gaussian distribution, where σ will be assigned later according to β. More precisely, since the prior cannot depend on the β which is associated with the learned parameters W, M and Y, we will set σ based on some discrete choices of $\tilde{\beta}$ that approximates β. For each value of $\tilde{\beta}$ of our choice, we will compute the PAC-Bayes bound, establishing the generalization guarantee for all w for which $\frac{1}{e}\beta \le \tilde{\beta} \le e\beta$, and ensuring that each relevant value of β is covered by some $\tilde{\beta}$ on the grid. We will then take a union bound over all $\tilde{\beta}$ of our choice.

For a random matrix $X \in \mathbb{R}^{n_1 \times n_2}$ with individual entries following normal distribution, (Tropp et al., 2005) provides the following bound of its spectral norm:

$$\mathbb{P}_{X\sim\mathcal{N}(0,\sigma^2 I)}[\|X\|_2 > t] \le 2ne^{-t^2/2n\sigma^2}, \forall n \ge n_1, n_2 \tag{26}$$

Therefore for $\delta W, \delta M, \delta Y$, the probability of their spectral norm being greater than t is bounded by $2he^{-t^2/2h\sigma^2}$, where $h = \max\{n, n_i, n_y\}$. Therefore with probability $\ge \frac{1}{2}$, $\|\delta W\|_2, \|\delta Y\|_2, \|\delta M\|_2 \le \sigma\sqrt{2h\ln(12h)}$.

Plugging into Lemma 3 we have with probability at least $\frac{1}{2}$,

$$\begin{aligned}
& \max_{\|x^{(t)}\| \le B, \forall t \le T} \|\hat{y}_{w+u} - \hat{y}_w\| \\
\le \quad & TB\max\{\|W\|^{T-1}, 1\}(\|Y\|\|\delta W\|\|M\|Te + \|Y\|\|\delta M\|e + \|\delta Y\|\|M\|) \\
\le \quad & TB\max\{\|W\|^{T-1}, 1\}\max\{\|Y\|, 1\}\max\{\|M\|, 1\}(\|\delta W\|Te + \|\delta M\|e + \|\delta Y\|) \\
\le \quad & TB\sqrt{2h\ln(12h)}\tilde{\beta}\phi(Te + e + 1) \\
\le \quad & \frac{\gamma}{4},
\end{aligned}$$

where we choose $\sigma = \frac{\gamma}{12\sqrt{2h\ln(12h)}TB(Te+e+1)\tilde{\beta}}$. Therefore now the perturbation u satisfies assumptions in Lemma 2.

We next compute the KL-divergence of distributions for P and u for the sake of Lemma 2.

$$\begin{aligned}
& KL(w + u \parallel P) \le \frac{\|w\|^2}{2\sigma^2} \\
& \le \mathcal{O}\left(\frac{B^2T^4h\ln(h)\max\{\|W\|^{2T-2}, 1\}\max\{\|M\|_2^2, 1\}\max\{\|Y\|_2^2, 1\}}{\gamma^2}(\|W\|_F^2 + \|M\|_F^2 + \|Y\|_F^2)\right)
\end{aligned}$$

□

Hence, with probability $\ge 1 - \delta$ and for all w such that, $\frac{1}{e}\beta \le \tilde{\beta} \le e\beta$, we have:

$$L_0(\hat{y}_w) \le \hat{L}_\gamma(\hat{y}_w) + \mathcal{O}(\sqrt{\frac{B(w) + \ln\frac{m}{\delta}}{m}}), \tag{27}$$

where $B(w) = \frac{B^2T^4h\ln(h)\max\{\|W\|^{2T-2},1\}\max\{\|M\|_2^2,1\}\max\{\|Y\|_2^2,1\}}{\gamma^2}(\|W\|_F^2 + \|M\|_F^2 + \|Y\|_F^2)$.

Since $\tilde{\beta}$ should be independent of the learned models. We finally take a union bound over different choices of the parameter. We will choose discrete set of $\tilde{\beta}$ such that they cover the real W, M, Y that satisfies $\frac{1}{e}\beta \le \tilde{\beta} \le e\beta$. Firstly we notice for some range of β inequality (27) holds trivially, when either term of its RHS is greater or equal to 1, since the expected margin loss is less or equal to 1.

$\hat{y}^{(T)} = Yh_w^{(T)}$, therefore if $\beta \leq \frac{\gamma}{2BT}$,

$$\begin{aligned}
\|\hat{y}\|_\infty &\leq \|\hat{y}\|_2 && \text{(by definition of } \ell_\infty \text{ norm and } \ell_2 \text{ norm)} \\
&\leq \|Y\|\|h_w^{(T)}\| && \text{(by definition of spectral norm)} \\
&\leq \|Y\|B\|M\|T\max\{\|W\|^{t-1}, 1\} \\
& && \text{(by activation bound (24))} \\
&\leq BT\beta < \frac{\gamma}{2}
\end{aligned}$$

Therefore $\hat{L}_\gamma = 1$ from definition of margin loss and the bound is satisfied trivially. Meanwhile, when $\beta \geq \frac{\gamma\sqrt{m}}{2BT}$, then the second term of (27) ≥ 1 and it also holds trivially. Therefore, we only need to consider $\tilde{\beta}$ such that $\tilde{\beta} \in [\frac{\gamma}{2BT}, \frac{\gamma\sqrt{m}}{2BT}]$. Therefore we could respectively set $\tilde{\beta}$ to be $\frac{\gamma}{2BT} + se\frac{\gamma}{2BT}, s = 0, 1, 2, \cdots$, and the size of the cover we need to consider is only $\frac{\sqrt{m}}{e}$. Therefore we replace δ by $e\frac{\delta}{\sqrt{m}}$ in (27) and take a union bound over all the $\tilde{\beta}$ on the grid to complete the proof.

B. Details of Forward and Backward Propagation Algorithms

Algorithm 1 Local forward/backward propagation

```
Input: $h^{(t-1)}, \frac{\partial L}{\partial \tilde{h}^{(t)}}, U = (u_n|...|u_{n-m_1+1})$,
$\Sigma, V = (v_n|...|v_{n-m_2+1})$
Output: $\tilde{h}^{(t)} = Wh^{(t-1)}, \frac{\partial L}{\partial U}, \frac{\partial L}{\partial V}, \frac{\partial L}{\partial \hat{\sigma}}, \frac{\partial L}{\partial h^{(t-1)}}$
// Begin forward propagation
$h^{(v)}_{n+1} \leftarrow h^{(t-1)}$
for $k = n, n-1, ..., n-m_2+1$ do
    $h^{(v)}_k \leftarrow Hprod(h^{(v)}_{k+1}, v_k)$    // Compute $\hat{V}^\top h$
end for
$h^{(u)}_{k_1-1} \leftarrow \Sigma h^{(v)}_{k_2}$    // Compute $\Sigma\hat{V}^\top h$
for $k = n-m_1+1, ..., n$ do
    $h^{(u)}_k \leftarrow Hprod(h^{(u)}_{k-1}, u_k)$    // Compute $\hat{U}\Sigma\hat{V}^\top h$
end for
$\tilde{h}^{(t)} \leftarrow h^{(u)}_n$
//Begin backward propagation
$g \leftarrow \frac{\partial L}{\partial \tilde{h}^{(t)}}$
for $k = n, n-1, ..., n-m_1+1$ do
    $g, G^{(u)}_{*,n-k+1} \leftarrow Hgrad(h^{(u)}_k, u_k, g)$    // Compute $\frac{\partial L}{\partial u_k}$
end for
$\bar{\Sigma} \leftarrow diag(g \circ h^{(v)}_{k_2}), g \leftarrow \Sigma g$    // Compute $\frac{\partial L}{\partial \Sigma}$
$g^{(\hat{\sigma})} \leftarrow \frac{\partial diag(\Sigma)}{\partial \hat{\sigma}} \circ diag(\bar{\Sigma})$    // Compute $\frac{\partial L}{\partial \hat{\sigma}}$
for $k = n-m_2+1, ..., n$ do
    $g, G^{(v)}_{*,n-k+1} \leftarrow Hgrad(h^{(u)}_{k+1}, v_k, g)$    // Compute $\frac{\partial L}{\partial v_k}$
end for
$\frac{\partial L}{\partial U} \leftarrow G^{(u)}, \frac{\partial L}{\partial V} \leftarrow G^{(v)}, \frac{\partial L}{\partial \hat{\sigma}} \leftarrow g^{(\hat{\sigma})}, \frac{\partial L}{\partial h^{(t-1)}} \leftarrow g$
```

Algorithm 2
$\hat{h} = Hprod(h, u_k)$

```
Input: $h, u_k$
Output: $\hat{h} = \mathcal{H}_k(u_k)h$
// Compute $\hat{h} = (I - \frac{2u_k u_k^\top}{u_k^\top u_k})h$
$\alpha \leftarrow \frac{2}{\|u_k\|^2} u_k^\top h$
$\hat{h} \leftarrow h - \alpha u_k$
```

Algorithm 3
$\bar{h}, \bar{u}_k = Hgrad(h, u_k, g)$

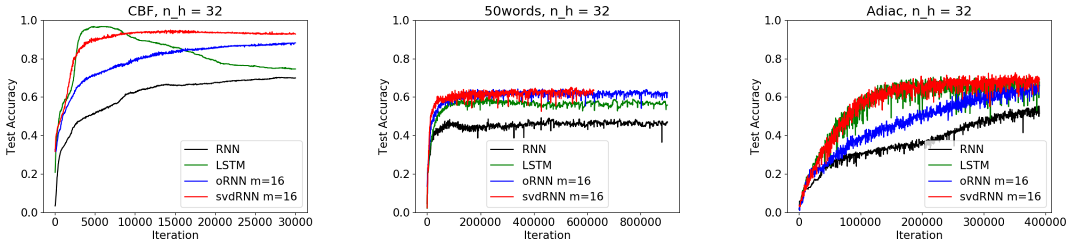
```
Input: $h, u_k, g = \frac{\partial L}{\partial \tilde{h}}$ where $\tilde{h} = \mathcal{H}_k(u_k)h$
Output: $\bar{h} = \frac{\partial L}{\partial h}, \bar{u}_k = \frac{\partial L}{\partial u_k}$
$\alpha = \frac{2}{\|u_k\|^2} u_k^\top h$
$\beta = \frac{2}{\|u_k\|^2} u_k^\top g$
$\bar{h} \leftarrow g - \beta u_k$
$\bar{u}_k \leftarrow -\alpha g - \beta h + \alpha\beta u_k$
```

C. More Experimental Details

C.1. Time Series Classification

In this experiment, we focus on the time series classification problem, where time series are fed into RNN sequentially, which then tries to predict the right class upon receiving the sequence end (Hüsken & Stagge, 2003). The dataset we choose is the largest public collection of class-labeled time-series with widely varying length, namely, the UCR time-series collection from (Chen et al., 2015). We use the training and testing sets directly from the UCR time series archive `http://www.cs.ucr.edu/~eamonn/time_series_data/`, and randomly choose 20% of the training set as validation data. We provide the statistical descriptions of the datasets and experimental results in Table 4.

In all experiments, we used hidden dimension $n_h = 32$, and chose total number of reflectors for oRNN and Spectral-RNN to be $m = 16$ (for Spectral-RNN $m_1 = m_2 = 8$). We choose proper depth t as well as input size n_i. Given sequence length L, since $tn_i = L$, we choose n_i to be the maximum divisor of L that satisfies $depth \leq \sqrt{L}$. To have a fair comparison of how the proposed principle itself influences the training procedure, we did not use dropout in any of these models. As illustrated in the optimization process in Figure 6, this resulted in some overfitting (see (a) CBF), but on the other hand it shows that Spectral-RNN is able to prevent overfitting. This supports our claim that since generalization is bounded by the spectral norm of the weights (Bartlett et al., 2017), Spectral-RNN will potentially generalize better than other schemes. This phenomenon is more drastic when the depth is large (e.g. ArrowHead(251 length) and FaceAll(131 length)), since regular RNN, and even LSTM, have no control over the spectral norms. Also note that there are substantially fewer parameters in oRNN and Spectral-RNN as compared to LSTM.



(a) (b) (c)

Figure 6. Performance comparisons of the RNN based models on three UCR datasets.

Datasets	Data Descriptions				Depth	RNN		LSTM		oRNN		Spectral-RNN	
	training/testing size		*length*	*#class*		acc	(n_{param})	acc	(n_{param})	acc	(n_{param})	acc	(n_{param})
50words	450	455	270	50	27	0.492	(3058)	0.598	(7218)	0.642	(2426)	**0.651**	(2850)
Adiac	390	391	176	37	16	0.552	(2694)	0.706	(6950)	0.668	(2062)	**0.726**	(2486)
ArrowHead	36	175	251	3	251	0.509	(1219)	0.537	(4515)	0.669	(587)	**0.800**	(1011)
Beef	30	30	470	5	47	0.600	(1606)	0.700	(5766)	**0.733**	(974)	**0.733**	(1398)
BeetleFly	20	20	512	2	32	**0.950**	(1699)	0.850	(6435)	0.900	(1067)	**0.950**	(1491)
CBF	30	900	128	3	16	0.702	(1476)	**0.967**	(5444)	0.881	(844)	0.948	(1268)
Coffee	28	28	286	2	22	**1.000**	(1570)	**1.000**	(6018)	**1.000**	(938)	**1.000**	(1362)
Cricket X	390	390	300	12	20	0.310	(1997)	0.456	(6637)	0.495	(1365)	**0.500**	(1789)
DistalPhalanxOutlineCorrect	276	600	80	2	10	0.790	(1410)	0.798	(5378)	0.830	(778)	**0.840**	(1202)
DistalPhalanxTW	154	399	80	6	10	**0.815**	(1641)	0.795	(5609)	0.807	(1009)	**0.815**	(1433)
ECG200	100	100	96	2	12	**0.640**	(1410)	**0.640**	(5378)	**0.640**	(778)	**0.640**	(1202)
ECG5000	500	4500	140	5	14	0.941	(1606)	0.936	(5766)	0.940	(974)	**0.945**	(1398)
ECGFiveDays	23	861	136	2	17	0.947	(1443)	0.790	(5411)	**0.976**	(811)	0.948	(1235)
FaceAll	560	1690	131	14	131	0.549	(1615)	0.455	(4911)	**0.714**	(983)	**0.714**	(1407)
FaceFour	24	88	350	4	25	0.625	(1701)	0.477	(6245)	0.511	(1069)	**0.716**	(1493)
FacesUCR	200	2050	131	14	131	0.449	(1615)	0.629	(4911)	0.710	(983)	**0.727**	(1407)
Gun Point	50	150	150	2	15	0.947	(1507)	0.920	(5667)	0.953	(875)	**0.960**	(1299)
InsectWingbeatSound	220	1980	256	11	16	0.534	(1996)	0.515	(6732)	**0.598**	(1364)	0.586	(1788)
ItalyPowerDemand	67	1029	24	2	6	0.970	(1315)	0.969	(4899)	0.972	(683)	**0.973**	(1107)
Lighting2	60	61	637	2	49	**0.541**	(1570)	**0.541**	(6018)	**0.541**	(938)	**0.541**	(1362)
MiddlePhalanxOutlineCorrect	291	600	80	2	10	0.793	(1410)	0.783	(5378)	0.712	(778)	**0.820**	(1202)

Table 4. Test accuracy (number of parameters) on UCR datasets. For each dataset, we present the testing accuracy when reaching the smallest validation error. The highest precision is in bold, and lowest two are colored gray.