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 $u_i \in \mathbb{R}^n$, such that $B = H_n^1(u_n) \cdots H_n^1(u_1) R$. (Note that we allow $u_i = 0$, in which case, $H_n^1(u_i) = I_n$ as in (1))

**Proof of Proposition 1.** For $n = 1$, note that $H_1^1(u_1) = \pm 1$. 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 $e_1 = (1, 0, ..., 0)^T$. Thus we have

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

where $\hat{B} \in \mathbb{R}^{k \times k}$. Note that $H_{k+1}^1(u_{k+1}) = I_{k+1}$ when $u_{k+1} = 0$ and the above still holds. By assumption we have $\hat{B} = H_k^1(u_k) \cdots H_1^1(u_1) \hat{R}$. Notice that $H_{k+1}^1(u_1) = \begin{pmatrix} 1 & H_k^1(u_1) \end{pmatrix}$, so we have that

$$
H_{k+1}^1(u_1) \cdots H_k^1(u_k) H_{k+1}^1(u_{k+1})B = \begin{pmatrix} \|B_1\| & \hat{B}_{1,2:k+1} \\
0 & \hat{R}
\end{pmatrix} = R
$$

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 $M_1$ is a subset of $O(n)$, and we now show that the converse is also true. Given $A \in O(n)$, by Proposition 1, there exists an upper triangular matrix $R$ with positive diagonal elements, and an orthogonal matrix $Q$ expressed as $Q = H_n^2(u_n) \cdots H_1^2(u_1)$ for some set of Householder vectors $\{u_i\}_{i=1}^n$, such that $A = QR$. Since $A$ is orthogonal, we have $A^TA = AAT = I_n$, thus:

$$
A^TA = R^TQ^TQR = R^TR = I_n; \ Q^TAAT = Q^TQRR^TQ = RR^T = I_n
$$

Thus $R$ is orthogonal and upper triangular matrix with positive diagonal elements. So $R = I_n$ and $A = Q = H_n^2(u_n) \cdots H_1^2(u_1)$. □

A.3. Proof of Theorem 2

**Proof.** It is easy to see that the image of $M_{1,1}$ is a subset of $\mathbb{R}^{n \times n}$. For any $W \in \mathbb{R}^{n \times n}$, we have its SVD, $W = U \Sigma V^T$, where $\Sigma = \text{diag}(\sigma)$. 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 $U = M_1(u_1, ..., u_n)$ and $V = M_1(v_1, ..., v_n)$, then we have:

$$
W = H_n^2(u_n) \cdots H_1^2(u_1) \Sigma H_1^1(v_1) \cdots H_n^1(v_n) = M_{1,1}(u_1, ..., u_n, v_1, ..., v_n, \sigma)
$$

□

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 $A = M_1(a_1, ..., a_n)$. Since $A^T$ is also orthogonal, for the same reason, there exist $\{b_i\}_{i=1}^n$, such that $A^T = M_1(b_1, ..., b_n)$. Thus we have:

$$
A = H_n(a_n) \cdots H_1(a_1) = H_1(b_1) \cdots H_n(b_n)
$$

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

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

(18)
We are looking at a recurrent neural network with:
\[ \Sigma T \phi \]
where
\[ \Sigma \]
To get a generalization bound for RNN, we need to use the following lemma from (Neyshabur et al., 2017).

**A.5. Proof of Theorem 4**

**Proof.** It is easy to see that the image of \( \mathcal{M}_{n,m}^{*} \) is a subset of \( \mathbb{R}^{m \times n} \). For any \( W \in \mathbb{R}^{m \times n} \), we have its SVD, \( W = U \Sigma V^T \), where \( \Sigma \) 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 \( U = H_m^m(u_m) \ldots H_1^m(u_1) \) and \( V = H_n^n(v_n) \ldots H_1^n(v_1) \). By Lemma 1, if \( m < n \) we have:
\[
W = H_m^m(u_m) \ldots H_1^m(u_1) \Sigma H_1^m(v_1) \ldots H_n^n(v_n) \\
= H_m^m(u_m) \ldots H_1^m(u_1) \Sigma H_{n-m+1}^n(v_{n-m+1}) \ldots H_n^n(v_n).
\]
Similarly, for \( n < m \), we have:
\[
W = H_m^m(u_m) \ldots H_1^m(u_1) \Sigma H_1^m(v_1) \ldots H_n^n(v_n) \\
= H_m^m(u_m) \ldots H_{m-n+1}^m(u_{m-n+1}) \Sigma H_1^n(v_1) \ldots H_n^n(v_n).
\]

**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 \( \gamma \) with \( m \) samples, i.e.,
\[
\hat{L}_\gamma(f_w) = \frac{1}{m} \sum_{i=1}^{m} \left[ f_w(x_i)[y_i] \leq \gamma + \max_{j \neq y_i} f_w(x_i)[j] \right].
\]
We are looking at a recurrent neural network with \( T \) time steps:
\[
h^{(t)} = \phi(W h^{(t-1)} + M x^{(t)}), h^{(0)} = 0, t = 1, 2, \ldots T \\
\hat{y}^{(t)} = Y h^{(t)},
\]
where \( \phi \) is the activation function. The dimensions are as follows: \( x^{(t)} \in \mathbb{R}^{n_t}, \hat{y}^{(t)} \in \mathbb{R}^{n_y} \), and \( h^{(t)} \in \mathbb{R}^{n} \). Therefore \( W \in \mathbb{R}^{n \times n}, M \in \mathbb{R}^{n_t \times n}, Y \in \mathbb{R}^{n_y \times n} \). 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, \( h^{(t)}_w \) denotes the activation that takes \( w = \text{vec} \{W, Y, M\} \) as parameters, and similar notation also holds for the output \( \hat{y}^{(t)}_w \). 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 $f_w(x) : \mathcal{X} \rightarrow \mathbb{R}^k$ 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. $P_u[\max_{x \in \mathcal{X}} \|f_w(x) - f_w(x)\|_\infty < \frac{2}{\delta}] \geq \frac{1}{2}$, 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 $w = \text{vec}(\{W, Y, M\})$, and perturbation $u = \text{vec}(\{\delta W, \delta Y, \delta M\})$ such that $\|\delta W\| \leq \frac{1}{T} \|W\|$, $\|\delta Y\| \leq \frac{2}{\delta} \|Y\|$, $\|\delta M\| \leq \frac{1}{T} \|M\|$. For a recurrent neural network (17) with $T$ time steps that satisfies Assumption 1, the perturbation in the activation is bounded by

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

while the perturbation in the output satisfies:

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

Here $e$ is the natural logarithm base.

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

$$\|h_w^{(t)}\| = \|\phi(Wh_w^{(t-1)} + Mx^{(t)})\|$$

$$\leq \|W\|\|h_w^{(t-1)}\| + \|M\|\|x^{(t)}\|$$

(by Assumption 1.2)

$$\leq \|W\|\|h_w^{(t-1)}\| + \|M\|\|x^{(t)}\|$$

(by triangle inequality)  \hspace{1cm} (23)

$$\leq \|W\|\left(\|W\|\|h_w^{(t-2)}\| + \|M\|\|x^{(t-1)}\|\right) + \|M\|\|x^{(t)}\|$$

(applying (23) to $\|h_w^{(t-1)}\|$)

$$\leq \cdots$$

$$\leq \|W\|^t\|h_w^{(0)}\| + \|M\| \sum_{j=0}^{t-1} \|W\|^{t-1-j} \|x^{(j+1)}\|$$

$$= \|M\| \sum_{j=0}^{t-1} \|W\|^{t-1-j} \|x^{(j+1)}\|$$

(since $h_w^{(0)} = 0$)

$$\leq B\|M\| \sum_{j=0}^{t-1} \|W\|^{t-1-j}$$

(by Assumption 1.1)

$$\Rightarrow \|h_w^{(t)}\| \leq B\|M\|t \max\{\|W\|^{t-1}, 1\}$$

(since $\sum_{i=0}^{t-1} \|W\|^i \leq t \max\{\|W\|^{t-1-i}, 1\}$)  \hspace{1cm} (24)

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,

$$\Delta_t \leq Bt(1 + \frac{1}{T})^{t-1} (\|\delta W\|\|M\|T + \|\delta M\|) \max\{\|W\|^{t-1}, 1\}, \forall t \leq T$$  \hspace{1cm} (25)
Then by induction and the bound of the activations, we have:

\[
\Delta_t \leq (1 + \frac{1}{T})||W|| \left( B(t-1)(1 + \frac{1}{T})^{t-2}(||W|| M||T + ||\delta M||) \max\{||W||^{t-2}, 1\} \right)
\]

(by induction)

\[
+ ||\delta W|| \left( B(t-1)||M|| \max\{||W||^{t-2}, 1\} \right) + B||\delta M||
\]

(by activation bound (24))

\[
= B(t-1)T(1 + \frac{1}{T})^{t-1}||\delta W|| ||M|| ||W|| \max\{||W||^{t-2}, 1\} + B(t-1)||\delta W|| ||M|| \max\{||W||^{t-2}, 1\}
\]

\[
+ (1 + \frac{1}{T})^{t-1}||W|| B(t-1) \max\{||W||^{t-2}, 1\} ||\delta M|| + B||\delta M||
\]

\[
= B(t-1)||\delta W|| ||M|| \max\{||W||^{t-2}, 1\} \left( 1 + \frac{1}{T} \right)^{t-1}||W|| + 1
\]

\[
+ B||\delta M|| \left( 1 + \frac{1}{T} \right)^{t-1} \max\{||W||^{t-2}, 1\} + 1\]

\[
\leq B(t-1)||\delta W|| ||M|| \max\{||W||^{t-2}, 1\} \left( 1 + \frac{1}{T} \right)^{t-1} \max\{||W||, 1\}
\]

\[
+ B||\delta M|| \left( 1 + \frac{1}{T} \right)^{t-1} \max\{||W||^{t-2}, 1\} + 1 \max\{||W||, 1\}
\]

\[
\leq B||\delta W|| ||M|| \max\{||W||^{t-1}, 1\}
\]

\[
+ B||\delta M|| \left( 1 + \frac{1}{T} \right)^{t-1} \max\{||W||^{t-2}, 1\}
\]

\[
\leq Bt(1 + \frac{1}{T})^{t-1}(T||\delta W|| ||M|| + ||\delta M||) \max\{||W||^{t-1}, 1\}
\]

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

\[
||\tilde{y}_{w+u}^{(T)} - \tilde{y}_{w}^{(T)}||
\]

\[
= ||(Y + \delta Y)h_{w+u}^{(T)} - Yh_{w}^{(T)}||
\]

\[
= ||(Y + \delta Y)(h_{w+u}^{(T)} - h_{w}^{(T)}) + (Y + \delta Y)h_{w}^{(T)} - Yh_{w}^{(T)}||
\]

\[
\leq ||(Y + \delta Y)||\Delta_T + ||\delta Yh_{w}^{(T)}||
\]

\[
\leq ||Y||\left( 1 + \frac{1}{T} \right) BT(1 + \frac{1}{T})^{T-1}(T||\delta W|| ||M|| + ||\delta M||) \max\{||W||^{T-1}, 1\}
\]

(by perturbation bound (25))

\[
+ ||\delta Y|| TB ||M|| \max\{||W||^{T-1}, 1\}
\]

(by activation bound (24))

\[
\leq TB \max\{||W||^{T-1}, 1\} \left( ||Y|| ||\delta W|| ||M|| ||T|| + ||Y|| ||\delta M|| ||e|| + ||\delta Y|| ||M|| \right)
\]

(since \((1 + \frac{1}{T})^{T} \leq e\)
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 $\beta = \max\{\|W\|_2^{-1}, 1\} \max\{\|Y\|_2, 1\} \max\{\|M\|_2, 1\}$. 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 $\sigma$ will be assigned later according to $\beta$. More precisely, since the prior cannot depend on the $\beta$ which is associated with the learned parameters $W, M$ and $Y$, we will set $\sigma$ based on some discrete choices of $\beta$ that approximates $\beta$. For each value of $\beta$ of our choice, we will compute the PAC-Bayes bound, establishing the generalization guarantee for all $\sigma$ for which $\frac{1}{2} \beta \leq \beta \leq e \beta$, and ensuring that each relevant value of $\beta$ is covered by some $\beta$ 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:

$$
P_{X \sim \mathcal{N}(0, \sigma^2 I)}(\|X\|_2 > t) \leq 2ne^{-t^2/2\sigma^2}, \forall n \geq n_1, n_2$$  \hspace{1cm} (26)

Therefore for $\delta W, \delta M, \delta Y$, the probability of their spectral norm being greater than $t$ is bounded by $2he^{-t^2/2\sigma^2}$, where $h = \max\{n, n_1, n_2\}$. Therefore with probability $\geq \frac{1}{2}$, $\|\delta W\|_2, \|\delta Y\|_2, \|\delta M\|_2 \leq \sigma \sqrt{2h \ln (12h)}$.

Plugging into Lemma 3 we have with probability at least $1/2$,

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

where we choose $\sigma = \frac{\sqrt{2}}{12 \sqrt{2h \ln (12h)}TB(Te+e+1)}$. 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.

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

Hence, with probability $\geq 1 - \delta$ and for all $w$ such that, $\frac{1}{2} \beta \leq \tilde{\beta} \leq e \beta$, we have:

$$
L_0(\hat{y}_w) \leq \hat{L}_\gamma(\hat{y}_w) + O\left( \sqrt{\frac{B(w) + \ln \frac{2m}{\delta}}{m}} \right),
$$

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

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 $\frac{1}{e} \beta \leq \tilde{\beta} \leq e \beta$. Firstly we notice for some range of $\beta$ 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.
\( \hat{y}^{(T)} = Y h_\gamma^{(T)} \), therefore if \( \beta \leq \frac{\gamma}{2BT} \),

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

Therefore \( \hat{L}_\gamma = 1 \) from definition of margin loss and the bound is satisfied trivially. Meanwhile, when \( \beta \geq \frac{\gamma}{2BT} \), then the second term of (27) \( \geq 1 \) and it also holds trivially. Therefore, we only need to consider \( \tilde{\beta} \) such that \( \tilde{\beta} \in [\frac{\gamma}{2BT}, \frac{2\gamma}{\sqrt{m}}] \).

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 \( \delta \) 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: $h^{(t-1)}$, $\frac{\partial L}{\partial h^{(t-1)}}$, $U = (u_n, u_{n-m_1+1})$, $\Sigma, V = (v_n, v_{n-m_2+1})$

Output: $\tilde{h}(t) = W h^{(t-1)}$, $\frac{\partial L}{\partial V}, \frac{\partial L}{\partial V}, \frac{\partial L}{\partial \Sigma}$

// Begin forward propagation

$h_{n+1}^{(v)} \leftarrow h^{(t-1)}$

for $k = n, n-1, ..., n-m_2+1$ do

$h_k^{(v)} \leftarrow H prod(h_{k+1}^{(v)}, v_k) // Compute \tilde{V}^T h$

end for

$h_{k-1}^{(v)} \leftarrow \Sigma h_k^{(v)} // Compute \Sigma \tilde{V}^T h$

for $k = n-m_1+1, ..., n$ do

$h_k^{(u)} \leftarrow H prod(h_{k-1}^{(u)}, u_k) // Compute \tilde{U} \tilde{V}^T h$

end for

// Begin backward propagation

$g \leftarrow \frac{\partial L}{\partial h^{(t-1)}}$

for $k = n, n-1, ..., n-m_1+1$ do

$g, G_{k+1}^{(u)} \leftarrow H grad(h_k^{(u)}, u_k, g) // Compute \frac{\partial L}{\partial u_k}$

end for

$\Sigma \leftarrow diag(g \circ h_k^{(v)})$, $g \leftarrow \Sigma g // Compute \frac{\partial L}{\partial \Sigma}$

$g_{(\Sigma)} \leftarrow \frac{\partial diag(\Sigma)}{\partial \Sigma} \circ diag(\Sigma) // Compute \frac{\partial L}{\partial \Sigma}$

for $k = n-m_2+1, ..., n$ do

$g, G_{k+1}^{(u)} \leftarrow H grad(h_k^{(u)}, v_k, g) // Compute \frac{\partial L}{\partial v_k}$

end for

$\frac{\partial L}{\partial V} \leftarrow G^{(v)}, \frac{\partial L}{\partial \Sigma} \leftarrow G^{(\Sigma)}$

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 use 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.
Stabilizing Gradients for Deep Neural Networks via Efficient SVD Parameterization

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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Data Descriptions</th>
<th>Depth</th>
<th>RNN acc (n_{param})</th>
<th>LSTM acc (n_{param})</th>
<th>oRNN acc (n_{param})</th>
<th>Spectral-RNN acc (n_{param})</th>
</tr>
</thead>
<tbody>
<tr>
<td>50words</td>
<td>450 455 270 50</td>
<td>27</td>
<td>0.492 (3058)</td>
<td>0.598 (7218)</td>
<td>0.642 (2426)</td>
<td>0.651 (2850)</td>
</tr>
<tr>
<td>Adiac</td>
<td>390 391 176 37</td>
<td>16</td>
<td>0.552 (2694)</td>
<td>0.706 (6950)</td>
<td>0.668 (2062)</td>
<td>0.726 (2486)</td>
</tr>
<tr>
<td>ArrowHead</td>
<td>36 175 251 3</td>
<td>251</td>
<td>0.509 (1219)</td>
<td>0.537 (4515)</td>
<td>0.669 (587)</td>
<td>0.800 (1011)</td>
</tr>
<tr>
<td>Beef</td>
<td>30 30 470 5</td>
<td>47</td>
<td>0.600 (1606)</td>
<td>0.700 (5766)</td>
<td>0.733 (974)</td>
<td>0.733 (1398)</td>
</tr>
<tr>
<td>BeetleFly</td>
<td>20 20 512 2</td>
<td>32</td>
<td>0.950 (1699)</td>
<td>0.850 (6435)</td>
<td>0.900 (1067)</td>
<td>0.950 (1491)</td>
</tr>
<tr>
<td>CBF</td>
<td>30 900 128 3</td>
<td>16</td>
<td>0.702 (1476)</td>
<td>0.967 (5444)</td>
<td>0.881 (844)</td>
<td>0.948 (1268)</td>
</tr>
<tr>
<td>Coffee</td>
<td>28 28 286 2 2</td>
<td>22</td>
<td>1.000 (1570)</td>
<td>1.000 (6018)</td>
<td>1.000 (938)</td>
<td>1.000 (1362)</td>
</tr>
<tr>
<td>Cricket X</td>
<td>390 390 300 12</td>
<td>20</td>
<td>0.310 (1997)</td>
<td>0.456 (6637)</td>
<td>0.495 (1365)</td>
<td>0.500 (1789)</td>
</tr>
<tr>
<td>DistalPhalanxOutlineCorrect</td>
<td>276 600 80 2</td>
<td>10</td>
<td>0.790 (1410)</td>
<td>0.798 (5378)</td>
<td>0.830 (778)</td>
<td>0.840 (1202)</td>
</tr>
<tr>
<td>DistalPhalanxTW</td>
<td>154 399 80 6</td>
<td>10</td>
<td>0.815 (1641)</td>
<td>0.795 (5609)</td>
<td>0.807 (1009)</td>
<td>0.815 (1433)</td>
</tr>
<tr>
<td>ECG200</td>
<td>100 100 96 2</td>
<td>12</td>
<td>0.640 (1410)</td>
<td>0.640 (5378)</td>
<td>0.640 (778)</td>
<td>0.640 (1202)</td>
</tr>
<tr>
<td>ECG5000</td>
<td>500 4500 140 5</td>
<td>14</td>
<td>0.941 (1606)</td>
<td>0.936 (5766)</td>
<td>0.940 (798)</td>
<td>0.945 (1398)</td>
</tr>
<tr>
<td>ECGFiveDays</td>
<td>23 861 136 2</td>
<td>17</td>
<td>0.947 (1443)</td>
<td>0.798 (5411)</td>
<td>0.976 (811)</td>
<td>0.948 (1235)</td>
</tr>
<tr>
<td>FaceAll</td>
<td>560 1690 131 14</td>
<td>131</td>
<td>0.549 (1615)</td>
<td>0.455 (4911)</td>
<td>0.714 (983)</td>
<td>0.714 (1407)</td>
</tr>
<tr>
<td>FaceFour</td>
<td>24 88 350 4</td>
<td>25</td>
<td>0.625 (1701)</td>
<td>0.477 (6245)</td>
<td>0.511 (1069)</td>
<td>0.716 (1493)</td>
</tr>
<tr>
<td>FaceUCR</td>
<td>200 2050 131 14</td>
<td>131</td>
<td>0.449 (1615)</td>
<td>0.629 (4911)</td>
<td>0.710 (983)</td>
<td>0.727 (1407)</td>
</tr>
<tr>
<td>GunPoint</td>
<td>50 150 150 2</td>
<td>15</td>
<td>0.947 (1507)</td>
<td>0.920 (5667)</td>
<td>0.953 (875)</td>
<td>0.960 (1299)</td>
</tr>
<tr>
<td>InsectWingbeatSound</td>
<td>220 1980 256 11</td>
<td>16</td>
<td>0.534 (1996)</td>
<td>0.515 (6732)</td>
<td>0.598 (1364)</td>
<td>0.586 (1788)</td>
</tr>
<tr>
<td>ItalyPowerDemand</td>
<td>67 1029 24 2</td>
<td>6</td>
<td>0.970 (1315)</td>
<td>0.969 (4899)</td>
<td>0.972 (683)</td>
<td>0.973 (1107)</td>
</tr>
<tr>
<td>Lighting2</td>
<td>60 61 637 2</td>
<td>49</td>
<td>0.541 (1570)</td>
<td>0.541 (6018)</td>
<td>0.541 (938)</td>
<td>0.541 (1362)</td>
</tr>
<tr>
<td>MiddlePhalanxOutlineCorrect</td>
<td>291 600 80 2</td>
<td>10</td>
<td>0.793 (1410)</td>
<td>0.783 (5378)</td>
<td>0.712 (778)</td>
<td>0.820 (1202)</td>
</tr>
</tbody>
</table>

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.