Abstract

This supplementary material contains proofs, detailed remarks and additional theoretic and numerical results to support the theory and claims in the main paper. We also repeat some necessary contents here for easy reference.

A. Detailed Theoretical Results

A.1. Proof of Theorem 1

**Assumption 1.** (i) The joint density function of \( \varepsilon(t) \) is continuous and positive everywhere; (ii) For some \( \kappa \geq 2 \), \( E[|\varepsilon(t)|^\kappa] < \infty \).

**Theorem 1.** Under Assumptions 1, if there exist real numbers \( 0 < a < 1 \) and \( b \) such that \( \|M(x)\| \leq a \|x\| + b \), then recurrent network process (7) is geometrically ergodic, and hence has short memory.

**Proof.** Let \( Y(t) = (y(t), s(t)) \) and \( r = p + q \). Rewrite model (7) as

\[
Y(t) = M(Y(t-1)) + \varepsilon(t),
\]

where \( Y(t), \varepsilon(t) \in \mathbb{R}^r \) and \( M : \mathbb{R}^r \to \mathbb{R}^r \) is a general nonlinear function.

Let \( B' \) be the class of Borel sets of \( \mathbb{R}^r \) and \( \nu_r \) be the Lebesgue measure on \( (\mathbb{R}^r, B') \). Then, \( \{Y(t)\} \) is a homogeneous Markov chain on the state space \( (\mathbb{R}^r, B', \nu_r) \) with the transition probability

\[
P(x, A) = \int_A f(z - M(x)) dz, \quad x \in \mathbb{R}^r \text{ and } A \in B',
\]

where \( f(\cdot) \) is the density of \( \varepsilon(t) \). Observe that, from Assumption 1, the transition density kernel in (2) is positive everywhere, and thus \( \{Y(t)\} \) is \( \nu_r \)-irreducible.

We prove by showing that Tweedie’s drift criterion (Tweedie, 1983) holds, i.e. there exists a small set \( G \) with \( \nu_r(G) > 0 \) and a non-negative continuous function \( \psi(x) \) such that

\[
E\{\psi(Y(t)) | Y(t-1) = x\} \leq (1 - \epsilon) \psi(x), \quad x \notin G,
\]

and

\[
E\{\psi(Y(t)) | Y(t-1) = x\} \leq M, \quad x \in G,
\]

for some \( 0 < \epsilon < 1 \) and \( 0 < M < \infty \).

Given that \( \|M(x)\| \leq a \|x\| + b \), where \( a < 1 \), we have

\[
E\left( \|Y(t)\|^\kappa \right) \leq \|M(x)\|^\kappa + E[|\varepsilon(t)|^\kappa]
\]

Define test function \( \psi(x) = 1 + \|x\|^\kappa > 0 \). Then,

\[
E\left( \psi(Y(t)) | Y(t-1) = x\right)
\]

\[
\leq 1 + |a|^\kappa \|x\|^\kappa + |b|^\kappa + E[|\varepsilon(t)|^\kappa]
\]

\[
\leq \rho \psi(x) + 1 + |a|^\kappa \|x\|^\kappa + |b|^\kappa + E[|\varepsilon(t)|^\kappa],
\]

where \( \rho = |a| < 1 \).

Denote \( \epsilon = 1 - \rho - \frac{1 - \rho + |b|^\kappa + E[|\varepsilon(t)|^\kappa]}{\psi(x)} \) and \( G = \{x : \|x\| \leq L\} \) such that \( \psi(x) > 1 + \frac{|b|^\kappa + E[|\varepsilon(t)|^\kappa]}{\psi(x)} \) for all \( \|x\| > L \). We obtain that conditions (3) and (4) hold.

Moreover, \( E(\phi(Y(t))) | Y(t-1) = x\) is continuous with respect to \( x \) for any bounded continuous function \( \phi(\cdot) \), then \( \{Y(t)\} \) is a Feller chain. By Feigin & Tweedie (1985), \( G \) is a small set. By referring to Theorem 4(ii) in Tweedie (1983) and Theorem 1 in Feigin & Tweedie (1985), \( \{Y(t)\} \) is geometrically ergodic with a unique strictly stationary solution.

A.2. Proof of Theorem 2

**Theorem 2.** Under Assumption 1, linear recurrent network process (8) is geometrically ergodic if and only if spectral radius \( \rho(W) < 1 \). Model (8) hence has short memory.

**Proof.** Proof of a similar result might exist in the literature, but we are unaware of the specific paper(s). For the convenience of the readers, we outline the proof here.
Let the Markov chain \( \{Y^{(t)}\} \) and its state space be defined as in (i). Under the linear setting, model (7) can be written as
\[
Y^{(t)} = WY^{(t-1)} + e^{(t)},
\]
where \( Y^{(t)} \in \mathbb{R}^r \) and \( W \in \mathbb{R}^{r \times r} \), and the transition probability can be written as
\[
P(x, A) = \int_A f(z - Wx) dx, \quad x \in \mathbb{R}^r \text{ and } A \in \mathcal{B}^r. \tag{6}
\]
Under Assumption 1, \( \{Y^{(t)}\} \) is \( \nu_r \)-irreducible.

First, suppose \( \rho(W) < 1 \). Then, there exists an integer \( s \) such that \( \|W^s\| < 1 \). In the following, we prove that \( s \)-step Markov chain \( \{Y^{(ts)}\} \) satisfies Tweedie’s drift criterion (Tweedie, 1983), i.e., there exists a small set \( G \) with \( \nu_r(G) > 0 \) and a non-negative continuous function \( \psi(x) \) such that
\[
E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq (1 - \epsilon)\psi(x), \quad x \notin G, \tag{7}
\]
and
\[
E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq M, \quad x \in G, \tag{8}
\]
for some constant \( 0 < \epsilon < 1 \) and \( 0 < M < \infty \).

We iterate (5) \( s \) times and obtain
\[
Y^{(ts)} = W^s Y^{((t-1)s)} + \left( e^{(ts)} + \sum_{j=1}^{s-1} W^j e^{(ts-j)} \right).
\]
Let \( g(x) = 1 + \|x\|^\alpha \), and it can be verified that
\[
E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} = x
\]
\[
\leq 1 + \|W^s\|^\alpha \|x\|^\alpha + E\left(e^{(ts)} + \sum_{j=1}^{s-1} W^j e^{(ts-j)}\right)
\]
\[
\leq \psi(x) \|W^s\|^\alpha + C,
\]
where \( C = 1 + E(e^{(ts)} + \sum_{j=1}^{s-1} W^j e^{(ts-j)}) - \|W^s\|^\alpha < \infty \).

Note that \( \|W^s\|^\alpha \) is finite. Then there exists \( L > 0 \), such that
\[
E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq (1 - \epsilon)\psi(x), \quad \forall \|x\| > L,
\]
and
\[
E\{\psi(Y^{(ts)})|Y^{((t-1)s)} = x\} \leq M < \infty, \quad \forall \|x\| \leq L.
\]
and \( G = \{x : \|x\| \leq L\} \) with \( \nu_r(G) > 0 \).

Moreover, because for each bounded continuous function \( \phi(\cdot) \), \( E\phi(Y^{(ts)})|Y^{((t-1)s)} = x \) is continuous with respect to \( x \), \( \{Y^{(ts)}\} \) is a Feller chain. And \( \{Y^{(ts)}\} \) is \( \nu_{r^*} \)-irreducible. This implies that \( G \) is a small set (Feigin & Tweedie, 1985). By referring to Theorem 4(ii) in Tweedie (1983), we can show that \( \{Y^{(ts)}\} \) is geometrically ergodic with a unique strictly stationary solution. By Lemma 3.1 of Tjøstheim (1990), \( \{Y^{(t)}\} \) is geometrically ergodic.

Then, we prove the necessity. Suppose that model (5) is geometrically ergodic, then there exists a strictly stationary solution \( \{Y^{(t)}\} \) to model (5) (Feigin & Tweedie, 1985).

And then the Markov chain \( Y^{(t)} \) have a stationary distribution \( \pi(\cdot) \), from which we can generate \( Y^{(0)} \), and iteratively obtain the sequence \( \{Y^{(t)}\} \). It is nonanticipative and equation (5) holds.

From (6), it holds that
\[
P(Y^{(t)} \in A|Y^{(t-1)} = x) = P(x, A) > 0
\]
as \( \nu_r(A) > 0 \). Let \( H \) be any affine invariant subspace of \( \mathbb{R}^r \) under model (5), i.e., \( \{Wx + e^{(t)} : x \in H\} \subseteq H \) with probability one. If \( \nu_r(R^r - H) \neq 0 \), then for any \( x \in H \), \( P(Wx + e^{(t)} \in H) < 1 \). As a result, \( \mathbb{R}^r \) is the unique affine invariant subspace, and hence model (5) is irreducible. Thus, by Theorem 2.5 in Bougerol & Picard (1992), we have that the the top Lyapounov exponent is strictly negative, and thus spectral radius \( \rho(W) = \|W^s\|^{1/s} < 1 \). This completes the proof of (ii).

A.3. Proof of Corollary 1

**Corollary 1.** Suppose that the output and activation functions, \( g(\cdot) \) and \( \sigma(\cdot) \), at (10) are continuous and bounded. If Assumption 1 holds, then the RNN process is geometrically ergodic and has short memory.

**Proof.** Need to show that there always exist real numbers \( a < 1 \) and \( b \) such that \( \|M_\text{RNN}(u, v)\| \leq a \|(u', v')\| + b \).

Since \( g(\cdot) \) and \( \sigma(\cdot) \) are bounded, there exist positive constants \( M_1 \) and \( M_2 \) such that \( \|g(Wzh\sigma(W_{hh}v + W_{bh}u + b_h) + b_z)\|_{l_1} \leq M_1, \|\sigma(W_{hh}v + W_{bh}u + b_h)\|_{l_1} \leq M_2 \) for any \( u \in \mathbb{R}^p, v \in \mathbb{R}^q \).

Let \( a = a_0 \in (0, 1) \) and \( b = M_1 + M_2 \), we have
\[
\|M_\text{RNN}(u, v)\|_{l_1} - a_0 \|(u', v')\|_{l_1} \leq M_1 + M_2 - a_0
\]
\[
\|u\|_{l_1} - a_0 \|v\|_{l_1} \leq b = M_1 + M_2. \tag{10}
\]

By Theorem 1, model (10) with bounded and continuous output and activation function is geometrically ergodic and has short memory.

A.4. Apply Theorem 1 to LSTM networks with \( p = q = 1 \)

We use an LSTM process with \( p = q = 1 \) as an example to illustrate the application of Theorem 1 to LSTM networks, and prepare readers for Corollary 2. Assume that the norm \( \| \cdot \| \) in Theorem 1 is the \( l_1 \) norm. Although sigmoid is used by default as the activation functions for the gates, we also
consider \( \sigma(\cdot) \) as ReLU or tanh for theoretical interests here. For output function \( g(\cdot) \), we consider commonly used linear, sigmoid and softmax functions. We summarize our results in Table A1.

A.5. Proof of Corollary 2

**Corollary 2.** The input series features \( \{y^{(t-1)}\} \) are scaled to the range of \([-1, 1]\). Suppose that \( M := \sup_{x \in B^2_\infty} \|g(W_{zh}x + b_z)\|_1 < \infty \) and \( \sigma(|W_{fh}||l_\infty + \|W_{fy}\|l_\infty + \|b_f\|l_\infty) \leq A \) for some \( A > 1 \), where \( B^2_\infty \) is the \( d \)-dimensional \( l_\infty \)-ball and \( \|W\|l_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |w_{ij}| \) is the matrix \( l_\infty \)-norm. If Assumption 1 holds, then the LSTM process at (14) is geometrically ergodic and has short memory.

**Proof.** Let \( a = a_0 \in (0, 1) \) and \( b = M + 2q \), we have

\[
\begin{align*}
\|M_{LSTM}(u, v, w)\|_{l_1} - a_0 \|u', v', w'\|_{l_1} &\leq \|g(W_{zh}x + b_z)\|_{l_1} + \|x\|_{l_1} + \|q + f(u, v) \omega w\|_{l_1} - a_0 \|u', v', w'\|_{l_1} \\
&\leq M + q + \|f(u, v)\|_{l_\infty} \|w\|_{l_1} - a_0 \|u', v', w'\|_{l_1} \leq M + 2q - a_0 \|u\|_{l_1} - a_0 \|w\|_{l_1} \leq b = M + 2q,
\end{align*}
\]

where \( x = o(u, v) \odot \tanh(i(u, v) \odot \tanh(W_{ch}v + W_{cy}u + b_c) + f(u, v) \omega w) \in B^2_\infty, v = h(t) \in B^2_\infty, \) and \( u = y^{(t-1)} \in B^2_\infty \). The second inequality holds due to the definition of \( M \) and \( x \in B^2_\infty \). The forth inequality holds due to \( \|f(u, v)\|_{l_\infty} = \sigma(|W_{fh}v + W_{fy}u + b_f|) \leq \sigma(|W_{fh}|_{l_\infty} + |W_{fy}|_{l_\infty} + \|b_f\|_{l_\infty}) \leq A_0 \). By Theorem 1, model (14) is geometrically ergodic and has short memory.

A.6. Proof of Theorem 3

**Theorem 3.** In terms of Definition 3, the MRNNF has the capability of handling long-range dependence data, while the RNN cannot.

**Proof.** Without loss of generality, assume that the linear activation and output functions are identity.

(1) The RNN process can be written as

\[
\begin{align*}
y(t) &= W_{zh}h(t) + \varepsilon(t) \\
h(t) &= W_{hh}h(t-1) + W_{hx}x(t)
\end{align*}
\]

Then, \( h(t) = (I - W_{hh}B)^{-1}W_{hx}x(t) \), and we have
\[
y(t) = W_{zh}(I - W_{hh}B)^{-1}W_{hx}x(t) + \varepsilon(t).
\]

Let \( y(t) = \sum_{k=0}^{\infty} A_k x^{(t-k)} + \varepsilon(t) \). Since \( (I - W_{hh}B)^{-1} = \sum_{k=0}^{\infty} W_{kh}^k B^k \), we have \( A_k = W_{zh}W_{kh}W_{hx} \), and \( (A_k)_{ij} \) decays exponentially for all \( i, j \).

(2) The MRNNF process can be written as

\[
\begin{align*}
y(t) &= W_{zh}h(t) + W_{zm}m(t) + \varepsilon(t) \\
h(t) &= W_{hh}h(t-1) + W_{hz}x(t) \\
m(t) &= W_{mm}m(t-1) + W_{mf}((I - B)^d - I)x(t)
\end{align*}
\]

Then,
\[
\begin{align*}
h(t) &= (I - W_{hh}B)^{-1}W_{hz}x(t) \\
m(t) &= (I - W_{mm}B)^{-1}W_{mf}((I - B)^d - I)x(t)
\end{align*}
\]

Let \( y(t) = \sum_{k=0}^{\infty} C_k x^{(t-k)} + \varepsilon(t) \), then \( A_k = C_k + D_k \), where
\[
\begin{align*}
\sum_{k=0}^{\infty} C_k x^{(t-k)} &= W_{zh}(I - W_{hh}B)^{-1}W_{hz}x(t) \\
\sum_{k=0}^{\infty} D_k x^{(t-k)} &= W_{zm}(I - W_{mm}B)^{-1}W_{mf}((I - B)^d - I)x(t)
\end{align*}
\]

From part (1) we know that the entries in \( C_k \) decay exponentially as well as the entries in the first part \( W_{zm}(I - W_{mm}B)^{-1} \) in \( D_k \). Since \( (I - B)^d - I = \sum_{k=0}^{\infty} W_k B^k \) and \( W_k \)’s are diagonal matrices with \( (W_k)_{ii} \sim k^{-d_i-1} \), the decay of \( D_k \) is dominated by \( (I - B)^d - I \), and entries of \( C_k \) decay by some polynomial rate \( k^{-d_i-1} \).

A.7. Property of Constant-gates-LSTM and Constant-gates-MLSTM

**Theorem 4.** In terms of Definition 3, the constant-gates-MLSTM has the capability of handling long-range dependent data, while the constant-gates-LSTM cannot.

**Proof.** Without loss of generality, assume that the linear activation and output functions are identity.

(1) The constant-gates-LSTM process can be written as

\[
\begin{align*}
y(t) &= W_{zh}h(t) + \varepsilon(t) \\
oc(t) &= W_{ch}h^{(t-1)} + W_{cx}x^{(t)} \\
\hat{c}(t) &= D_t \hat{c}^{(t-1)} + D_f c^{(t-1)} \\
\hat{h}(t) &= D_o \hat{c}(t)
\end{align*}
\]

where \( D_t, D_f \) and \( D_o \) are matrices obtained by diagonalize the constant gates.

Then, \( (I - D_f B) c^{(t)} = D_t \hat{c}(t) = D_t(W_{ch}h^{(t-1)} + W_{cx}x^{(t)}) = D_t W_{ch}D_o \hat{c}^{(t-1)} + D_t W_{cx}x^{(t)} \), and we have
\[
(I - (D_f + D_t W_{ch}D_o)B)\hat{c}^{(t-1)} = D_t W_{cx}x^{(t)}.
\]

Thus, \( c^{(t)} = (I - (D_f + D_t W_{ch}D_o)B)^{-1}D_t W_{cx}x^{(t)} \), then \( y^{(t)} = W_{zh}D_o(I - (D_f + D_t W_{ch}D_o)B)^{-1}D_t W_{cx}x^{(t)} + \varepsilon(t) \).

From the proof of Theorem 3 (1) we know that writing
The constant-gates-MLSTM process can be written as
\[ (I - B)^d c(t) = D_i c(t), \]
\[ h(t) = D_0 c(t). \]

Then, \((I - B)^d c(t) = D_i(W_{ch} h(t-1) + W_{cx} x(t)) = D_i W_{ch} D_0 c(t-1) + D_i W_{cx} x(t),\) and we have \((I - B)^d - D_i W_{ch} D_0 B)^c(t) = D_i W_{cx} x(t).\) Thus, \(c(t) = ((I - B)^d - D_i W_{ch} D_0 B)^{-1} D_i W_{cx} x(t),\) then
\[ y(t) = W_{zh} D_0 ((I - B)^d - D_i W_{ch} D_0 B)^{-1} D_i W_{cx} x(t) + \varepsilon(t). \]

Now we need to obtain the rate of polynomial \((I - B)^d - CB)^{-1}\) for some matrix \(C = D_i W_{ch} D_0.\) Let \((I - B)^d - CB)^{-1} = \sum_{j=0}^{\infty} \Theta_j B^j,\) then \((\sum_{j=0}^{\infty} \Theta_j B^j) ((I - B)^d - CB) = I.\) Thus,
\[
\begin{align*}
\sum_{j=0}^{\infty} \Theta_j B^j ((I - B)^d - CB)^{-1} &= I + C \sum_{j=0}^{\infty} \Theta_j B^{j+1} \\
\sum_{j=0}^{\infty} \Theta_j B^j (\sum_{k=0}^{\infty} W_k B^k) &= I + C \sum_{j=0}^{\infty} \Theta_j B^{j+1} \\
\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \Theta_j B^j W_k B^k &= I + C \sum_{j=0}^{\infty} \Theta_j B^{j+1}.
\end{align*}
\]

Equate the coefficients for each \(B^j\) term for \(j = 0, 1, 2, \ldots,\) we have
\[
\begin{align*}
\Theta_0 &= I \\
\Theta_1 &= C - W_1 \\
\Theta_2 &= C \Theta_1 - W_1 \Theta_1 - W_2 \\
\Theta_3 &= C \Theta_2 - W_1 \Theta_2 - W_2 \Theta_1 - W_3 \\
&\quad \vdots \\
\Theta_k &= C \Theta_{k-1} - \sum_{j=1}^{k} W_j \Theta_{k-j}
\end{align*}
\]
The \(\Theta_k\)'s are dominated by the \(W_k\) term and the elements decay at the same rate as \(W_k,\) which is \(k^{-d_j-1}.\)
Supplementary Materials for Do RNN and LSTM have Long Memory?

Figure 1. Autocorrelation plots of all 4 datasets.

(a) ARFIMA  
(b) DJI  
(c) traffic  
(d) tree

in Figure 3 for datasets ARFIMA, DJI, traffic and tree.

MAPE  Average MAPE and standard deviation of one-step forecasting is shown in Table 3.

Boxplot of MAPE for 100 different initializations are shown in Figure 4 for datasets ARFIMA, DJI, traffic and tree.

B.3. Best Performance of the Models

Best performance of the models, in terms of MAE and MAPE, are shown in Table 4 & 5.

B.4. Performance on a Dataset without Long Memory

We generated a sequence of length 4001 (2000 + 1200 + 800) using model (10), which does not have long memory according to Corollary 1. We refer to this synthetic dataset as the RNN dataset. The boxplots of error measures are presented in Figure 5. From the boxplots we can see that the performance of our proposed models is comparable with that of the true model RNN, except that the variation of the error measures is a bit larger.

B.5. Experiment on Parameter $K$

Boxplot of RMSE for 100 different initializations are shown in Figure 6, 7, 8 and 9 for datasets ARFIMA, DJI, traffic and tree, respectively. Values of $K$ are appended to the abbreviations of the proposed models to distinguish the settings. For example, model “MRNN25” means the MRNN model with $K = 25$. There are 20 models with different settings in total, and they are sorted by the average RMSE in ascending order from left to right.

For MRNN and MRNNF, the prediction is generally better for a larger $K$, and they have smaller average RMSE than all the baseline models regardless of the choice of $K$. Interestingly, the performance of MLSTM and MLSTMF gets better when $K$ becomes smaller, and with $K = 25$, they can outperform LSTM on ARFIMA and traffic datasets. Thus, we recommend a large $K$ for MRNN and MRNNF models, while for the more complicated MLSTM models, $K$ deserves more investigation to balance expressiveness and optimization.
Table 3. Overall performance in terms of MAPE. Average MAPE and the standard deviation (in brackets) are reported.

<table>
<thead>
<tr>
<th>Model</th>
<th>ARFIMA</th>
<th>DJI (x100)</th>
<th>Traffic</th>
<th>Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>2.5760</td>
<td>1.4371</td>
<td>1.3943</td>
<td>0.2747</td>
</tr>
<tr>
<td></td>
<td>(0.4030)</td>
<td>(0.2566)</td>
<td>(0.1998)</td>
<td>(0.0079)</td>
</tr>
<tr>
<td>RNN2</td>
<td>2.5570</td>
<td>1.4407</td>
<td>1.4092</td>
<td>0.2739</td>
</tr>
<tr>
<td></td>
<td>(0.4420)</td>
<td>(0.2106)</td>
<td>(0.1789)</td>
<td>(0.0071)</td>
</tr>
<tr>
<td>RWA</td>
<td>2.2370</td>
<td>1.2733</td>
<td>1.3745</td>
<td>0.2939</td>
</tr>
<tr>
<td></td>
<td>(0.1950)</td>
<td>(0.1702)</td>
<td>(0.1457)</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>MRNNF</td>
<td>2.6430</td>
<td>1.5561</td>
<td>1.4270</td>
<td>0.2714</td>
</tr>
<tr>
<td></td>
<td>(0.3380)</td>
<td>(0.2243)</td>
<td>(0.1834)</td>
<td>(0.0042)</td>
</tr>
<tr>
<td>MRNN</td>
<td>2.7010</td>
<td>1.5031</td>
<td>1.4253</td>
<td>0.2706</td>
</tr>
<tr>
<td></td>
<td>(0.2680)</td>
<td>(0.2045)</td>
<td>(0.1586)</td>
<td>(0.0044)</td>
</tr>
<tr>
<td>LSTM</td>
<td>2.5660</td>
<td>1.5725</td>
<td>1.3632</td>
<td>0.2727</td>
</tr>
<tr>
<td></td>
<td>(0.3750)</td>
<td>(0.2283)</td>
<td>(0.1807)</td>
<td>(0.0060)</td>
</tr>
<tr>
<td>MLSTM</td>
<td>2.5100</td>
<td>1.3141</td>
<td>1.3462</td>
<td>0.2750</td>
</tr>
<tr>
<td></td>
<td>(0.4690)</td>
<td>(0.1369)</td>
<td>(0.1769)</td>
<td>(0.0074)</td>
</tr>
<tr>
<td>MLSTM</td>
<td>2.5500</td>
<td>1.3123</td>
<td>1.3353</td>
<td>0.2748</td>
</tr>
<tr>
<td></td>
<td>(0.4370)</td>
<td>(0.1281)</td>
<td>(0.1926)</td>
<td>(0.0075)</td>
</tr>
</tbody>
</table>

Table 4. Best performance in terms of MAE.

<table>
<thead>
<tr>
<th>Model</th>
<th>ARFIMA</th>
<th>DJI (x100)</th>
<th>Traffic</th>
<th>Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARFIMA</td>
<td>0.8190</td>
<td>0.1800</td>
<td>230.99</td>
<td>0.2174</td>
</tr>
<tr>
<td>RNN</td>
<td>0.8378</td>
<td>0.1667</td>
<td>213.96</td>
<td>0.2185</td>
</tr>
<tr>
<td>RNN2</td>
<td>0.8196</td>
<td>0.1667</td>
<td>212.69</td>
<td>0.2186</td>
</tr>
<tr>
<td>RWA</td>
<td>1.3307</td>
<td>0.1862</td>
<td>227.01</td>
<td>0.2378</td>
</tr>
<tr>
<td>MRNNF</td>
<td>0.8179</td>
<td>0.1649</td>
<td>214.88</td>
<td>0.2172</td>
</tr>
<tr>
<td>MRNN</td>
<td>0.8171</td>
<td>0.1654</td>
<td>213.79</td>
<td>0.2170</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8197</td>
<td>0.1671</td>
<td>214.22</td>
<td>0.2170</td>
</tr>
<tr>
<td>MLSTM</td>
<td>0.8191</td>
<td>0.1716</td>
<td>216.15</td>
<td>0.2177</td>
</tr>
<tr>
<td>MLSTM</td>
<td>0.8193</td>
<td>0.1702</td>
<td>216.12</td>
<td>0.2175</td>
</tr>
</tbody>
</table>

References


**Supplementary Materials for Do RNN and LSTM have Long Memory?**

Figure 3. Boxplot of MAE for 100 different initializations.

Table 5. Best performance in terms of MAPE.

<table>
<thead>
<tr>
<th></th>
<th>ARFIMA</th>
<th>DJI</th>
<th>Traffic</th>
<th>Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARFIMA</td>
<td>2.8424</td>
<td>1.8334</td>
<td>1.6942</td>
<td>0.2676</td>
</tr>
<tr>
<td>RNN</td>
<td>1.5729</td>
<td>1.0789</td>
<td>1.0075</td>
<td>0.2680</td>
</tr>
<tr>
<td>RNN2</td>
<td>1.5905</td>
<td>1.0714</td>
<td>1.0215</td>
<td>0.2680</td>
</tr>
<tr>
<td>RWA</td>
<td>1.4408</td>
<td>1.0091</td>
<td><strong>0.9986</strong></td>
<td>0.2923</td>
</tr>
<tr>
<td>MRNNF</td>
<td>1.6508</td>
<td>1.0304</td>
<td>1.0816</td>
<td>0.2670</td>
</tr>
<tr>
<td>MRNN</td>
<td>1.5967</td>
<td>1.1303</td>
<td>1.0938</td>
<td><strong>0.2668</strong></td>
</tr>
<tr>
<td>LSTM</td>
<td>1.5282</td>
<td><strong>0.9918</strong></td>
<td>1.0099</td>
<td>0.2675</td>
</tr>
<tr>
<td>MLSTMF</td>
<td>1.5565</td>
<td>1.0368</td>
<td>0.9990</td>
<td>0.2673</td>
</tr>
<tr>
<td>MLSTM</td>
<td>1.5597</td>
<td>1.0518</td>
<td>1.0098</td>
<td>0.2673</td>
</tr>
</tbody>
</table>
Figure 4. Boxplot of MAPE for 100 different initializations.
Figure 5. Boxplot of RMSE, MAE and MAPE for 100 different initializations. Dataset: RNN.

Figure 6. Boxplot of RMSE for 100 different initializations. Dataset: ARFIMA.
Supplementary Materials for Do RNN and LSTM have Long Memory?

Figure 7. Boxplot of RMSE for 100 different initializations. Dataset: DJI.

Figure 8. Boxplot of RMSE for 100 different initializations. Dataset: traffic.

Figure 9. Boxplot of RMSE for 100 different initializations. Dataset: tree.