Foundations of Sequence-to-Sequence Modeling for Time Series

Zelda Mariet
zelda@csail.mit.edu
Massachusetts Institute of Techology

Vitaly Kuznetsov
vitalyk@google.com
Google AI

Abstract

The availability of large amounts of time series data, paired with the performance of deep-learning algorithms on a broad class of problems, has recently led to significant interest in the use of sequence-to-sequence models for time series forecasting. We provide the first theoretical analysis of this time series forecasting framework. We include a comparison of sequence-to-sequence modeling to classical time series models, and as such our theory can serve as a quantitative guide for practitioners choosing between different modeling methodologies.

1 Introduction

Time series analysis is a critical component of real-world applications such as climate modeling, web traffic prediction, neuroscience, as well as economics and finance. We focus on the fundamental question of time series forecasting. Specifically, we study the task of forecasting the next ℓ steps of an m-dimensional time series $\mathbf{Y}$, where m is assumed to be very large. For example, in climate modeling, m may correspond to the number of locations at which we collect historical observations, and more generally to the number of sources which provide us with time series.

Often, the simplest way to tackle this problem is to approach it as m separate tasks, where for each of the m dimensions we build a model to forecast the univariate time series corresponding to that dimension. Autoregressive and state-space models (Engle, 1982; Bollerslev, 1986; Brockwell and Davis, 1986; Box and Jenkins, 1990; Hamilton, 1994), as well as non-parametric approaches such as RNNs (Bianchi et al., 2017), are often used in this setting. To account for correlations between different time series, these models have also been generalized to the multivariate case (Lütkepohl, 2006, 2007; Song and Bickel, 2011; Han et al., 2015a,b; Banbura et al., 2010; Basu and Michailidis, 2015; Song and Bickel, 2011; Basu and Michailidis, 2015; W. Sun, 2015; Negahban and Wainwright, 2011; Yu et al., 2016; Lv et al., 2015). In both univariate and multivariate settings, an observation at time t is treated as a single sample point, and the model tries to capture relations between observations at times t and $t+1$. Therefore, we refer to these models as *local*.

In contrast, an alternative methodology based on treating m univariate time series as m samples drawn from some unknown distribution has also gained popularity in recent years. In this setting, each of the m dimensions of $\mathbf{Y}$ is treated as a separate example and a single model is learned from these m observations. Given m time series of length T, this model learns to map past vectors of length $T-\ell$ to corresponding future vectors of length ℓ. LSTMs and RNNs (Hochreiter and Schmidhuber, 1997) are a popular choice of model class for this setup (Flunkert et al., 2017; Goel et al., 2017; Yu et al., 2017; Li et al., 2017; Laptev et al., 2017; Zhu and Laptev, 2017).[1] Consequently, we refer to this framework as *sequence-to-sequence modeling*.

While there has been progress in understanding the generalization ability of local models (Yu, 1994; Meir and Hellerstein, 2000; Mohri and Rostamizadeh, 2009, 2010; Kuznetsov and Mohri, 2014, 2015, 2016, 2017; Zimin and Lampert, 2017), to the best of our knowledge the generalization properties of sequence-to-sequence modeling have not yet been studied, raising the following natural questions:

- What is the generalization ability of sequence-to-sequence models and how is it affected by the statistical properties of the underlying stochastic processes (e.g. non-stationarity, correlations)?

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

[1] Sequence-to-sequence models are also among the winning solutions in the recent time series forecasting competition: https://www.kaggle.com/c/web-traffic-time-series-forecasting.

– When is sequence-to-sequence modeling preferable to local modeling, and vice versa?

We provide the first generalization guarantees for time series forecasting with sequence-to-sequence models. Our results are expressed in terms of simple, intuitive measures of non-stationarity and correlation strength between different time series and hence explicitly depend on the key components of the learning problem.

We compare our generalization bounds to guarantees for local models and identify regimes under which one methodology is superior to the other. Therefore, our theory may also serve as a quantitative guide for a practitioner choosing the right modeling approach.

The rest of the paper is organized as follows: in Section 2, we formally define sequence-to-sequence and local modeling. In Section 3, we define the key tools that we require for our analysis. Generalization bounds for sequence-to-sequence models are given in Section 4. We compare sequence-to-sequence and local models in Section 5. Section 6 concludes this paper with a study of a setup that is a hybrid of the local and sequence-to-sequence models.

2 Sequence-to-sequence modeling

We begin by providing a formal definition of sequence-to-sequence modeling. The learner receives a multi-dimensional time series $\mathbf{Y} \in \mathcal{Y}^{m \times T}$ which we view as m time series of same length T. We denote by $Y_t(i)$ the value of the i-th time series at time t and write $Y_a^b(i)$ to denote the sequence $(Y_a(i), Y_{a+1}(i), \ldots, Y_b(i))$. Similarly, we let $Y_t(\cdot) = (Y_t(1), \ldots, Y_t(m))$ and $Y_a^b(\cdot) = (Y_a(\cdot), \ldots, Y_b(\cdot))$. In particular, $\mathbf{Y} \equiv Y_1^T(\cdot)$. In addition, the sequence $Y_1^{T-1}(\cdot)$ is of a particular importance in our analysis and we denote it by $\mathbf{Y}'$.

The goal of the learner is to predict $Y_{T+1}(\cdot)$.[2] We further assume that our input $\mathbf{Y}$ is partitioned into a training set of m examples $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$, where each $Z_i = (Y_1^{T-1}(i), Y_T(i)) \in \mathcal{Y}^T$. The learner's objective is to select a hypothesis $h : \mathcal{Y}^T \to \mathcal{Y}$ from a given hypothesis set $\mathcal{H}$ that achieves a small generalization error:

$$\mathcal{L}(h \mid \mathbf{Y}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}} \left[L(h(Y_1^T(i)), Y_{T+1}(i)) \mid \mathbf{Y} \right],$$

where $L : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ is a bounded[3] loss function and $\mathcal{D}$ is the distribution of Y_{T+1} conditioned on $\mathbf{Y}$.

In other words, the learner seeks a hypothesis h that maps sequences of past $Y_1(i), \ldots, Y_T(i)$ values to sequences of future values $Y_{T+1}(i), \ldots, Y_{T+\ell}(i)$, justifying our choice of "sequence-to-sequence" terminology.[4] Incidentally, the machine translation problem studied by Sutskever et al. (2014) under the same name represents a special case of our problem when sequences (sentences) are independent and data is stationary. In fact, LSTM-based approaches used in aforementioned translation problem are also common for time series forecasting (Flunkert et al., 2017; Goel et al., 2017; Yu et al., 2017; Li et al., 2017; Laptev et al., 2017; Zhu and Laptev, 2017). However, feed-forward NNs have also been successfully applied in this framework (Romeu et al., 2013) and in practice, our definition allows for any set of functions $\mathcal{H}$ that map input sequences to output sequences. For instance, we can train a feed-forward NN to map $Y_1^{T-1}(i)$ to $Y_T(i)$ and at inference time use $Y_2^T(i)$ as input to obtain a forecast for $Y_T(i)$.[5]

We contrast sequence-to-sequence modeling to *local* modeling, which consists of splitting each time series $Y(i)$ into a training set $\mathbf{Z}_i = \{Z_{i,1}, \ldots, Z_{i,T}\}$, where $Z_{i,t} = (Y_{t-p}^{t-1}(i), Y_t(i))$ for some $p \in \mathbb{N}$, then learning a separate hypothesis h_i for each $\mathbf{Z}_i$. Each h_i models relations between observations that are close in time, which is why we refer to this framework as local modeling. As in sequence-to-sequence modeling, the goal of a local learner is to achieve a small generalization error for $h_{\text{loc}} = (h_1, \ldots, h_m)$, given by:

$$\mathcal{L}(h_{\text{loc}} \mid \mathbf{Y}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}} \left[L(h_i(Y_{T-p}^T(i)), Y_{T+1}(i)) \mid \mathbf{Y} \right].$$

In order to model correlations between different time series, it is also common to split $\mathbf{Y}$ into one single set of multivariate examples $\mathbf{Z} = \{Z_1, \ldots, Z_T\}$, where $Z_i = (Y_{t-p}^{t-1}(\cdot), Y_t(\cdot))$, and to learn a single hypothesis h that maps $\mathcal{Y}^{m \times p} \to \mathcal{Y}^m$. As mentioned earlier, we consider this approach a variant of local modeling, since h in this case again models relations between observations that are close in time.

Finally, *hybrid* or *local sequence-to-sequence* models, which interpolate between local and sequence-to-sequence approaches, have also been considered in the

[2] We are often interested in long term forecasting, i.e. predicting $Y_{T+1}^{T+\ell}(\cdot)$ for $\ell \geq 1$. For simplicity, we only consider the case of $\ell = 1$. However, all our results extend immediately to $\ell \geq 1$.

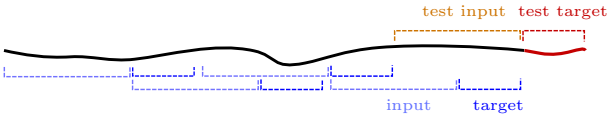
[3] Most of the results in this paper can be straightforwardly extended to unbounded case assuming Y is sub-Gaussian.

[4] In practice, each Z_i may start at a different, arbitrary time t_i, and may furthermore include some additional features X_i, i.e. $Z_i = (Y_{t_i}^{T-1}(i), X_i, Y_T(i))$. Our results can be extended to this case as well using an appropriate choice of the hypothesis set.

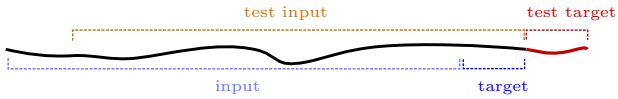
[5] As another example, a runner-up in the Kaggle forecasting competition (`https://www.kaggle.com/c/web-traffic-time-series-forecasting`) used a combination of boosted decision trees and feed-forward networks, and as such employs the sequence-to-sequence approach.

Table 1: Summary of local, sequence-to-sequence, and hybrid models.

Learning Model	Training set	Hypothesis	Example
UniVar. local	$\mathbf{Z}_i = \{(Y_{t-1-p}^{t-1}(i), Y_t(i)) : p \le t \le T\}$	$h_i : \mathcal{Y}^p \to \mathcal{Y}$	ARIMA
MultiVar. local	$\mathbf{Z} = \{(Y_{t-1-p}^{t-1}(\cdot), Y_t(\cdot)) : p \le t \le T\}$	$h : \mathcal{Y}^{m \times p} \to \mathcal{Y}^m$	VARMA
Seq-to-seq	$\mathbf{Z} = \{(Y_1^{T-1}(i), Y_T(i)) : 1 \le i \le m\}$	$h : \mathcal{Y}^T \to \mathcal{Y}$	Neural nets
Hybrid	$\mathbf{Z} = \{(Y_{t-p}^{t-1}(i), Y_t(i)) : 1 \le i \le m, p \le t \le T\}$	$h : \mathcal{Y}^p \to \mathcal{Y}$	Neural nets



(a) The local model trains each $h_{\text{loc},i}$ on time series $Y(i)$ split into multiple (partly overlapping) examples.



(b) The sequence-to-sequence trains h_{s2s} on m time series split into (input, target) pairs.

Figure 1: Local and sequence-to-sequence splits of a one dimensional time series into training and test examples.

literature (Zhu and Laptev, 2017). In this setting, each local example is split across the temporal dimension into smaller examples of length p, which are then used to train a single sequence-to-sequence model h. We discuss bounds for this specific case in Section 6.

Our work focuses on the statistical properties of the sequence-to-sequence model. We provide the first generalization bounds for sequence-to-sequence and hybrid models, and compare these to similar bounds for local models, allowing us to identify regimes in which one methodology is more likely to succeed.

Aside from learning guarantees, there are other important considerations that may lead a practitioner to choose one approach over others. For instance, the local approach is trivially parallelizable; on the other hand, when additional features X_i are available, sequence-to-sequence modeling provides an elegant solution to the *cold start problem* in which at test time we are required to make predictions on time series for which no historical data is available.

3 Correlations and non-stationarity

In the standard supervised learning scenario, it is common to assume that training and test data are drawn *i.i.d.* from some unknown distribution. However, this assumption does not hold for time series, where observations at different times as well as across different series may be correlated. Furthermore, the data-generating distribution may also evolve over time.

These phenomena present a significant challenge to providing guarantees in time series forecasting. To quantify non-stationarity and correlations, we introduce the notions of mixing coefficients and discrepancy, which are defined below.

The final ingredient we need to analyze sequence-to-sequence learning is the notion of Rademacher complexity $\mathfrak{R}_m(\mathcal{F})$ of a family of functions $\mathcal{F}$ on a sample of size m, which has been previously used to characterize learning in the i.i.d. setting (Koltchinskii and Panchenko, 2002; Mohri et al., 2012). In Appendix A, we include a brief discussion of its properties.

3.1 Expected mixing coefficients

To measure the strength of dependency between time series, we extend the notion of β-mixing coefficients (Doukhan, 1994) to *expected β-mixing coefficients*, which are a more appropriate measure of correlation in sequence-to-sequence modeling.

Definition 1 (Expected β_{s2s} coefficients)**.** Let $i, j \in [m] \triangleq \{1, \ldots, m\}$. We define

$$\beta_{\text{s2s}}(i,j) = \mathbb{E}_{\mathbf{Y}'}\Big[\|P(Y_T(i)|\mathbf{Y}')P(Y_T(j)|\mathbf{Y}') - P(Y_T(i), Y_T(j)|\mathbf{Y}')\|_{TV}\Big],$$

where TV denotes the total variations norm. For a subset $I \subseteq [m]$, we define

$$\beta_{\text{s2s}}(I) = \sup_{i,j \in C} \beta_{\text{s2s}}(i,j).$$

The coefficient $\beta_{\text{s2s}}(i,j)$ captures how close $Y_{T+1}(i)$ and $Y_{T+1}(j)$ are to being independent, given $\mathbf{Y}'$ (and averaged over all realizations of $\mathbf{Y}'$). We further study these coefficients in Section 4, where we derive explicit upper bounds on expected β_{s2s}-mixing coefficients for various standard classes of stochastic processes, including spatio-temporal and hierarchical time series.

We also define the following related notion of $\bar{\beta}$-coefficients.

Definition 2 (Unconditional $\bar{\beta}$ coefficients)**.** Let $i, j \in$

$[m] \triangleq \{1, \dots, m\}$. We define

$$\bar{\beta}(i,j) = \|\Pr(Y_1^T(i), Y_1^T(j)) - \Pr(Y_1^T(i)) \Pr(Y_1^T(j))\|_{TV}$$
$$\bar{\beta}'(i,j) = \|\Pr(Y_1^{T-1}(i), Y_1^{T-1}(j)) - \Pr(Y_1^{T-1}(i)) \Pr(Y_1^{T-1}(j))\|_{TV}$$

and as before, for a subset I of $[m]$, write $\bar{\beta}(I) = \sup_{i,j \in I} \bar{\beta}(i,j)$ (and similarly for $\bar{\beta}'$).

Note that β_{s2s} coefficients measure the strength of dependence between time series conditioned on the history observed so far, while $\bar{\beta}$ coefficients measure the (unconditional) strength of dependence between time series. The following result relates these two notions.

Lemma 1. *For $\bar{\beta}$ (and $\bar{\beta}'$ similarly), we have the following upper bound:*

$$\bar{\beta}(i,j) \leq \beta_{s2s}(i,j) + \mathbb{E}_{\mathbf{Y}'}\Big[Cov\Big(\Pr(Y_T(i) \mid \mathbf{Y}'), \Pr(Y_T(j) \mid \mathbf{Y}')\Big)\Big]$$

The proof of this result (as well as all other proofs in this paper) is deferred to the supplementary material.

Finally, we require the notion of *tangent collections*, within which time series are independent.

Definition 3 (Tangent collection)**.** Given a collection of time series $C = \{Y(1), \dots, Y(c)\}$, we define the tangent collection $\widetilde{C}$ as $\{\widetilde{Y}(1), \dots, \widetilde{Y}(c)\}$ such that $\widetilde{Y}(i)$ is drawn according to the marginal $\Pr(Y(i))$ and such that $\widetilde{Y}(i)$ and $\widetilde{Y}(i')$ are independent for $i \neq i'$.

The notion of tangent collections, combined with mixing coefficients, allows us to reduce the analysis of correlated time series in C to the analysis of independent time series in $\widetilde{C}$ (see Proposition 6 in Appendix for details).

3.2 Discrepancy

Various notions of discrepancy have been previously used to measure the non-stationarity of the underlying stochastic processes with respect to the hypothesis set $\mathcal{H}$ and loss function L in the analysis of local models (Kuznetsov and Mohri, 2015; Zimin and Lampert, 2017). In this work, we introduce a notion of discrepancy specifically tailored to sequence-to-sequence modeling scenario, taking into account both the hypothesis set and the loss function.

Definition 4 (Discrepancy)**.** Let $\mathcal{D}$ be the distribution of Y_{T+1} conditioned on $\mathbf{Y}$ and let $\mathcal{D}'$ be the distribution of Y_T conditioned on $\mathbf{Y}'$. We define the discrepancy Δ as $\Delta = \sup_{h \in \mathcal{H}} |\mathcal{L}(h \mid \mathbf{Y}) - \mathcal{L}(h \mid \mathbf{Y}')|$ where $\mathcal{L}(h \mid \mathbf{Y}') = \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\mathcal{D}'}\left[L(h(Y_1^{T-1}(i)), Y_T(i)) \mid \mathbf{Y}\right]$.

The discrepancy forms a pseudo-metric on the space of probability distributions and can be completed to a Wasserstein metric (by extending $\mathcal{H}$ to all Lipschitz functions). This also immediately implies that the discrepancy can be further upper bounded by the l_1-distance and by relative entropy between conditional distributions of Y_T and Y_{T+1} (via Pinsker's inequality). However, unlike these other divergences, the discrepancy takes into account both the hypothesis set and the loss function, making it a finer measure of non-stationarity.

However, the most important property of the discrepancy is that it can be upper bounded by the related notion of *symmetric discrepancy*, which can be estimated from data.

Definition 5 (Symmetric discrepancy)**.** We define the symmetric discrepancy Δ_s as

$$\Delta_s = \frac{1}{m} \sup_{h,h' \in \mathcal{H}} \Big| \sum_{i=1}^m L(h(Y_1^T(i)), h'(Y_1^T(i))) - L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i)))\Big|.$$

Proposition 1. *Let $\mathcal{H}$ be a hypothesis space and let L be a bounded loss function which respects the triangle inequality. Let $h \in \mathcal{H}$ be any hypothesis. Then, $\Delta \leq \Delta_s + \mathcal{L}(h \mid \mathbf{Y}) + \mathcal{L}(h \mid \mathbf{Y}')$.*

We do not require test labels to evaluate Δ_s. Since Δ_s only depends on the observed data, Δ_s can be computed directly from samples, making it a useful tool to assess the non-stationarity of the learning problem.

Another useful property of Δ_s is that, for certain classes of stochastic processes, we can provide a direct analysis of this quantity.

Proposition 2. *Let $I_1, \cdots, I_k$ be a partition of $\{1, \dots, m\}$, $C_1, \dots, C_k$ be the corresponding partition of $\mathbf{Y}$ and $C'_1, \dots, C'_k$ be the corresponding partition of $\mathbf{Y}'$. Write $c = \min_j |C_j|$, and define the* expected discrepancy

$$\Delta_e = \sup_{h,h' \in \mathcal{H}} \Big[\mathbb{E}_Y[L(h(Y_1^T), h'(Y_1^T))] - \mathbb{E}_Y[L(h(Y_1^{T-1}), h'(Y_1^{T-1}))]\Big].$$

Then, writing $\mathfrak{R}$ the Rademacher complexity (see Appendix A) we have with probability $1 - \delta$,

$$\Delta_s \leq \Delta_e + \max\Big(\max_j \mathfrak{R}_{|C_j|}(\widetilde{C}'_j), \max_j \mathfrak{R}_{|C_j|}(\widetilde{C}_j)\Big) + \sqrt{\frac{1}{2c} \log \frac{2k}{\delta - \sum_j (|I_j| - 1)[\bar{\beta}(I_j) + \bar{\beta}'(I_j)]}}.$$

The expected discrepancy Δ_e can be computed analytically for many classes of stochastic processes. For

example, for stationary processes, we can show that it is negligible. Similarly, for covariance-stationary[6] processes with linear hypothesis sets and the squared loss function, the discrepancy is once again negligible. These examples justify our use of the discrepancy as a natural measure of non-stationarity. In particular, the covariance-stationary example highlights that the discrepancy takes into account not only the distribution of the stochastic processes but also $\mathcal{H}$ and L.

Proposition 3. *If $Y(i)$ is stationary for all $1 \le i \le m$, and $\mathcal{H}$ is a hypothesis space such that $h \in \mathcal{H} : \mathcal{Y}^{T-1} \to \mathcal{Y}$ (i.e. the hypotheses only consider the last $T-1$ values of Y), then $\Delta_e = 0$.*

Proposition 4. *If $\mathbf{Y}$ is covariance stationary for all $1 \le i \le m$, L is the squared loss, and $\mathcal{H}$ is a linear hypothesis space $\{x \to w \cdot x \mid \|w\| \in \mathbb{R}^p \le \Lambda\}$, $\Delta_e = 0$.*

Another insightful example is the case when $\mathcal{H} = \{h\}$: then, $\Delta = 0$ even if $\mathbf{Y}$ is non-stationary, which illustrates that learning is trivial for trivial hypothesis sets, even in non-stationary settings.

The final example that we consider in this section is the case of non-stationary periodic time series. Remarkably, we show that the discrepancy is still negligible in this case provided that we observe all periods with equal probability.

Proposition 5. *If the $Y(i)$ are periodic with period p and the observed starting time of each $Y(i)$ is distributed uniformly at random in $[p]$, then $\Delta_e = 0$.*

4 Generalization bounds

We now present our generalization bounds for time series prediction with sequence-to-sequence models. We write $\mathcal{F} = \{L \circ h : \ h \in \mathcal{H}\}$, where $f = L \circ h$ is the loss of hypothesis h given by $f(h, Z_i) = L(h(Y_1^{T-1}(i)), Y_T)$. To obtain bounds on the generalization error $\mathcal{L}(h \mid \mathbf{Y})$, we study the gap between $\mathcal{L}(h \mid \mathbf{Y})$ and the empirical error $\widehat{\mathcal{L}}(h)$ of a hypothesis h, where

$$\widehat{\mathcal{L}}(h) = \frac{1}{m} \sum_{i=1}^{m} f(h, Z_i).$$

That is, we aim to give a high probability bound on the supremum of the empirical process $\Phi(\mathbf{Y}) = \sup_h [\mathcal{L}(h \mid \mathbf{Y}) - \widehat{\mathcal{L}}(h)]$. We take the following high-level approach: we first partition the training set $\mathbf{Z}$ into k collections $C_1, \ldots, C_k$ such that within each collection, correlations between different time series are as weak as possible. We then analyze each collection C_j by comparing the generalization error of sequence-to-sequence learning on C_j to the sequence-to-sequence generalization error on the *tangent* collection $\widetilde{C}_j$.

Theorem 4.1. *Let $C_1, \ldots, C_k$ form a partition of the training input $\mathbf{Z}$ and let I_j denote the set of indices of time series that belong to C_j. Assume that the loss function L is bounded by 1. Then, we have for any $\delta > \sum_j (|I_j|-1)\beta(I_j)$, with probability $1-\delta$,*

$$\Phi(\mathbf{Y}) \le \max_j \left[\widehat{\mathfrak{R}}_{\widetilde{C}_j}(\mathcal{F})\right] + \Delta + \frac{1}{\sqrt{2 \min_j |I_j|}} \sqrt{\log\left(\frac{k}{\delta - \sum_j (|I_j|-1)\beta_{s2s}(I_j)}\right)}.$$

Theorem 4.1 illustrates the trade-offs that are involved in sequence-to-sequence learning for time series forecasting. As $\sum_j (|I_j|-1)\beta(I_j)$ is a function of m, we expect it to decrease as m grows (*i.e.* more time series we have), allowing for smaller δ as m increases.

Assuming that the C_j are of the same size, if $\mathcal{H}$ is a collection of neural networks of bounded depth and width then $\mathfrak{R}_{\widetilde{C}_j}(\mathcal{F}) = \mathcal{O}\left(\sqrt{kT/m}\right)$ (see Appendix A). Therefore,

$$\mathcal{L}(h \mid \mathbf{Y}) \le \widehat{\mathcal{L}}(h) + \Delta + \mathcal{O}\left(\sqrt{\tfrac{kT}{m}}\right)$$

with high probability uniformly over $h \in \mathcal{H}$, provided that $\frac{m}{k} \sum_{j=1}^{k} \beta_{\mathrm{s2s}}(I_j) = o(1)$. This shows that extremely high-dimensional ($m \gg 1$) time series are beneficial for sequence-to-sequence models, whereas series with a long histories $T \gg m$ will generally not benefit from sequence-to-sequence learning. Note also that correlations in data reduce the effective sample size from m to m/k.

Furtermore, Theorem 4.1 indicates that balancing the complexity of the model (e.g. depth and width of a neural net) with the fit it provides to the data is critical for controlling both the discrepancy and Rademacher complexity terms. We further illustrate this bound with several examples below.

4.1 Independent time series

We begin by considering the case where all dimensions of $\mathbf{Y}$ are independent. Although this may seem a restrictive assumption, it arises in a variety of applications: in neuroscience, different dimensions may represent brain scans of different patients; in reinforcement learning, they may correspond to different trajectories of a robotic arm.

Theorem 4.2. *Let $\mathcal{H}$ be a given hypothesis space with associated function family $\mathcal{F}$ corresponding to a loss*

[6] Recall that a process $X_1, X_2, \ldots$ is called stationary if for any l, k, m, the distributions of $(X_k, \ldots, X_{k+l})$ and $(X_{k+m}, \ldots, X_{k+m+l})$ are the same. Covariance stationarity is a weaker condition that requires that $\mathbb{E}[X_k]$ be independent of k and that $\mathbb{E}[X_k X_m] = f(k-m)$ for some f.

function L bounded by 1. Suppose that all dimensions of $\mathbf{Y}$ are independent and let $I_1 = [m]$; then $\beta(I_1) = 0$ and so for any $\delta > 0$, with probability at least $1 - \delta$ and for any $h \in \mathcal{H}$:

$$\mathcal{L}(h|\mathbf{Y}) \leq \widehat{\mathcal{L}}(h) + 2\mathfrak{R}_m(\mathcal{F}) + \Delta + \sqrt{\frac{\log(1/\delta)}{m}}.$$

Theorem 4.2 shows that when time series are independent, learning is not affected by correlations in the samples and can only be obstructed by the non-stationarity of the problem, captured via Δ.

Note that when examples are drawn *i.i.d.*, we have $\Delta = 0$ in Theorem 4.2: we recover the standard standard generalization results for regression problems.

4.2 Correlated time series

We now consider several concrete examples of high-dimensional time series in which different dimensions may be correlated. This setting is common in a variety of applications including stock market indicators, traffic conditions, climate observations at different locations, and energy demand.

Suppose that each $Y(i)$ is generated by the auto-regressive (AR) processes with correlated noise

$$y_{t+1}(i) = \Theta_i(y_0^t(i)) + \varepsilon_{t+1}(i) \tag{4.1}$$

where the $w_i \in \mathbb{R}^p$ are unknown parameters and the noise vectors $\epsilon_t \in \mathbb{R}^m$ are drawn from a Gaussian distribution $\mathcal{N}(0, \Sigma)$ where, crucially, Σ is not diagonal. The following lemma is key to our analysis.

Lemma 2. *Two AR processes $Y(i), Y(j)$ generated by (4.1) such that $\sigma = Cov(Y(i), Y(j)) \leq \sigma_0 < 1$ verify $\beta_{s2s}(i,j) = \max\left(\frac{3}{2(1-\sigma_0^2)}, \frac{1}{1-2\sigma_0}\right)\sigma = \mathcal{O}(\sigma)$.*

Hierarchical time series. As our first example, we consider the case of hierarchical time series that arises in many real-world applications (Wickramasuriya et al., 2015; Taieb et al., 2017). Consider the problem of energy demand forecasting: frequently, one observes a sequence of energy demands at a variety of levels: single household, local neighborhood, city, region and country. This imposes a natural hierarchical structure on these time series.

Formally, we consider the following hierarchical scenario: a binary tree of total depth D, where time series are generated at each of the leaves. At each leaf, $Y(i)$ is given by the AR process (4.1) where we impose $\Sigma_{ij} = (\frac{1}{m})^{d(i,j)}$ given $d(i,j)$ the length of the shortest path from either leaf to the closest common ancestor between i and j. Hence, as $d(i,j)$ increases, $Y(i)$ and $Y(j)$ grow more independent.

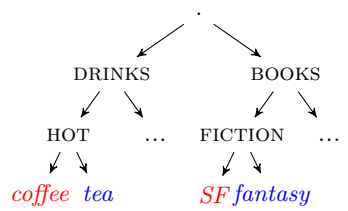


Figure 2: Hierarchical time series. Collections for $d = 1$ are C_1 which contains the (red) left-side leaves and C_2 which contains the (blue) right-side leaves.

For the bound of Theorem 4.1 to be non-trivial, we require a partition $C_1, \ldots, C_k$ of $\mathbf{Z}$ such that within a given C_j the time series are close to being independent. One such construction is the following: fix a depth $d \leq D$ and construct $C_1, \ldots, C_{2^d}$ such that each C_i contains exactly one time series from each sub-tree of depth $D - d$; hence, $|C_i| = 2^{D-d}$. Lemma 2 shows that for each C_i, we have $\beta(C_i) = \mathcal{O}(m^{d-D})$. For example, setting $d = \frac{D}{2} = \frac{\log m}{2}$, it follows that for any $\delta > 0$, with probability $1 - \delta$,

$$\mathcal{L}(h|\mathbf{Y}) \leq \widehat{\mathcal{L}}(h) + \max_j \left[\mathfrak{R}_{\widetilde{C}_j}(\mathcal{F})\right] + \Delta + \frac{1}{\sqrt{2}\sqrt[4]{m}}\sqrt{\log\left(\frac{\sqrt{m}}{\delta - \frac{m}{\mathcal{O}(\sqrt{m^{\log m}})}}\right)}.$$

Furthermore, suppose the model is a linear AR process given by $y_{t+1}(i) = w_i \cdot (y_{t-p}^t(i)) + \varepsilon_{t+1}(i)$. In that case, the underlying stochastic process is weakly stationary and by Proposition 3 our bound reduces to: $\mathcal{L}(h|\mathbf{Y}) \leq \widehat{\mathcal{L}}(h) + \max_j \left[\mathfrak{R}_{\widetilde{C}_j}(\mathcal{F})\right] + \mathcal{O}\left(\frac{\sqrt{\log m}}{m^{1/4}}\right)$. By Proposition 5, similar results holds when Θ_i is periodic.

Spatio-temporal processes. Another common task is spatio-temporal forecasting, in which historical observation are available at different locations. For instance, these observations may represent temperature at different locations, as in the case of climate modeling (McQuade and Monteleoni, 2012; Ghafarianzadeh and Monteleoni, 2013), or car traffic at different locations (Li et al., 2017).

It is natural to expect correlations between time series to decay as the geographical distance between them increases. As a simplified example, consider that the sphere $\mathbb{S}^3$ is subdivided according to a geodesic grid and a time series is drawn from the center of each patch according to (4.1), also with $\Sigma_{ij} = m^{-d(i,j)}$ but this time with $d(i,j)$ equal to the (geodesic) distance between the center of two cell centers. We choose subsets C_i with the goal of minimizing the strength of dependencies between time series within each subsets. Assuming we divide the sphere into $\sqrt{m}$ collections size

approximately $c = \sqrt{m}$ such that the minimal distance between two points in a set is d_0, we obtain

$$\mathcal{L}(h \mid \mathbf{Y}) \leq \widehat{\mathcal{L}}(h) + \max_j \left[\mathfrak{R}_{\widetilde{C}_j}(\mathcal{F})\right] + \Delta + \frac{1}{\sqrt{2}\sqrt[4]{m}} \sqrt{\log\left(\frac{\sqrt{m}}{\delta - \mathcal{O}(m^{1-d_0})}\right)}.$$

As in the case of hierarchical time series, Proposition 3 or Proposition 5 can be used to remove the dependence on Δ for certain families of stochastic processes.

5 Comparison to local models

This section provides comparison of learning guarantees for sequence-to-sequence models with those of local models. In particular, we will compare our bounds on the generalization gap $\Phi(\mathbf{Y})$ for sequence-to-sequence models and local models, where the gap is given by

$$\Phi_{\text{loc}}(\mathbf{Y}) = \sup_{(h_1,\ldots,h_m)\in H^m} \left[\mathcal{L}(h_{\text{loc}} \mid \mathbf{Y}) - \widehat{\mathcal{L}}(h_{\text{loc}})\right] \tag{5.1}$$

where $\widehat{\mathcal{L}}(h_{\text{loc}})$ is the average empirical error of h_i on the sample $\mathbf{Z}_i$, defined as $\widehat{\mathcal{L}}(h_{\text{loc}}) = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T} f(h_i, Z_{t,i})$ where $f(h_i, Z_{t,i}) = L(h_i(Y_{t-p}^{t-1}(i)), Y_t(i))$.

To give a high probability bound for this setting, we take advantage of existing results for the single local model h_i (Kuznetsov and Mohri, 2015). These results are given in terms of a slightly different notion of discrepancy Δ which is defined by

$$\Delta(\mathbf{Z}_i) = \sup_{h\in\mathcal{H}} \left[\mathbb{E}\left[L(h(Y_{t-p+1}^{T}), Y_{T+1}) \mid Y_1^T\right] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[L(h(Y_{t-p}^{t-1}), Y_t) \mid Y_1^{t-1}\right]\right].$$

Another required ingredient to state these results is the expected sequential covering number $\mathbb{E}_{v\sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{F}, v)]$ (Kuznetsov and Mohri, 2015). For many hypothesis sets, the log of the sequential covering number admits upper bounds similar to those presented earlier for the Rademacher complexity. We provide some examples below and refer the interested reader to (Rakhlin et al., 2015) for a details.

Theorem 5.1. *For $\delta > 0$ and $\alpha > 0$, with probability at least $1 - \delta$, for any $(h_1, \ldots, h_m)$, and any $\alpha > 0$,*

$$\Phi_{loc}(\mathbf{Y}) \leq \frac{1}{m}\sum_{i=1}^{m} \Delta(\mathbf{Z}_i) + 2\alpha + \sqrt{\frac{2}{T}\log\frac{m \max_i \mathbb{E}_{v\sim T(\mathbf{Z}_i)}[\mathcal{N}_1(\alpha, \mathcal{F}, v)]}{\delta}}.$$

Choosing $\alpha = 1/\sqrt{T}$, we can show that, for standard local models such as the linear hypothesis space $\{x \to w^\top x, w \in \mathbb{R}^p, \|w\|_2 \leq \Lambda\}$, we have

$$\sqrt{\frac{1}{T}\log\frac{2m\,\mathbb{E}_{v\sim T(Z)}[\mathcal{N}_1(\alpha, \mathcal{F}, v)]}{\delta}} = \mathcal{O}\left(\sqrt{\frac{\log m}{T}}\right).$$

In this case, it follows that $\Phi_{\text{loc}}(\mathbf{Y}) \leq \frac{1}{m}\sum_{i=1}^{m}\Delta(\mathbf{Z}_i) + \mathcal{O}\left(\sqrt{\frac{\log m}{T}}\right)$. where the last term in this bound should be compared with corresponding (non-discrepancy) terms in the bound of Theorem. 4.1, which, as discussed above, scales as $\mathcal{O}(\sqrt{T/m})$ for a variety of different hypothesis sets.

Hence, when we have access to relatively few time series compared to their length ($m \ll T$), learning to predict each time series as its own independent problem will with high probability lead to a better generalization bound. On the other hand, in extremely high-dimensional settings when we have significantly more time series than time steps ($m \gg T$), sequence-to-sequence learning will (with high probability) provide superior performance. We also expect the performance of sequence-to-sequence models to deteriorate as the correlation between time series increases.

A direct comparison of bounds in Theorem 4.1 and Theorem 5.1 is complicated by the fact that discrepancies that appear in these results are different. In fact, it is possible to design examples where $\frac{1}{m}\sum_{i=1}^{m}\Delta(\mathbf{Z}_i)$ is constant and Δ is negligible, and vice-versa.

Consider a tent function g_b such that $g_b(s) = 2bs/T$ for $s \in [0, T/2]$ and $g_b(s) = -2bs/T + 2b$ for $s \in [T/2, T]$. Let f_b be its periodic extension to the real line, and define $\mathcal{S} = \{f_b : b \in [0, 1]\}$. Suppose that we sample uniformly $b \in [0, 1]$ and $s \in \{0, T/2\}$ m times, and observe time series $f_{b_i}(s_i), \ldots, f_{b_i}(s_i + T)$. Then, as we have shown in Proposition 5, Δ is negligible for sequence-to-sequence models. However, unless the model class is trivial, it can be shown that $\Delta(\mathbf{Z}_i)$ is bounded away from zero for all i.

Conversely, suppose we sample uniformly $b \in [0, 1]$ m times and observe time series $f_{b_i}(0), \ldots, f_{b_i}(T/2 + 1)$. Consider a set of local models that learn an offset from the previous point $\{h : x \mapsto x + c, c \in [0, 1]\}$. It can be shown that in this case $\Delta(\mathbf{Z}_i) = 0$, whereas Δ is bounded away from zero for any non-trivial class of sequence-to-sequence models.

From a practical perspective, we can simply use Δ_s and empirical estimates of $\Delta(\mathbf{Z}_i)$ to decide whether to choose sequence-to-sequence or local models.

We conclude this section with an observation that similar results to Theorem 4.1 can be proved for multivariate local models with the only difference that the sample complexity of the problem scales as $O(\sqrt{m/T})$,

and hence these models are even more prone to the curse of dimensionality.

6 Hybrid models

In this section, we discuss models that interpolate between local and sequence-to-sequence models. This hybrid approach trains a single model h on the union of *local* training sets $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ used to train m models in the local approach. The bounds that we state here require the following extension of the discrepancy to Δ_t, defined as

$$\Delta_t = \frac{1}{m} \sup_{h \in \mathcal{H}} \Big| \sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}}[L(h(Y_{t-p-1}^{t-1}(i)), Y_t(i)) | Y_1^{t-1}] - \mathbb{E}_{\mathcal{D}'}[L(h(Y_{T-p}^{T}(i)), Y_{T+1}(i)) | \mathbf{Y}] \Big|$$

Many of the properties that were discussed for the discrepancy Δ carry over to Δ_t as well. The empirical error in this case is the same as for the local models:

$$\widehat{\mathcal{L}}(h) = \frac{1}{mT} \sum_{i=1}^{m} \sum_{t=1}^{T} f(h, Z_{t,i}).$$

Observe that one straightforward way to obtain a bound for hybrid models is to apply Theorem 5.1 with $(h, \ldots, h) \in \mathcal{H}^m$. Alternatively, we can apply Theorem 4.1 at every time point $t = 1, \ldots, T$.

Combining these results via union bound leads to the following learning guarantee for hybrid models.

Theorem 6.1. *Let $C_1, \ldots, C_k$ form a partition of the training input $\mathbf{Z}$ and let I_j denote the set of indices of time series that belong to C_j. Assume that the loss function L is bounded by 1. Then, for any $\delta > 0$, with probability $1 - \delta$, for any $h \in \mathcal{H}$ and any $\alpha > 0$*

$$\mathcal{L}(h \mid \mathbf{Y}) \leq \widehat{\mathcal{L}}(h) + \min(B_1, B_2),$$

where

$$B_1 = \frac{1}{T} \sum_{t=1}^{T} \Delta_t + \max_j \widehat{\mathfrak{R}}_{\widetilde{C}_j}(\mathcal{F}) + \frac{1}{\sqrt{2 \min_j |I_j|}} \sqrt{\log \left(\frac{2Tk}{\delta - 2 \sum_j (|I_j| - 1) \beta_{s \otimes s}(I_j)} \right)}$$

$$B_2 = \frac{1}{m} \sum_{i=1}^{m} \Delta(\mathbf{Z}_i) + 2\alpha + \sqrt{\frac{2}{T} \log \frac{2m \max_i \mathbb{E}_{v \sim T(\mathbf{Z}_i)}[\mathcal{N}_1(\alpha, \mathcal{F}, v)]}{\delta}}.$$

Using the same arguments for the complexity terms as in the case of sequence-to-sequence and local models, this result shows that hybrid models are successful with high probability when $m \gg T$ or correlations between time series are strong, as well as when $T \gg m$.

Potential costs for this model are hidden in the new discrepancy term $\frac{1}{T} \sum_{t=1}^{T} \Delta_t$. This term leads to different bounds depending on the particular non-stationarity in the given problem. As before this trade-off can be accessed empirically using the data-dependent version of discrepancy.

Note that the above bound does not imply that hybrid models are strictly better than local models: using m different hypotheses $h_1, \ldots, h_m$ can help us achieve a better trade-off between $\widehat{\mathcal{L}}(h)$ and B_2, and vice versa.

7 Conclusion

We formally introduce sequence-to-sequence learning for time series, a framework in which a model learns to map past sequences of length T to their next values. We provide the first generalization bounds for sequence-to-sequence modeling. Our results are stated in terms of new notions of discrepancy and expected mixing coefficients. We study these new notions for several different families of stochastic processes including stationary, weakly stationary, periodic, hierarchical and spatio-temporal time series.

Furthermore, we show that our discrepancy can be computed from data, making it a useful tool for practitioners to empirically assess the non-stationarity of their problem. In particular, the discrepancy can be used to determine whether the sequence-to-sequence methodology is likely to succeed based on the inherent non-stationarity of the problem.

Furthermore, compared to the local framework for time series forecasting, in which independent models for each one-dimensional time series are learned, our analysis shows that the sample complexity of sequence-to-sequence models scales as $\mathcal{O}(\sqrt{T/m})$, providing superior guarantees when the number m of time series is significantly greater than the length T of each series, provided that different series are weakly correlated.

Conversely, we show that the sample complexity of local models scales as $\mathcal{O}(\sqrt{\log(m)/T})$, and should be preferred when $m \ll T$ or when time series are strongly correlated. We also study hybrid models for which learning guarantees are favorable both when $m \gg T$ and $T \gg m$, but which have a more complex trade-off in terms of discrepancy.

As a final note, the analysis we have carried through is easily extended to show similar results for the sequence-to-sequence scenario when the test data includes *new* series not observed during training, as is often the case in a variety of applications.

References

Marta Banbura, Domenico Giannone, and Lucrezia Reichlin. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 2015.

Filippo Maria Bianchi, Enrico Maiorino, Michael C. Kampffmeyer, Antonello Rizzi, and Robert Jenssen. *Recurrent Neural Networks for Short-Term Load Forecasting - An Overview and Comparative Analysis*. Springer Briefs in Computer Science. Springer, 2017.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3): 307 – 327, 1986.

George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer-Verlag New York, Inc., 1986.

P. Doukhan. *Mixing: Properties and Examples*. Lecture notes in statistics. Springer, 1994.

Robert Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *Arxiv:1704.04110*, 2017.

Mahsa Ghafarianzadeh and Claire Monteleoni. Climate prediction via matrix completion. In *Late-Breaking Developments in the Field of Artificial Intelligence*, volume WS-13-17 of *AAAI Workshops*. AAAI, 2013.

Hardik Goel, Igor Melnyk, and Arindam Banerjee. R2N2: Residual recurrent neural networks for multivariate time series forecasting. *arXiv:1709.03159*, 2017.

James Douglas Hamilton. *Time series analysis*. Princeton Univ. Press, 1994.

Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150, 2015a.

Fang Han, Sheng Xu, and Han Liu. Rate optimal estimation of high dimensional time series. Technical report, Technical Report, Johns Hopkins University, 2015b.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. ISSN 0899-7667.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 2002.

Vitaly Kuznetsov and Mehryar Mohri. *Generalization Bounds for Time Series Prediction with Non-stationary Processes*, pages 260–274. Springer International Publishing, 2014.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems 28*, pages 541–549. Curran Associates, Inc., 2015.

Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and online learning. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1190–1213. PMLR, 2016.

Vitaly Kuznetsov and Mehryar Mohri. Discriminative state space models. In *NIPS, Long Beach, CA, USA*, 2017.

Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. Time-series extreme event forecasting with neural networks at uber. In *ICML Workshop*, 2017.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv:1707.01926*, 2017.

Helmut Lütkepohl. Chapter 6 Forecasting with VARMA models. In *Handbook of Economic Forecasting*, volume 1, pages 287 – 325. Elsevier, 2006.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated, 2007.

Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16:865–873, 2015.

Scott McQuade and Claire Monteleoni. Global climate model tracking using geospatial neighborhoods. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.* AAAI Press, 2012.

Ron Meir and Lisa Hellerstein. Nonparametric time series prediction through adaptive model selection. In *Machine Learning*, pages 5–34, 2000.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems 21*, pages 1097–1104. Curran Associates, Inc., 2009.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *J. Mach. Learn. Res.*, 11:789–814, 2010. ISSN 1532-4435.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097, 2011.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401. PMLR, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

Pablo Romeu, Francisco Zamora-Martínez, Paloma Botella-Rocamora, and Juan Pardo. *Time-Series Forecasting of Indoor Temperature Using Pre-trained Deep Neural Networks*, pages 451–458. Springer Berlin Heidelberg, 2013.

Song Song and Peter J. Bickel. Large vector auto regressions. *arXiv:1106.3915*, 2011.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

Souhaib Ben Taieb, James W. Taylor, and Rob J. Hyndman. Coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3348–3357. PMLR, 2017.

D. Malioutov W. Sun. Time series forecasting with shared seasonality patterns using non-negative matrix factorization. In *The 29th Annual Conference on Neural Information Processing Systems (NIPS)*. Time Series Workshop, 2015.

Shanika L Wickramasuriya, George Athanasopoulos, and Rob J Hyndman. Forecasting hierarchical and grouped time series through trace minimization. Technical Report 15/15, Monash University, Department of Econometrics and Business Statistics, 2015.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.*, 22(1): 94–116, 1994.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in Neural Information Processing Systems 29*, pages 847–855. Curran Associates, Inc., 2016.

Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. Long-term forecasting using tensor-train RNNs. *Arxiv:1711.00073*, 2017.

Lingxue Zhu and Nikolay Laptev. Deep and confident prediction for time series at Uber. *arXiv:1709.01907*, 2017.

Alexander Zimin and Christoph H. Lampert. Learning theory for conditional risk minimization. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, pages 213–222, 2017.