Discriminative State-Space Models

Vitaly Kuznetsov
Google Research
New York, NY 10011, USA
`vitaly@cims.nyu.edu`

Mehryar Mohri
Courant Institute and Google Research
New York, NY 10011, USA
`mohri@cims.nyu.edu`

Abstract

We introduce and analyze Discriminative State-Space Models for forecasting non-stationary time series. We provide data-dependent generalization guarantees for learning these models based on the recently introduced notion of discrepancy. We provide an in-depth analysis of the complexity of such models. We also study the generalization guarantees for several structural risk minimization approaches to this problem and provide an efficient implementation for one of them which is based on a convex objective.

1 Introduction

Time series data is ubiquitous in many domains including such diverse areas as finance, economics, climate science, healthcare, transportation and online advertisement. The field of time series analysis consists of many different problems, ranging from analysis to classification, anomaly detection, and forecasting. In this work, we focus on the problem of forecasting, which is probably one of the most challenging and important problems in the field.

Traditionally, time series analysis and time series prediction, in particular, have been approached from the perspective of generative modeling: particular generative parametric model is postulated that is assumed to generate the observations and these observations are then used to estimate unknown parameters of the model. Autoregressive models are among the most commonly used types of generative models for time series [Engle, 1982, Bollerslev, 1986, Brockwell and Davis, 1986, Box and Jenkins, 1990, Hamilton, 1994]. These models typically assume that the stochastic process that generates the data is stationary up to some known transformation, such as differencing or composition with natural logarithms.

In many modern real world applications, the stationarity assumption does not hold, which has led to the development of more flexible generative models that can account for non-stationarity in the underlying stochastic process. *State-Space Models* [Durbin and Koopman, 2012, Commandeur and Koopman, 2007, Kalman, 1960] provide a flexible framework that captures many of such generative models as special cases, including autoregressive models, hidden Markov models, Gaussian linear dynamical systems and many other models. This framework typically assumes that the time series $\mathbf{Y}$ is a noisy observation of some dynamical system $\mathbf{S}$ that is hidden from the practitioner:

$$Y_t = h(S_t) + \epsilon_t, \quad S_t = g(S_{t-1}) + \eta_t \quad \text{for all } t. \tag{1}$$

In (1), h, g are some unknown functions estimated from data, $\{\epsilon_t\}$, $\{\eta_t\}$ are sequences of random variables and $\{S_t\}$ is an unobserved sequence of states of a hidden dynamical system.[1] While this class of models provides a powerful and flexible framework for time series analysis, the theoretical learning properties of these models is not sufficiently well understood. The statistical guarantees available in the literature rely on strong assumptions about the noise terms (e.g. $\{\epsilon_t\}$ and $\{\eta_t\}$ are Gaussian white noise). Furthermore, these results are typically asymptotic and require the model

---

[1]A more general formulation is given in terms of distribution of Y_t: $p_h(Y_t|S_t)p_g(S_t|S_{t-1})$.

to be correctly specified. This last requirement places a significant burden on a practitioner since the choice of the hidden state-space is often a challenging problem and typically requires extensive domain knowledge.

In this work, we introduce and study *Discriminative State-Space Models (DSSMs)*. We provide the precise mathematical definition of this class of models in Section 2. Roughly speaking, a DSSM follows the same general structure as in (1) and consists of a state predictor g and an observation predictor h. However, no assumption is made about the form of the stochastic process used to generate observations. This family of models includes existing generative models and other state-based discriminative models (e.g. RNNs) as special cases, but also consists of some novel algorithmic solutions explored in this paper.

The material we present is organized as follows. In Section 3, we generalize the notion of discrepancy, recently introduced by Kuznetsov and Mohri [2015] to derive learning guarantees for DSSMs. We show that our results can be viewed as a generalization of those of these authors. Our notion of discrepancy is finer, taking into account the structure of state-space representations, and leads to tighter learning guarantees. Additionally, our results provide the first high-probability generalization guarantees for state-space models with possibly incorrectly specified models. Structural Risk Minimization (SRM) for DSSMs is analyzed in Section 4. As mentioned above, the choice of the state-space representation is a challenging problem since it requires carefully balancing the accuracy of the model on the training sample with the complexity of DSSM to avoid overfitting. We show that it is possible to adaptively learn a state-space representation in a principled manner using the SRM technique. This requires analyzing the complexity of several families of DSSMs of interest in Appendix B. In Section 5, we use our theory to design an efficient implementation of our SRM technique. Remarkably, the resulting optimization problem turns out to be convex. This should be contrasted with traditional SSMs that are often derived via Maximum Likelihood Estimation (MLE) with a non-convex objective. We conclude with some promising preliminary experimental results in Appendix D.

2 Preliminaries

In this section, we introduce the general scenario of time series prediction as well as the broad family of DSSMs considered in this paper.

We study the problem of time series forecasting in which the learner observes a realization $(X_1, Y_1), \ldots, (X_T, Y_T)$ of some stochastic process, with $(X_t, Y_t) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We assume that the learner has access to a family of observation predictors $\mathcal{H} = \{h \colon \mathcal{X} \times \mathcal{S} \to \mathcal{Y}\}$ and state predictors $\mathcal{G} = \{g \colon \mathcal{X} \times \mathcal{S} \to \mathcal{S}\}$, where $\mathcal{S}$ is some pre-defined space. We refer to any pair $f = (h, g) \in \mathcal{H} \times \mathcal{G} = \mathcal{F}$ as a DSSM, which is used to make predictions as follows:

$$y_t = h(X_t, s_t), \quad s_t = g(X_t, s_{t-1}) \quad \text{for all } t. \tag{2}$$

Observe that this formulation includes the hypothesis sets used in (1) as special cases. In our setting, h and g both accept an additional argument $x \in \mathcal{X}$. In practice, if $X_t = (Y_{t-1}, \ldots, Y_{t-p}) \in \mathcal{X} = \mathcal{Y}^p$ for some p, then X_t represents some recent history of the stochastic process that is used to make a prediction of Y_t. More generally, $\mathcal{X}$ may also contain some additional side information. Elements of the output space $\mathcal{Y}$ may further be multi-dimensional, which covers both multi-variate time series forecasting and multi-step forecasting.

The performance of the learner is measured using a bounded loss function $L \colon \mathcal{H} \times \mathcal{S} \times \mathcal{Z} \to [0, M]$, for some upper bound $M \geq 0$. A commonly used loss function is the squared loss: $L(h, s, z) = (h(x, s) - y)^2$.

The objective of the learner is to use the observed realization of the stochastic process up to time T to determine a DSSM $f = (h, g) \in \mathcal{F}$ that has the smallest expected loss at time $T + 1$, conditioned on the given realization of the stochastic process:[2]

$$\mathcal{L}_{T+1}(f | \mathbf{Z}_1^T) = \mathbb{E}[L(h, s_{T+1}, Z_{T+1}) | \mathbf{Z}_1^T], \tag{3}$$

---

[2]An alternative performance metric commonly considered in the time series literature is the *averaged* generalization error $\mathcal{L}_{T+1}(f) = \mathbb{E}[L(f, s_{T+1}, Z_{T+1})]$. The path-dependent generalization error that we consider in this work is a finer measure of performance since it only takes into consideration the realized history of the stochastic process, as opposed to an average trajectory.

where s_t for all t is specified by g via the recursive computation in (2). We will use the notation $\mathbf{a}_s^r$ to denote $(a_s, a_{s+1}, \ldots a_r)$.

In the rest of this section, we will introduce the tools needed for the analysis of this problem. The key technical tool that we require is the notion of *state-space discrepancy*:

$$\text{disc}(\mathbf{s}) = \sup_{h \in \mathcal{H}} \left(\mathbb{E}[L(h, s_{T+1}, Z_{T+1})|\mathbf{Z}_1^T] - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L(h, s_t, Z_t)|\mathbf{Z}_1^{t-1}] \right), \tag{4}$$

where, for simplicity, we used the shorthand $\mathbf{s} = \mathbf{s}_1^{T+1}$. This definition is a strict generalization of the $\mathbf{q}$-weighted discrepancy of Kuznetsov and Mohri [2015]. In particular, redefining $L(h, s, z) = s\widetilde{L}(h, z)$ and setting $s_t = Tq_t$ for $1 \leq t \leq T$ and $s_{T+1} = 1$ recovers the definition of $\mathbf{q}$-weighted discrepancy. The discrepancy disc defines an integral probability pseudo-metric on the space of probability distributions that serves as a measure of the non-stationarity of the stochastic process $\mathbf{Z}$ with respect to both the loss function L and the hypothesis set $\mathcal{H}$, conditioned on the given state sequence $\mathbf{s}$. For example, if the process $\mathbf{Z}$ is i.i.d., then we simply have $\text{disc}(\mathbf{s}) = 0$ provided that $\mathbf{s}$ is a constant sequence. See [Cortes et al., 2017, Kuznetsov and Mohri, 2014, 2017, 2016, Zimin and Lampert, 2017] for further examples and bounds on discrepancy in terms of other divergences. However, the most important property of the discrepancy $\text{disc}(\mathbf{s})$ is that, as shown in Appendix C, under some additional mild assumptions, it can be estimated from data.

The learning guarantees that we present are given in terms of data-dependent measures of sequential complexity, such as expected sequential covering number [Rakhlin et al., 2010], that are modified to account for the state-space structure in the hypothesis set. The following definition of a complete binary tree is used throughout this paper: a $\mathcal{Z}$-valued complete binary tree $\mathbf{z}$ is a sequence $(z_1, \ldots, z_T)$ of T mappings $z_t \colon \{\pm 1\}^{t-1} \to \mathcal{Z}, t \in [1, T]$. A path in the tree is $\sigma = (\sigma_1, \ldots, \sigma_{T-1}) \in \{\pm 1\}^{T-1}$. We write $z_t(\boldsymbol{\sigma})$ instead of $z_t(\sigma_1, \ldots, \sigma_{t-1})$ to simplify the notation. Let $\mathcal{R} = \mathcal{R}' \times \mathcal{G}$ be any function class where $\mathcal{G}$ is a family of state predictors and $\mathcal{R}' = \{r \colon \mathcal{Z} \times \mathcal{S} \to \mathbb{R}\}$. A set V of $\mathbb{R}$-valued trees of depth T is a *sequential α-cover* (with respect to ℓ_p norm) of $\mathcal{R}$ on a tree $\mathbf{z}$ of depth T if for all $(r, g) \in \mathcal{R}$ and all $\boldsymbol{\sigma} \in \{\pm 1\}^T$, there is $\mathbf{v} \in V$ such that

$$\left[\frac{1}{T} \sum_{t=1}^{T} \left| \mathbf{v}_t(\boldsymbol{\sigma}) - r(\mathbf{z}_t(\boldsymbol{\sigma}), s_t) \right|^p \right]^{\frac{1}{p}} \leq \alpha,$$

where $s_t = g(\mathbf{z}_t(\boldsymbol{\sigma}), s_{t-1})$. The *(sequential) covering number* $\mathcal{N}_p(\alpha, \mathcal{R}, \mathbf{z})$ on a given tree $\mathbf{z}$ is defined to be the size of the minimal sequential cover. We call $\mathcal{N}_p(\alpha, \mathcal{R}) = \sup_{\mathbf{z}} \mathcal{N}_p(\alpha, \mathcal{R}, \mathbf{z})$ the *maximal covering number*. See Figure 1 for an example.

We define the *expected covering number* to be $\mathbb{E}_{\mathbf{z} \sim T(\mathbf{p})}[\mathcal{N}_p(\alpha, \mathcal{R}, \mathbf{z})]$, where $T(\mathbf{p})$ denotes the distribution of $\mathbf{z}$ implicitly defined via the following sampling procedure. Given a stochastic process distributed according to the distribution $\mathbf{p}$ with $\mathbf{p}_t(\cdot|\mathbf{z}_1^{t-1})$ denoting the conditional distribution at time t, sample Z_1, Z_1' from $\mathbf{p}_1$ independently. In the left child of the root sample Z_2, Z_2' according to $\mathbf{p}_2(\cdot|Z_1)$ and in the right child according to $\mathbf{p}_2(\cdot|Z_2')$ all independent from each other. For a node that can be reached by a path $(\sigma_1, \ldots, \sigma_t)$, we draw Z_t, Z_t' according to $\mathbf{p}_t(\cdot|S_1(\sigma_1), \ldots, S_{t-1}(\sigma_{t-1}))$, where $S_t(1) = Z_t$ and $S_t(-1) = Z_t'$. Expected sequential covering numbers are a finer measure of complexity since they directly take into account the distribution of the underlying stochastic process.

For further details on sequential complexity measures, we refer the reader to [Littlestone, 1987, Rakhlin et al., 2010, 2011, 2015a,b].

3 Theory

In this section, we present our generalization bounds for learning with DSSMs. For our first result, we assume that the sequence of states $\mathbf{s}$ (or equivalently state predictor g) is fixed and we are only learning the observation predictor h.

Theorem 1. *Fix $\mathbf{s} \in \mathcal{S}^{T+1}$. For any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$ and all $\alpha > 0$, the following inequality holds:*

$$\mathcal{L}(f|\mathbf{Z}_1^T) \leq \frac{1}{T} \sum_{t=1}^{T} L(h, X_t, s_t) + \text{disc}(\mathbf{s}) + 2\alpha + M\sqrt{\frac{2 \log \frac{\mathbb{E}_{\mathbf{v} \sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{R}_{\mathbf{s}}, \mathbf{v})]}{\delta}}{T}},$$

3

where $\mathcal{R}_\mathbf{s} = \{(z,s) \mapsto L(h,s,z) \colon h \in \mathcal{H}\} \times \{\mathbf{s}\}$.

The proof of Theorem 1 (as well as the proofs of all other results in this paper) is given in Appendix A. Note that this result is a generalization of the learning guarantees of Kuznetsov and Mohri [2015]. Indeed, setting $\mathbf{s} = (Tq_1, \dots, Tq_T, 1)$ for some weight vector $\mathbf{q}$ and $L(h,s,z) = s\widetilde{L}(h,z)$ recovers Corollary 2 of Kuznetsov and Mohri [2015]. Zimin and Lampert [2017] show that, under some additional assumptions on the underlying stochastic process (e.g. Markov processes, uniform martingales), it is possible to choose these weights to guarantee that the discrepancy $\mathrm{disc}(\mathbf{s})$ is small. Alternatively, Kuznetsov and Mohri [2015] show that if the distribution of the stochastic process at times $T+1$ and $[T-s,T]$ is sufficiently close (in terms of discrepancy) then $\mathrm{disc}(\mathbf{s})$ can be estimated from data. In Theorem 5 in Appendix C, we show that this property holds for arbitrary state sequences $\mathbf{s}$. Therefore, one can use the bound of Theorem 1 that can be computed from data to search for the predictor $h \in \mathcal{H}$ that minimizes this quantity. The quality of the result will depend on the given state-space sequence $\mathbf{s}$. Our next result shows that it is possible to learn $h \in \mathcal{H}$ and $\mathbf{s}$ generated by some state predictor $g \in \mathcal{G}$ jointly.

Theorem 2. *For any $\delta > 0$, with probability at least $1 - \delta$, for all $f = (h,g) \in \mathcal{H} \times \mathcal{G}$ and all $\alpha > 0$, the following inequality holds:*

$$\mathcal{L}(f|\mathbf{Z}_1^T) \leq \frac{1}{T}\sum_{t=1}^T L(h, X_t, s_t) + \mathrm{disc}(\mathbf{s}) + 2\alpha + M\sqrt{\frac{2\log\frac{\mathbb{E}_{\mathbf{v}\sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{R}, \mathbf{v})]}{\delta}}{T}},$$

where $s_t = g(X_t, s_{t-1})$ for all t and $\mathcal{R} = \{(z,s) \mapsto L(h,s,z) \colon h \in \mathcal{H}\} \times \mathcal{G}$.

The cost of this significantly more general result is a slightly larger complexity term $\mathcal{N}_1(\alpha, \mathcal{R}, \mathbf{v}) \geq \mathcal{N}_1(\alpha, \mathcal{R}_\mathbf{s}, \mathbf{v})$. This bound is also much tighter than the one that can be obtained by applying the result of Kuznetsov and Mohri [2015] directly to $\mathcal{F} = \mathcal{H} \times \mathcal{G}$, which would lead to the same bound as in Theorem 2 but with $\mathrm{disc}(\mathbf{s})$ replaced by $\sup_{g \in \mathcal{G}} \mathrm{disc}(\mathbf{s})$. Not only $\sup_{g \in \mathcal{G}} \mathrm{disc}(\mathbf{s})$ is an upper bound on $\mathrm{disc}(\mathbf{s})$, but it is possible to construct examples that lead to learning bounds that are too loose. Consider the stochastic process generated as follows. Let X be uniformly distributed on $\{\pm 1\}$. Suppose $Y_1 = 1$ and $Y_t = -Y_{t-1}$ for all $t > 1$ if $X = -1$ and $Y_t = Y_{t-1}$ for all $t > 1$ otherwise. In other words, $\mathbf{Y}$ is either periodic or a constant state sequence. If L is the squared loss, for $\mathcal{G} = \{g_1, g_2\}$ with $g_1(s) = s$ and $g_2(s) = -s$ and $\mathcal{H} = \{h\}$ with $h(s) = s$, for odd T, $\sup_{g \in \mathcal{G}} \mathrm{disc}(\mathbf{s}) \geq 1/2$. On the other hand, the bound in terms of $\mathrm{disc}(\mathbf{s})$ is much finer and helps us select g such that $\mathrm{disc}(\mathbf{s}) = 0$ for that g. This example shows that even for simple deterministic dynamics our learning bounds are finer than existing ones.

Since the guarantees of Theorem 2 are data-dependent and hold uniformly over $\mathcal{F}$, they allow us to seek a solution $f \in \mathcal{F}$ that would directly optimize this bound and that could be computed from the given sample. As our earlier example shows, the choice of the family of state predictors $\mathcal{G}$ is crucial to achieve good guarantees. For instance, if $\mathcal{G} = \{g_1\}$ then it may be impossible to have a non-trivial bound. In other words, if the family of state predictors is not rich enough, then, it may not be possible to handle the non-stationarity of the data. On the other hand, if $\mathcal{G}$ is chosen to be too large, then, the complexity term may be too large. In Section 4, we present an SRM technique that enables us to learn the state-space representation and adapt to non-stationarity in a principled way.

4 Structural Risk Minimization

Suppose we are given a sequence of families of observation predictors $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \mathcal{H}_n \dots$ and a sequence of families of state predictors $\mathcal{G}_1 \subset \mathcal{G}_2 \cdots G_n \dots$ Let $\mathcal{R}_k = \{(s,z) \mapsto L(h,s,z) \colon h \in \mathcal{H}_k\} \times \mathcal{G}_k$ and $\mathcal{R} = \cup_{k=1}^\infty \mathcal{R}_k$. Consider the following objective function:

$$F(h,g,k) = \frac{1}{T}\sum_{t=1}^T L(h, s_t, Z_t) + \Delta(\mathbf{s}) + B_k + M\sqrt{\frac{\log k}{T}}, \tag{5}$$

where $\Delta(\mathbf{s})$ is any upper bound on $\mathrm{disc}(\mathbf{s})$ and B_k is any upper bound on $M\sqrt{\frac{2\log\frac{\mathbb{E}_{\mathbf{v}\sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{R}_k, \mathbf{v})]}{\delta}}{T}}$. We are presenting an estimatable upper bound on $\mathrm{disc}(\mathbf{s})$ in Appendix C, which provides one

particular choice for $\Delta(\mathbf{s})$. In Appendix B, we also prove upper bounds on the expected sequential covering numbers for several families of hypothesis. Then, we define the SRM solution as follows:

$$(\widetilde{h}, \widetilde{g}, \widetilde{k}) = \operatorname{argmin}_{h,g \in \mathcal{H}_k \times \mathcal{G}_k, k \geq 1} F(h, g, k). \tag{6}$$

We also define f^* by $f^* = (h^*, g^*) \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_{T+1}(f|\mathbf{Z}_1^T)$. Then, the following result holds.

Theorem 3. *For any $\delta > 0$, with probability at least $1 - \delta$, for all $\alpha > 0$, the following bound holds:*

$$\mathcal{L}_{T+1}(\widetilde{h}, \widetilde{g}|\mathbf{Z}_1^T) \leq \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) + 2\Delta(\mathbf{s}^*) + 2\alpha + 2B_{k(f^*)} + M\sqrt{\frac{\log k(f^*)}{T}} + 2M\sqrt{\frac{\log \frac{2}{\delta}}{T}},$$

where $s_t^* = g^*(X_t, s_{t-1}^*)$, and where $k(f^*)$ is the smallest integer k such that $f^* \in \mathcal{H}_k \times \mathcal{G}_k$.

Theorem 3 provides a learning guarantee for the solution of SRM problem (5). This result guarantees for the SRM solution a performance close to that of the best-in-class model f^* modulo a penalty term that includes the discrepancy (of the best-in-class state predictor), similar to the guarantees of Section 3. This guarantee can be viewed as a worst case bound when we are unsure if the state-space predictor captures the non-stationarity of the problem correctly. However, in most cases, by introducing a state-space representation, we hope that it will help us model (at least to some degree) the non-stationarity of the underlying stochastic process. In what follows, we present a more optimistic best-case analysis which shows that, under some additional mild assumptions on the complexity of the hypothesis space with respect to stochastic process, we can simultaneously simplify the SRM optimization and give tighter learning guarantees for this modified version.

Assumption 1 (Stability of state trajectories). *Assume that there is a decreasing function r such that for any $\epsilon > 0$ and $\delta > 0$, with probability $1 - \delta$, if $h^*, g^* = \operatorname{argmin}_{(h,g) \in \mathcal{F}} \mathcal{L}_{T+1}(h, g|\mathbf{Z}_1^T)$ and $(h, g) \in \mathcal{F}$ is such that*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(h, g|\mathbf{Z}_1^{t-1}) - \mathcal{L}_t(h^*, g^*|\mathbf{Z}_1^{t-1}) \right| \leq \epsilon, \tag{7}$$

then, the following holds:

$$\mathcal{L}_{T+1}(h, g|\mathbf{Z}_1^T) - \mathcal{L}_{T+1}(h^*, g^*|\mathbf{Z}_1^T) \leq r(\epsilon). \tag{8}$$

Roughly speaking, this assumption states that, given a sequence of states $s_1, \ldots, s_T$ generated by g such that the performance of some observation predictor h along this sequence of states is close to the performance of the ideal pair h^* along the ideal sequence generated by g^*, the performance of h in the near future (at state s_{T+1}) will remain close to that of h^* (in state s_{T+1}^*). Note that, in most cases of interest, r has the form $r(\epsilon) = a\epsilon$, for some $a > 0$.

Consider the following optimization problem which is similar to (5) but omits the discrepancy upper bound Δ:

$$F_0(h, g, k) = \frac{1}{T} \sum_{t=1}^{T} L(h, s_t, Z_t) + B_k + M\sqrt{\frac{\log k}{T}}, \tag{9}$$

We will refer to F_0 as an *optimistic* SRM objective and we let (h_0, g_0) be a minimizer of F_0. Then, we have the following learning guarantee.

Theorem 4. *Under Assumption 1, for any $\delta > 0$, with probability at least $1 - \delta$, for all $\alpha > 0$, the inequality $\mathcal{L}_{T+1}(h_0, g_0|\mathbf{Z}_1^T) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) < r(\epsilon)$ holds with*

$$\epsilon = 2\alpha + 2B_{k(f^*)} + M\sqrt{\frac{\log k(f^*)}{T}} + 2M\sqrt{\frac{\log \frac{2}{\delta}}{T}},$$

where $s_t^* = g^*(X_t, s_{t-1}^*)$, and where $k(f^*)$ is the smallest integer k such that $f^* \in \mathcal{H}_k \times \mathcal{G}_k$.

We remark that a finer analysis can be used to show that Assumption 1 only need to be satisfied for $k \leq k(f^*)$ for the Theorem 4. Furthermore, observe that for linear functions $r(\epsilon) = a\epsilon$, one recovers a guarantee similar to the bound in Theorem 3, but the discrepancy term is omitted making this result tighter. This result suggests that in the optimistic scenarios where our hypothesis set contains a good

5

state predictor that can capture the data non-stationarity, it is possible to achieve a tighter guarantee that avoids the pessimistic discrepancy term. Note that, increasing the capacity of the family of state predictors makes it easier to find such a good state predictor but it also may make the learning problem harder and lead to the violation of Assumption 1. This further motivates the use of an SRM technique for this problem to find the right balance between capturing the non-stationarity in data and the complexity of the models that are being used. Theorem 4 formalizes this intuition by providing theoretical guarantees for this approach.

We now consider several illustrative examples showing that this assumption holds in a variety of cases of interest. In all our examples, we will use the squared loss but it is possible to generalize all of them to other sufficiently regular losses.

Linear models. Let $\mathcal{F}$ be defined by $\mathcal{F} = \{f \colon \mathbf{y} \mapsto \mathbf{w} \cdot \Psi(\mathbf{y}), \|\mathbf{w}\| \leq \Lambda\}$ for some $\Lambda > 0$ and some feature map Ψ. Consider a separable case where $Y_t = \mathbf{w}^* \cdot \Psi(\mathbf{Y}_{t-p}^{t-1}) + \epsilon_t$, where ϵ_t represents white noise. One can verify that the following equality holds:

$$\mathcal{L}_t(\mathbf{w}|\mathbf{Z}_1^{t-1}) = \mathbb{E}[(\mathbf{w} \cdot \Psi(\mathbf{Y}_{t-p}^{t-1}) - Y_t)|\mathbf{Y}_1^{t-1}] = \left[(\mathbf{w} - \mathbf{w}^*) \cdot \Psi(\mathbf{Y}_{t-p}^{t-1})\right]^2.$$

In view of that, it follows that (7) is equal to

$$\frac{1}{T}\sum_{t=1}^{T}\left[(\mathbf{w} - \mathbf{w}^*) \cdot \Psi(\mathbf{Y}_{t-p}^{t-1})\right]^2 \geq \frac{1}{T}\sum_{t=1}^{T}(\mathbf{w}_j - \mathbf{w}_j^*)^2\Psi_j(\mathbf{Y}_{t-p}^{t-1})^2$$

for any coordinate $j \in [1, N]$. Thus, for any coordinate $j \in [1, N]$, by Hölder's inequality, we have

$$\mathcal{L}_{T+1}(h, g|\mathbf{Z}_1^T) - \mathcal{L}_{T+1}(h^*, g^*|\mathbf{Z}_1^T) = \left[(\mathbf{w} - \mathbf{w}^*) \cdot \Psi(\mathbf{Y}_{T-p+1}^T)\right]^2 \leq r\epsilon \sum_{j=1}^{N}\frac{1}{\sigma_j},$$

where $\sigma_j = \frac{1}{T}\sum_{t=1}^{T}\Psi_j(\mathbf{Y}_{t-p}^{t-1})^2$ is the empirical variance of the j-th coordinate and where $r = \sup_{\mathbf{y}}\Psi(\mathbf{y})^2$ is the empirical ℓ_∞-norm radius of the data. The special case where Ψ is the identity map covers standard autoregressive models. These often serve as basic building blocks for other state-space models, as discussed below. More generally, other feature maps Ψ may be induced by a positive definite kernel K. Alternatively, we may take as our hypothesis set $\mathcal{F}$ the convex hull of all decision trees of certain depth d. In that case, we can view each coordinate Ψ_j as the output of a particular decision tree on the given input.

Linear trend models. For simplicity, in this example, we consider univariate time series with linear trend. However, this can be easily generalized to the multi-variate setting with different trend models. Define $\mathcal{G}$ as $\mathcal{G} = \{s \mapsto s + c \colon |c| \leq \Lambda\}$ for some $\Lambda > 0$ and let $\mathcal{H}$ be a singleton consisting of the identity map. Assume that $Y_t = c^*t + \epsilon_t$, where ϵ_t is white noise. As in the previous example, it is easy to check that $\mathcal{L}_t(h, g|\mathbf{Z}_1^{t-1}) = |c - c^*|^2t^2$. Therefore, in this case, one can show that (7) reduces to $\frac{1}{3}(T+1)(2T+1)|c - c^*|^2$ and therefore, if $\epsilon = O(\sqrt{1/T})$, then we have $|c - c^*|^2 = O(1/T^{5/2})$ and thus (8) is $|c - c^*|^2(T+1)^2 = O(\sqrt{1/T})$.

Periodic signals. We study a multi-resolution setting where the time series of interest are modeled as a linear combination of periodic signals at different frequencies. We express this as a state-space model as follows. Define

$$\mathbf{A}_d = \begin{bmatrix} -\mathbf{1} & -1 \\ \mathbf{I}_{d-1} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{1}$ is $d-1$-dimensional row vector of 1s, $\mathbf{0}$ is $d-1$-dimensional column vector of 0 and $\mathbf{I}_{d-1}$ is an identity matrix. It easy to verify that, under the map $\mathbf{s} \mapsto \mathbf{A}_d\mathbf{s}$, the sequence $\mathbf{s}_1 \cdot \mathbf{e_1}, \mathbf{s}_2 \cdot \mathbf{e}_1 \ldots, \mathbf{s}_t \cdot \mathbf{e}_1 \ldots$, where $\boldsymbol{\epsilon}_1 = (1, 0, \ldots, 0)^T$, is a periodic sequence with period d. Let $D = d_1, \ldots, d_k$ be any collection of positive integers and let $\mathbf{A}$ be a block-diagonal matrix with $\mathbf{A}_{d_1}, \ldots, \mathbf{A}_{d_k}$ on the diagonal. We set $\mathcal{G} = \{\mathbf{s} \mapsto \mathbf{A} \cdot \mathbf{s}\}$ and $\mathcal{H} = \{\mathbf{s} \mapsto \mathbf{w} \cdot \mathbf{s} \colon \|\mathbf{w}\| < \Lambda\}$, where we also restrict $\mathbf{w}$s to be non-zero only at coordinates $1, 1 + d_1, 1 + d_1 + d_2, \ldots, 1 + \sum_{j=1}^{k-1}d_{k-1}$. Once again, to simplify our presentation, we assume that Y_t satisfies $Y_t = \mathbf{w}^* \cdot \mathbf{s}_t + \epsilon_t$. Using arguments similar to those of the previous examples, one can show that (7) is lower bounded by $(\mathbf{w}_j - \mathbf{w}_j^*)^2\frac{1}{T}\sum_{t=1}^{T}\mathbf{s}_{t,j}$ for any coordinate j. Therefore, as before, if (7) is upper bounded by $\epsilon > 0$, then (8) is upper bounded by $r\epsilon \sum_{j=1}^{N}\frac{1}{\sigma_j}$, where r is the maximal radius of any state and σ_j a variance of j-th state sequence.

Trajectory ensembles. Note that, in our previous example, we did not exploit the fact that the sequences were periodic. Indeed, our argument holds for any g that generates a multi-dimensional trajectory $h \in \mathcal{H} = \{\mathbf{s} \mapsto \mathbf{w} \cdot \mathbf{s} \colon \|\mathbf{w}\| < \Lambda\}$ which can be interpreted as learning an ensemble of different state-space trajectories.

Structural Time Series Models (STSMs). STSMs are a popular family of state-space models that combine all of the previous examples. For this model, we use $(h, g) \in \mathcal{H} \times \mathcal{G}$ that have the following structure: $h(\mathbf{x}_t, g(\mathbf{s}_t)) = \mathbf{w} \cdot \Psi(\mathbf{x}_t) + ct + \mathbf{w}' \cdot \mathbf{s}_t$, where $\mathbf{s}_t$ is a vector of periodic sequences described in the previous examples and $\mathbf{x}_t$ is the vector representing the most recent history of the time series. Note that our formulation is very general and allows for arbitrary feature maps Ψ that can correspond either to kernel-based or tree-based models. Arguments similar to those given in previous examples show that Assumption 1 holds in this case.

Shifting parameters. We consider the non-realizable case where $\mathcal{H}$ is a set of linear models but where the data is generated according to the following procedure. The first $T/2$ rounds obey the formula $Y_t = \mathbf{w}_0 Y_{t-1} + \epsilon_t$, the subsequent rounds the formula $Y_t = \mathbf{w}^* Y_{t-1} + \epsilon_t$. Note that, in this case, we have $|\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{w}_0 | \mathbf{Z}_1^{t-1}) - \mathcal{L}_t(\mathbf{w}^* | \mathbf{Z}_1^{t-1})| = 0$. However, if $\mathbf{w}_0$ and $\mathbf{w}^*$ are sufficiently far apart, it is possible to show that there is a constant lower bound on $\mathcal{L}_{T+1}(\mathbf{w}_0 | \mathbf{Z}_1^T) - \mathcal{L}_{T+1}(\mathbf{w}^* | \mathbf{Z}_1^T)$. One approach to making Assumption 1 hold for this stochastic process is to choose $\mathcal{H}$ such that the resulting learning problem is separable. However, that requires us to know the exact nature of the underlying stochastic process. An alternative agnostic approach, is to consider a sequence of states (or equivalently weights) that can assign different weights q_t to different training points.

Finally, observe that our learning guarantees in Section 3 and 4 are expressed in terms of the expected sequential covering numbers of the family of DSSMs that we are seeking to learn. A priori, it is not clear if it is possible to control the complexity of such models in a meaningful way. However, in Appendix B, we present explicit upper bounds on the expected sequential covering numbers of several families of DSSMs, including several of those discussed above: linear models, tree-based hypothesis, and trajectory ensembles.

5 Algorithmic Solutions

The generic SRM procedures described in Section 4 can lead to the design of a range of different algorithmic solutions for forecasting time series, depending on the choice of the families $\mathcal{H}_k$ and $\mathcal{F}_k$. The key challenge for the design of an algorithm design in this setting is to come up with a tractable procedure for searching through sets of increasing complexity. In this section, we describe one such procedure that leads to a boosting-style algorithm. Our algorithm learns a structural time series model by adaptively adding various structural subcomponents to the model in order to balance model complexity and the ability of the model to handle non-stationarity in data. We refer to our algorithm as *Boosted Structural Time Series Models* (BOOSTSM).

We will discuss BOOSTSM in the context of the squared loss, but most of our results can be straightforwardly extended to other convex loss functions. The hypothesis set used by our algorithm admits the following form: $\mathcal{H} = \{(\mathbf{x}, \mathbf{s}) \mapsto \mathbf{w} \cdot \Psi(\mathbf{x}) + \mathbf{w}' \cdot \mathbf{s} \colon \|\mathbf{w}\|_1 \leq \Lambda, \|\mathbf{w}'\|_1 \leq \Lambda'\}$. Each coordinate Ψ_j is a binary-valued decision tree maps its inputs to a bounded set. For simplicity, we also assume that $\Lambda = \Lambda' = 1$. We choose $\mathcal{G}$ to be any set of state trajectories. For instance, this set may include periodic or trend sequences as described in Section 4.

Note that, to make the discussion concrete, we impose an ℓ_1-constraint to the parameter vectors, but other regularization penalties are also possible. The particular choice of the regularization defined by $\mathcal{H}$ would also lead to sparser solutions, which is an additional advantage given that our state-space representation is high-dimensional.

For the squared loss and the aforementioned $\mathcal{H}$, the optimistic SRM objective (9) is given by

$$F(\mathbf{w}, \mathbf{w}') = \frac{1}{T} \sum_{t=1}^{T} \left(y_t - \mathbf{w} \cdot \Psi(\mathbf{x}_t) + \mathbf{w}' \cdot \mathbf{s}_t \right)^2 + \lambda(\|\mathbf{w}\|_1 + \|\mathbf{w}'\|_1), \tag{10}$$

where we omit $\log(k)$ because the index k in our setting tracks the maximal depth of the tree and it suffices to restrict the search to the case $k < T$ as, for deeper trees, we can achieve zero empirical error. With this upper bound on k, $O\left(\sqrt{\frac{\log T}{T}}\right)$ is small and hence not included in the objective.

BOOSTSM($S = ((x_i, y_i)_{t=1}^T)$)
1    $f_0 \leftarrow 0$
2   **for** $k \leftarrow 1$ **to** K **do**
3        $j \leftarrow \operatorname{argmin}_j \epsilon_{k,j} + \lambda \operatorname{sgn}(w_j)$
4        $j' \leftarrow \operatorname{argmin}_{j'} \delta_{k,j'} + \lambda \operatorname{sgn}(w'_j)$
5   **if** $\epsilon_{k,j} + \lambda \operatorname{sgn}(w_j) \leq \delta_{k,j'} + \lambda \operatorname{sgn}(w'_j)$ **then**
6            $\eta_k \leftarrow \operatorname{argmin}_\eta F(\mathbf{w} + \eta \boldsymbol{\epsilon}_j, \mathbf{w}')$
7            $f_k \leftarrow f_{k-1} + \eta_k \Psi_j$
8   **else** $\eta_k \leftarrow \operatorname{argmin}_\eta F(\mathbf{w}, \mathbf{w}' + \eta \boldsymbol{\epsilon}_{j'})$
9            $f_k \leftarrow f_{k-1} + \eta_t \boldsymbol{\epsilon}_{j'}$
10   **return** f_K

Figure 1: Pseudocode of the BOOSTSM algorithm. On line 3 and 4 two candidates are selected to be added to the ensemble: a state trajectory with j' or a tree-based predictor with index j. Both of these minimize their subgradients within their family of weak learners. Subgradients are defined by (11). The candidate with the smallest gradient is added to the ensemble. The weight of the new ensemble member is found via line search (line 6 and 8).

The regularization penalty is directly derived from the bounds on the expected sequential covering numbers of $\mathcal{H}$ given in Appendix B in Lemma 4 and Lemma 5.

Observe that (10) is a convex objective function. Our BOOSTSM algorithm is defined by the application of coordinate descent to this objective. Figure 1 gives its pseudocode. The algorithm proceeds in K rounds. At each round, we either add a new predictor tree or a new state-space trajectory to the model, depending on which results in a greater decrease in the objective. In particular, with the following definitions:

$$\epsilon_{k,j} = \frac{1}{T} \sum_{t=1}^T (y_t - f_{k-1}(\mathbf{x}_t, \mathbf{s}_t)) \Psi_j(\mathbf{x}_t), \quad \delta_{k,j} = \frac{1}{T} \sum_{t=1}^T (y_t - f_{k-1}(\mathbf{x}_t, \mathbf{s}_t)) s_{t,j}. \quad (11)$$

the subgradient in tree-space direction j at round k is given by $\epsilon_{k,j} + \lambda \operatorname{sgn}(w_{k,j})$. We use the notation $\mathbf{w}_k$ to denote the tree-space parameter vector after $k-1$ rounds. Similarly, the subgradient in the trajectory space direction j' is given by $\delta_{k,j'} + \lambda \operatorname{sgn}(w'_{k,j})$, where $\mathbf{w}'_k$ represents the trajectory space parameter vector after $k-1$ rounds.

By standard results in optimization theory [Luo and Tseng, 1992], BOOSTSM admits a linear convergence guarantee.

6 Conclusion

We introduced a new family of models for forecasting non-stationary time series, *Discriminative State-Space Models*. This family includes existing generative models and other state-based discriminative models (e.g. RNNs) as special cases, but also covers several novel algorithmic solutions explored in this paper. We presented an analysis of the problem of learning DSSMs in the most general setting of non-stationary stochastic processes and proved finite-sample data-dependent generalization bounds. These learning guarantees are novel even for traditional state-space models since the existing guarantees are only asymptotic and require the model to be correctly specified. We fully analyzed the complexity of several DSSMs that are useful in practice. Finally, we also studied the generalization guarantees of several structural risk minimization approaches to this problem and provided an efficient implementation of one such algorithm which is based on a convex objective. We report some promising preliminary experimental results in Appendix D.

References

Rakesh D. Barve and Philip M. Long. On the complexity of learning from drifting distributions. In *COLT*, 1996.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J Econometrics*, 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, 1986.

J.J.F. Commandeur and S.J. Koopman. *An Introduction to State Space Time Series Analysis*. OUP Oxford, 2007.

Corinna Cortes, Giulia DeSalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yand. Multi-armed bandits with non-stationary rewards. *CoRR*, abs/1710.10657, 2017.

Victor H. De la Peña and Evarist Giné. *Decoupling: from dependence to independence: randomly stopped processes, U-statistics and processes, martingales and beyond*. Probability and its applications. Springer, NY, 1999.

J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford, 2012.

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

James D. Hamilton. *Time series analysis*. Princeton, 1994.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D), 1960.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for time series prediction with non-stationary processes. In *ALT*, 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, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and on-line learning. In *Proceedings of The 29th Conference on Learning Theory, COLT 2016*, 2016.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. U.S. Government Printing Office, 1991.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 1987.

Zhi-Quan Luo and Paul Tseng. On the convergence of coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7 – 35, 1992.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, 2011.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *JMLR*, 16(1), January 2015b.

Alexander Zimin and Christopher H. Lampert. Learning theory for conditional risk minimization. In *AISTAT*, 2017.

A Proofs

One of the key quantities that is used in our proofs is the empirical process:

$$\Phi(h, g) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(h, g | \mathbf{Z}_1^{t-1}) - \frac{1}{T} \sum_{t=1}^{T} L(h, s_t, Z_t), \tag{12}$$

where (h, g) is an element of $\mathcal{F}$ and $\mathbf{s}$ the sequence generated by g. The following result gives a high-probability bound on the supremum of such empirical process.

Another definition that is required in the proofs below is the notion of decoupled tangent sequence. Given a sequence of random variables $\mathbf{Z}_1^T$ we say that $\mathbf{Z}_1'^T$ is a decoupled tangent sequence if Z_t' is distributed according to $\mathbb{P}(\cdot | \mathbf{Z}_1^{t-1})$ and is independent of $\mathbf{Z}_t^\infty$. It is always possible to construct such a sequence of random variables [De la Peña and Giné, 1999]. We begin with the following concentration result.

Lemma 1. *Fix $\epsilon > 2\alpha > 0$. Then, the following holds:*

$$\mathbb{P}\left(\sup_{h,g \in \mathcal{F}} \Phi(h, g) \geq \epsilon \right) \leq \mathop{\mathbb{E}}_{\mathbf{v} \sim T(\mathbf{p})} \left[\mathcal{N}_1(\alpha, \mathcal{R}, \mathbf{v}) \right] \exp\left(-\frac{T(\epsilon - 2\alpha)^2}{2M^2} \right).$$

Symmetric result with $\Phi(h, g)$ replaced by $-\Phi(h, g)$ also holds.

Proof. Note that, for any sequence $z_1, \ldots, z_T$, we can write sequence $L(h, s_1, z_1), \ldots, L(h, s_T, z_T)$, (where $\mathbf{s}$ is generated by g) as a sequence $f_1(z_1), \ldots, f_T(z_T)$. The rest of the proof follows the same arguments as in Theorem 1 in Kuznetsov and Mohri [2015] where instead of a fixed f at each time t, we use f_t. For completeness, we include the full derivation of the result.

By Markov's inequality, for any $\lambda > 0$, the following inequality holds:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^{T} (\mathbb{E}[f_t(Z_t) | \mathbf{Z}_1^{t-1}] - f_t(Z_t)) \right) \geq \epsilon \right)$$

$$\leq \exp(-\lambda \epsilon) \mathbb{E}\left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^{T} (\mathbb{E}[f_t(Z_t) | \mathbf{Z}_1^{t-1}] - f_t(Z_t)) \right) \right) \right].$$

If $\mathbf{Z}_1'^T$ is a tangent sequence the following equalities hold: $\mathbb{E}[f_t(Z_t) | \mathbf{Z}_1^{t-1}] = \mathbb{E}[f_t(Z_t') | \mathbf{Z}_1^{t-1}] = \mathbb{E}[f_t(Z_t') | \mathbf{Z}_1^T]$. Using these equalities and Jensen's inequality, we obtain the following:

$$\mathbb{E}\left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \frac{1}{T} \left(\mathbb{E}[f_t(Z_t) | \mathbf{Z}_1^{t-1}] - f_t(Z_t) \right) \right) \right]$$

$$= \mathbb{E}\left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{T} \left(f_t(Z_t') - f_t(Z_t) \right) | \mathbf{Z}_1^T \right] \right) \right]$$

$$\leq \mathbb{E}\left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \left(f_t(Z_t') - f_t(Z_t) \right) \right) \right],$$

where the last expectation is taken over the joint measure of $\mathbf{Z}_1^T$ and $\mathbf{Z}_1'^T$. Applying Lemma 2, we can further bound this expectation by

$$\mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}') \sim T(\mathbf{p})} \mathop{\mathbb{E}}_{\sigma} \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t \left(f_t(\mathbf{z}'_t(\boldsymbol{\sigma})) - f_t(\mathbf{z}_t(\boldsymbol{\sigma})) \right) \right) \right]$$

$$\leq \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}') \sim T(\mathbf{p})} \mathop{\mathbb{E}}_{\sigma} \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t f_t(\mathbf{z}'_t(\boldsymbol{\sigma})) + \lambda \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} -\sigma_t f_t(\mathbf{z}_t(\boldsymbol{\sigma})) \right) \right]$$

$$\leq \tfrac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')} \mathop{\mathbb{E}}_{\sigma} \left[\exp\left(2\lambda \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t f_t(\mathbf{z}'_t(\boldsymbol{\sigma})) \right) \right] + \tfrac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')} \mathop{\mathbb{E}}_{\sigma} \left[\exp\left(2\lambda \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t f_t(\mathbf{z}_t(\boldsymbol{\sigma})) \right) \right]$$

$$= \mathop{\mathbb{E}}_{\mathbf{z} \sim T(\mathbf{p})} \mathop{\mathbb{E}}_{\sigma} \left[\exp\left(2\lambda \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t f_t(\mathbf{z}_t(\boldsymbol{\sigma})) \right) \right],$$

where for the second inequality we used Young's inequality and for the last equality we used symmetry. Given $\mathbf{z}$ let C denote the minimal α-cover with respect to the ℓ_1-norm of $\mathcal{F}$ on $\mathbf{z}$. Then, the following bound holds

$$\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\sigma_t f_t(\mathbf{z}_t(\boldsymbol{\sigma})) \leq \max_{\mathbf{c}\in C}\frac{1}{T}\sum_{t=1}^{T}\sigma_t \mathbf{c}_t(\boldsymbol{\sigma}) + \alpha.$$

By the monotonicity of the exponential function,

$$\mathbb{E}_{\sigma}\left[\exp\left(2\lambda\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\sigma_t f_t(\mathbf{z}_t(\boldsymbol{\sigma}))\right)\right] \leq \exp(2\lambda\alpha)\,\mathbb{E}_{\sigma}\left[\exp\left(2\lambda\max_{\mathbf{c}\in C}\frac{1}{T}\sum_{t=1}^{T}\sigma_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\right]$$

$$\leq \exp(2\lambda\alpha)\sum_{\mathbf{c}\in C}\mathbb{E}_{\sigma}\left[\exp\left(2\lambda\sum_{t=1}^{T}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\right].$$

Since $\mathbf{c}_t(\boldsymbol{\sigma})$ depends only on $\sigma_1,\ldots,\sigma_{T-1}$, by Hoeffding's bound,

$$\mathbb{E}_{\sigma}\left[\exp\left(2\lambda\frac{1}{T}\sum_{t=1}^{T}\sigma_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\right] = \mathbb{E}\left[\exp\left(2\lambda\frac{1}{T}\sum_{t=1}^{T-1}\sigma_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\mathbb{E}_{\sigma_T}\left[\exp\left(2\lambda\sigma_T\frac{1}{T}\mathbf{c}_T(\boldsymbol{\sigma})\right)\Big|\boldsymbol{\sigma}_1^{T-1}\right]\right]$$

$$\leq \mathbb{E}\left[\exp\left(2\lambda\frac{1}{T}\sum_{t=1}^{T-1}\sigma_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\exp(2\lambda^2\frac{1}{T^2}M^2)\right]$$

and iterating this inequality and using the union bound, we obtain the following:

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}]-f(Z_t)) \geq \epsilon\right) \leq \mathbb{E}_{\mathbf{v}\sim T(\mathbf{p})}[\mathcal{N}_1(\alpha,\mathcal{R},\mathbf{v})]\exp\left(-\lambda(\epsilon-2\alpha)+2\frac{1}{T}\lambda^2 M^2\right).$$

Optimizing over λ completes the proof of the first statement. The second statement holds by symmetry. $\square$

The following lemma is used in the proof of Lemma 1.

Lemma 2. *Given a sequence of random variables $\mathbf{Z}_1^T$ with joint distribution $\mathbf{p}$, let $\mathbf{Z}'^T_1$ be a decoupled tangent sequence. Then, for any measurable function G, the following equality holds*

$$\mathbb{E}\left[G\left(\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}(f_t(Z'_t)-f_t(Z_t))\right)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\mathbb{E}_{\mathbf{z}\sim T(\mathbf{p})}\left[G\left(\sup_{f}\frac{1}{T}\sum_{t=1}^{T}\sigma_t(f_t(\mathbf{z}'_t(\boldsymbol{\sigma}))-f_t(\mathbf{z}_t(\boldsymbol{\sigma})))\right)\right].$$
(13)

Proof. The proof follows an argument in the proof of Theorem 3 of [Rakhlin et al., 2011]. We only need to check that every step holds for a time-varying f_t instead f fixed over time and for an arbitrary measurable function G, instead of the identity function. Observe that we can write the left-hand side of (13) as

$$\mathbb{E}\left[G\left(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\right)\right] = \mathbb{E}_{Z_1,Z'_1\sim\mathbf{p}_1}\mathbb{E}_{Z_2,Z'_2\sim\mathbf{p}_2(\cdot|Z_1)}\cdots\mathbb{E}_{Z_T,Z'_T\sim\mathbf{p}_T(\cdot|\mathbf{Z}_1^{T-1})}\left[G\left(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\right)\right],$$

where $\boldsymbol{\sigma} = (1,\ldots,1)\in\{\pm1\}^T$ and $\Sigma(\boldsymbol{\sigma}) = \sum_{t=1}^{T}\sigma_t\frac{1}{T}(f_t(Z'_t)-f_t(Z_t))$. Now, by definition of decoupled tangent sequences, the value of the last expression is unchanged if we swap the sign of any σ_{i-1} to -1 since that is equivalent to permuting Z_i and Z'_i. Thus, the last expression is in fact equal to

$$\mathbb{E}_{Z_1,Z'_1\sim\mathbf{p}_1}\mathbb{E}_{Z_2,Z'_2\sim\mathbf{p}_2(\cdot|S_1(\sigma_1))}\cdots\mathbb{E}_{Z_T,Z'_T\sim\mathbf{p}_T(\cdot|S_1(\sigma_1),\ldots,S_{T-1}(\sigma_{T-1}))}\left[G\left(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\right)\right]$$

for any sequence $\boldsymbol{\sigma}\in\{\pm1\}^T$, where $S_t(1) = Z_t$ and Z'_t otherwise. Since this equality holds for any $\boldsymbol{\sigma}$, it also holds for the mean with respect to uniformly distributed $\boldsymbol{\sigma}$. Therefore, the last expression is equal to

$$\mathbb{E}_{\boldsymbol{\sigma}}\mathbb{E}_{Z_1,Z'_1\sim\mathbf{p}_1}\mathbb{E}_{Z_2,Z'_2\sim\mathbf{p}_2(\cdot|S_1(\sigma_1))}\cdots\mathbb{E}_{Z_T,Z'_T\sim\mathbf{p}_T(\cdot|S_1(\sigma_1),\ldots,S_{T-1}(\sigma_{T-1}))}\left[G\left(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\right)\right].$$

This last expectation coincides with the expectation with respect to drawing a random tree $\mathbf{z}$ from $T(\mathbf{p})$ (and its tangent tree $\mathbf{z}'$) and a random path $\boldsymbol{\sigma}$ to follow in that tree. That is, the last expectation is equal to

$$\mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathbf{z} \sim T(\mathbf{p})} \left[G\left(\sup_f \frac{1}{T} \sum_{t=1}^{T} \sigma_t (f_t(\mathbf{z}'_t(\boldsymbol{\sigma})) - f_t(\mathbf{z}_t(\boldsymbol{\sigma}))) \right) \right],$$

which concludes the proof. $\qquad\square$

Now we proceed with the proofs of our main results.

Theorem 1. *Fix* $\mathbf{s} \in \mathcal{S}^{T+1}$. *For any* $\delta > 0$, *with probability at least* $1 - \delta$, *for all* $h \in \mathcal{H}$ *and all* $\alpha > 0$,

$$\mathcal{L}(f|\mathbf{Z}_1^T) \leq \frac{1}{T} \sum_{t=1}^{T} L(h, X_t, s_t) + \mathrm{disc}(\mathbf{s}) + 2\alpha + M \sqrt{\frac{2 \log \frac{\mathbb{E}_{\mathbf{v} \sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{R}_\mathbf{s}, \mathbf{v})]}{\delta}}{T}},$$

where $\mathcal{R}_\mathbf{s} = \{(z, s) \mapsto L(h, s, z) \colon h \in \mathcal{H}\} \times \{\mathbf{s}\}$.

Proof. The proof of this result immediately follows from Theorem 2 by fixing $\mathcal{G}$ to be a singelton $\{g\}$, where g generates $\mathbf{s}$ on any input. $\qquad\square$

Theorem 2. *For any* $\delta > 0$, *with probability at least* $1 - \delta$, *for all* $f = (h, g) \in \mathcal{H} \times \mathcal{G}$ *and all* $\alpha > 0$,

$$\mathcal{L}(f|\mathbf{Z}_1^T) \leq \frac{1}{T} \sum_{t=1}^{T} L(h, X_t, s_t) + \mathrm{disc}(\mathbf{s}) + 2\alpha + M \sqrt{\frac{2 \log \frac{\mathbb{E}_{\mathbf{v} \sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{R}, \mathbf{v})]}{\delta}}{T}},$$

where $s_t = g(X_t, s_{t-1})$ *for all* t *and* $\mathcal{R} = \{(z, s) \mapsto L(h, s, z) \colon h \in \mathcal{H}\} \times \mathcal{G}$.

Proof. Define the following empirical process

$$\Phi_0(h, g) = \mathcal{L}(f|\mathbf{Z}_1^T) - \frac{1}{T} \sum_{t=1}^{T} L(h, X_t, s_t).$$

Observe that since difference of supermums is upper bounded by the supremum of the difference the following inequality holds

$$\sup_{g \in \mathcal{G}} \left(\sup_{h \in \mathcal{H}} \Phi_0(h, g) - \mathrm{disc}(\mathbf{s}_g) \right) \leq \sup_{g \in \mathcal{G}} \sup_{h \in \mathcal{H}} \Phi(h, g),$$

where we use notation $\mathbf{s}_g$ to emphasize dependence of state sequences $\mathbf{s}$ on g. The result now follows by application of Lemma 1. $\qquad\square$

Theorem 3. *For any* $\delta > 0$, *with probability at least* $1 - \delta$, *for all* $\alpha > 0$, *the following bound holds:*

$$\mathcal{L}_{T+1}(\widetilde{h}, \widetilde{g}|\mathbf{Z}_1^T) \leq \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) + 2\Delta(\mathbf{s}^*) + 2\alpha + 2B_{k(f^*)} + M \sqrt{\frac{\log k(f^*)}{T}} + 2M \sqrt{\frac{\log \frac{2}{\delta}}{T}},$$

where $s_t^* = g^*(X_t, s_{t-1}^*)$ *and* $k(f^*)$ *is the smallest integer* k *such that* $f^* \in \mathcal{H}_k \times \mathcal{G}_k$.

Proof. To simplify the notation, we let $\mathcal{N}(k) = \mathbb{E}_{\mathbf{v} \sim T(\mathbb{P})}[\mathcal{N}_1(\alpha, \mathcal{R}_k, \mathbf{v})]$. Observe that the following chain of inequalities holds:

$$
\mathbb{P}\left(\sup_{k \geq 1} \sup_{f \in \mathcal{F}_k} \mathcal{L}_{T+1}(f|\mathbf{Z}_1^T) - F(f, k) - 2\alpha \geq \epsilon \right)
$$

$$
\leq \sum_{k=1}^{\infty} \mathbb{P}\left(\sup_{f \in \mathcal{F}_k} \mathcal{L}_{T+1}(f|\mathbf{Z}_1^T) - F(f, k) - 2\alpha \geq \epsilon \right)
$$

$$
= \sum_{k=1}^{\infty} \mathbb{P}\left(\sup_{g \in \mathcal{G}_k} \Phi_0(h, g) - \mathrm{disc}(\mathbf{s}) \geq \epsilon + 2\alpha + M\sqrt{\frac{\log k}{T}} + M\sqrt{\frac{2 \log \mathcal{N}(k)}{T}} \right)
$$

$$
\leq \sum_{k=1}^{\infty} \mathcal{N}(k) \exp\left(-\frac{1}{2M^2} T\left(\epsilon + M\sqrt{\frac{\log k}{T}} + M\sqrt{\frac{2 \log \mathcal{N}(k)}{T}} \right)^2 \right)
$$

$$
\leq \sum_{k=1}^{\infty} \mathcal{N}(k) \exp\left(-\frac{1}{2M^2} T\epsilon^2 - 2 \log k - \log \mathcal{N}(k) \right)
$$

$$
= \exp\left(-\frac{2T\epsilon^2}{M^2} \right) \sum_{k=1}^{\infty} \frac{1}{k^2}
$$

$$
= \frac{\pi^2}{6} \exp\left(-\frac{2T\epsilon^2}{M^2} \right)
$$

$$
\leq 2 \exp\left(-\frac{2T\epsilon^2}{M^2} \right),
$$

where we use the union bound for the first inequality, Theorem 2 for the second inequality and $(a + b)^2 \leq 2a^2 + 2b^2$ for the third inequality. Next, we observe that

$$
\mathbb{P}\left(\mathcal{L}_{T+1}(\widetilde{h}, \widetilde{g}|\mathbf{Z}_1^T) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) - 2\Delta(\mathbf{s}^*) - 2\alpha - 2B_{k(f^*)} - M\sqrt{\frac{\log k(f^*)}{T}} > \epsilon \right)
$$

$$
\leq \mathbb{P}\left(\mathcal{L}_{T+1}(\widetilde{h}, \widetilde{g}|\mathbf{Z}_1^T) - F(\widetilde{f}, \widetilde{k}) - \alpha \geq \frac{\epsilon}{2} \right)
$$

$$
+ \mathbb{P}\left(F(f^*, k(f^*)) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) - 2\Delta(\mathbf{s}^*) - \alpha - 2B_{k(f^*)} - M\sqrt{\frac{\log k(f^*)}{T}} > \frac{\epsilon}{2} \right),
$$

where we used the union bound and the fact that $F(\widetilde{f}, \widetilde{k}) - F(f^*, k(f^*)) \leq 0$ since $\widetilde{f}$ is minimizer of F. The first term is bounded by $2 \exp\left(-\frac{2T\epsilon^2}{M^2} \right)$ by the previous argument. Similarly, by Lemma 1, the second term is bounded by $2 \exp\left(-\frac{2T\epsilon^2}{M^2} \right)$ since

$$
F(f^*, k(f^*)) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) - 2\Delta(\mathbf{s}^*) - \alpha - 2B_{k(f^*)} - M\sqrt{\frac{\log k(f^*)}{T}}
$$

can be upper bounded by

$$
\frac{1}{T} \sum_{t=1}^{T} L(h, s_t, Z_t) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) - \alpha - \mathrm{disc}(\mathbf{s}^*) - M\sqrt{\frac{2 \log \mathcal{N}(k(f^*))}{T}}.
$$

This completes the proof of this result. $\qquad\square$

Theorem 4. *Under Assumption 1, for any $\delta > 0$, with probability at least $1 - \delta$, for all $\alpha > 0$, $\mathcal{L}_{T+1}(h_0, g_0|\mathbf{Z}_1^T) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) < r(\epsilon)$, where*

$$
\epsilon = 2\alpha + 2B_{k(f^*)} + M\sqrt{\frac{\log k(f^*)}{T}} + 2M\sqrt{\frac{\log \frac{2}{\delta}}{T}},
$$

where $s_t^* = g^*(X_t, s_{t-1}^*)$ and $k(f^*)$ is the smallest integer k such that $f^* \in \mathcal{H}_k \times \mathcal{G}_k$.

Proof. Observe that the following decomposition holds

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h_0,g_0|\mathbf{Z}_1^{t-1}) - \mathcal{L}_t(h^*,g^*|\mathbf{Z}_1^{t-1}) = \left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h_0,g_0|\mathbf{Z}_1^{t-1}) - \widetilde{F}(h_0,g_0,k_0)\right)$$
$$+ \left(\widetilde{F}(h_0,g_0,k_0) - \widetilde{F}(h^*,g^*,k^*)\right)$$
$$+ \left(\widetilde{F}(h^*,g^*,k^*) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h^*,g^*|\mathbf{Z}_1^{t-1})\right).$$

Since $\widetilde{F}(h_0,g_0,k_0) - \widetilde{F}(h^*,g^*,k^*) \leq 0$ by definition of $\widetilde{F}$ and (h_0,g_0,k_0), it follows from the union bound that

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h_0,g_0|\mathbf{Z}_1^{t-1}) - \mathcal{L}_t(h^*,g^*|\mathbf{Z}_1^{t-1}) - 2\alpha - 2B_{k(f^*)} - M\sqrt{\frac{\log k(f^*)}{T}} > \epsilon\right)$$

$$\leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h_0,g_0|\mathbf{Z}_1^{t-1}) - \widetilde{F}(h_0,g_0,k_0) - \alpha \geq \frac{\epsilon}{2}\right)$$

$$+ \mathbb{P}\left(\widetilde{F}(h^*,g^*,k^*) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h^*,g^*|\mathbf{Z}_1^{t-1}) - \alpha - 2B_{k(f^*)} - M\sqrt{\frac{\log k(f^*)}{T}} > \frac{\epsilon}{2}\right).$$

Since we can upper bound

$$\widetilde{F}(h^*,g^*,k^*) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(h^*,g^*|\mathbf{Z}_1^{t-1}) - \alpha - 2B_{k(f^*)} - M\sqrt{\frac{\log k(f^*)}{T}}$$

$$\leq -\Phi(h^*,g^*) - \alpha - M\sqrt{\frac{2\log\mathcal{N}(k(f^*))}{T}},$$

the second term is bounded by $2\exp\left(-\frac{2T\epsilon^2}{M^2}\right)$. We are using notation $\mathcal{N}(k) = \mathbb{E}_{\mathbf{v}\sim T(\mathbb{P})}[\mathcal{N}_1(\alpha,\mathcal{R}_k,\mathbf{v})]$ to simplify the presentation. To complete the proof, observe that the following chain of inequalities holds, by union bound and Lemma 1:

$$\mathbb{P}\left(\sup_{k\geq 1}\sup_{f\in\mathcal{F}_k}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(f|\mathbf{Z}_1^{t-1}) - \widetilde{F}(f,k) - 2\alpha \geq \epsilon\right)$$

$$\leq \sum_{k=1}^{\infty}\mathbb{P}\left(\sup_{f\in\mathcal{F}_k}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(f|\mathbf{Z}_1^{t-1}) - \widetilde{F}(f,k) - 2\alpha \geq \epsilon\right)$$

$$= \sum_{k=1}^{\infty}\mathbb{P}\left(\sup_{(h,g)\in\mathcal{F}}\Phi(h,g) \geq \epsilon + 2\alpha + M\sqrt{\frac{\log k}{T}} + M\sqrt{\frac{2\log\mathcal{N}(k)}{T}}\right)$$

$$\leq \sum_{k=1}^{\infty}\mathcal{N}(k)\exp\left(-\frac{1}{2M^2}T\left(\epsilon + M\sqrt{\frac{\log k}{T}} + M\sqrt{\frac{2\log\mathcal{N}(k)}{T}}\right)^2\right).$$

Arguments similar to the ones in the proof of Theorem 3 show that this term also bounded $2\exp\left(-\frac{2T\epsilon^2}{M^2}\right)$. Therefore, by Assumption 1 it follows that for any $\delta > 0$, with probability at least $1-\delta$, for all $\alpha > 0$, $\mathcal{L}_{T+1}(h_0,g_0|\mathbf{Z}_1^T) - \mathcal{L}_{T+1}(f^*|\mathbf{Z}_1^T) < r(\epsilon)$ and this completes the proof. $\qquad\square$

B Complexities of DSSMs

In this Section, we prove upper bounds on the expected sequential covering numbers of some commonly used hypothesis sets. We start with a following lemma that gives a general upper bound in terms of the sequential Rademacher complexity:

$$\mathfrak{R}_T^{\text{seq}}(L\circ\mathcal{H}) = \sup_{\mathbf{z},\mathbf{s}}\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{T}\sum_{t=1}^{T}\sigma_t L(h,\mathbf{s}_t(\boldsymbol{\sigma}),z_t(\boldsymbol{\sigma}))\right)\right], \tag{14}$$
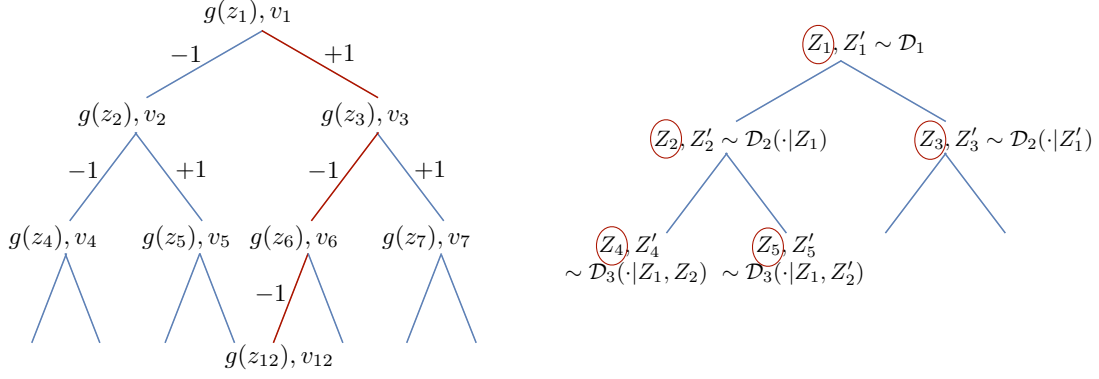
Table 1: Weighted sequential cover and sequential covering numbers.

where the supremum is taken over all pairs of complete $\mathcal{Z}$-valued binary tree $\mathbf{z}$ and $\mathcal{S}$-valued binary tree $\mathbf{s}$.

Lemma 3. *The following bound holds:*

$$\sup_{\alpha>0} \frac{\alpha}{2} \sqrt{\log \mathcal{N}_2(2\alpha, \mathcal{R})} \leq 3\sqrt{\frac{\pi}{2} \log T}\, \mathfrak{R}_T^{seq}(L \circ \mathcal{H}),$$

whenever $\mathcal{N}_2(2\alpha, \mathcal{R}) < \infty$.

Proof. As in the previous proofs, we use the convention that for any sequence of $z_1, \ldots, z_T$, we can write sequence $L(h, s_1, z_1), \ldots, L(h, s_T, z_T)$, (where $\mathbf{s}$ is generated by g) as a sequence $f_1(z_1), \ldots, f_T(z_T)$. We consider the following Gaussian-Rademacher sequential complexity:

$$\mathfrak{G}_T^{\text{seq}}(\mathcal{F}, \mathbf{z}) = \mathop{\mathbb{E}}_{\gamma, \sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^{T} \sigma_t \gamma_t f_t(z_t(\boldsymbol{\sigma})) \right) \right], \tag{15}$$

where $\boldsymbol{\sigma}$ is an independent sequence of Rademacher random variables, $\boldsymbol{\gamma}$ is an independent sequence of standard Gaussian random variables and $\mathbf{z}$ is a complete binary tree of depth T with values in $\mathcal{Z}$.

Observe that if V is any α-cover with respect to the ℓ_2-norm of $\mathcal{F}$ on $\mathbf{z}$, then, the following holds by independence of $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$:

$$\mathfrak{G}^{\text{seq}}(\mathcal{F}, \mathbf{z}) \geq \mathop{\mathbb{E}}_{\gamma} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\mathbf{v} \in V} \left(\frac{1}{T} \sum_{t=1}^{T} \sigma_t \gamma_t \mathbf{v}_t(\boldsymbol{\sigma}) \right) \right] = \mathop{\mathbb{E}}_{\sigma} \mathop{\mathbb{E}}_{\gamma} \left[\sup_{\mathbf{v} \in V} \left(\frac{1}{T} \sum_{t=1}^{T} \sigma_t \gamma_t \mathbf{v}_t(\boldsymbol{\sigma}) \right) \right].$$

Notice that V is also a 2α-cover with respect to the ℓ_2-norm of $\mathcal{F}$ on $\mathbf{z}$. We can obtain a smaller 2α-cover V_0 from V by eliminating $\mathbf{v}$s that are α-close to some other $\mathbf{v}' \in V$. Since V is finite, let $V = \{\mathbf{v}^1, \ldots, \mathbf{v}^{|V|}\}$, and for each $\mathbf{v}^i$ we delete $\mathbf{v}^j \in \{\mathbf{v}_{i+1}, \ldots, \mathbf{v}_{|V|}\}$ such that the following inequality holds:

$$\left[\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{v}_t^i(\boldsymbol{\sigma}) - \mathbf{v}_t^j(\boldsymbol{\sigma}) \right)^2 \right]^{1/2} \leq \alpha.$$

It is straightforward to verify that V_0 is a 2α-cover with respect to the ℓ_2-norm of $\mathcal{F}$ on $\mathbf{z}$. Furthermore, it follows that for a fixed $\boldsymbol{\sigma}$, the following holds:

$$\mathop{\mathbb{E}}_{\gamma} \left[\left(\frac{1}{T} \sum_{t=1}^{T} \sigma_t \gamma_t \mathbf{v}_t(\boldsymbol{\sigma}) - \sum_{t=1}^{T} \sigma_t \gamma_t \mathbf{v}'_t(\boldsymbol{\sigma}) \right)^2 \right] \geq \alpha^2$$

15

for any $\mathbf{v}', \mathbf{v} \in V_0$. Let $Z_i, i = 1, \ldots, |V_0|$ be a sequence of independent Gaussian random variables with $\mathbb{E}[Z_i] = 0$ and $\mathbb{E}[Z_i^2] = \alpha^2/2$. Observe that $\mathbb{E}[(Z_i - Z_j)] = \alpha^2$ and hence by the Sudakov-Fernique inequality it follows that

$$\mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\boldsymbol{\gamma}} \left[\sup_{\mathbf{v} \in V} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t \gamma_t \mathbf{v}_t(\boldsymbol{\sigma}) \right) \right] \geq \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\boldsymbol{\gamma}} \left[\sup_{\mathbf{v} \in V_0} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t \gamma_t \mathbf{v}_t(\boldsymbol{\sigma}) \right) \right]$$
$$\geq \mathbb{E} \left[\max_{i=1,\ldots,|V_0|} Z_i \right]$$
$$\geq \frac{\alpha}{2} \sqrt{\log |V_0|},$$

where the last inequality is the standard result for Gaussian random variables. Therefore, we conclude that $\mathfrak{G}^{\text{seq}}(\mathcal{F}, \mathbf{z}) \geq \sup_{\alpha > 0} \frac{\alpha}{2} \sqrt{\log \mathcal{N}_2(2\alpha, \mathcal{F}, \mathbf{z})}$. On the other hand, using standard properties of Gaussian complexity Ledoux and Talagrand [1991], we have

$$\mathfrak{G}_T^{\text{seq}}(\mathcal{F}, \mathbf{z}) \leq 3\sqrt{\frac{\pi}{2} \log T} \, \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t \epsilon_t f_t(z_t(\boldsymbol{\sigma})) \right) \right],$$

where $\boldsymbol{\epsilon}$ is an independent sequence of Rademacher variables. We re-arrange $\mathbf{z}$ into $\mathbf{z}^{\boldsymbol{\epsilon}}$ so that $z_t(\boldsymbol{\sigma}) = z_t^{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}\boldsymbol{\sigma})$ for all $\boldsymbol{\sigma} \in \{\pm 1\}^T$ and it follows that

$$\mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t \epsilon_t f_t(z_t(\boldsymbol{\sigma})) \right) \right] = \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t \epsilon_t f_t(z_t^{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}\boldsymbol{\sigma})) \right) \right]$$
$$\leq \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t \epsilon_t f_t(z_t(\boldsymbol{\epsilon}\boldsymbol{\sigma})) \right) \right] \right]$$
$$= \sup_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t f_t(z_t(\boldsymbol{\sigma})) \right) \right]$$
$$\leq \sup_{\mathbf{z}, \mathbf{s}} \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{T} \sum_{t=1}^T \sigma_t L(h, \mathbf{s}_t(\boldsymbol{\sigma}), z_t(\boldsymbol{\sigma})) \right) \right].$$

Therefore, the following inequality holds

$$\sup_{\alpha > 0} \frac{\alpha}{2} \sqrt{\log \mathcal{N}_2(2\alpha, \mathcal{F}, \mathbf{z})} \leq 3\sqrt{\frac{\pi}{2} \log T} \mathfrak{R}_T^{\text{seq}}(L \circ \mathcal{H}),$$

and the conclusion of the theorem follows by taking the supremum with respect to $\mathbf{z}$ on both sides of this inequality. $\square$

The following lemma decomposes the complexity of structural time series models into the complexities of its subcomponents.

Lemma 4. *Let $p \geq 1$ and $L \circ \mathcal{H} = \{(\mathbf{x}, \mathbf{s}, y) \to (\mathbf{w} \cdot \Psi(\mathbf{x}) + \mathbf{w}' \cdot \mathbf{s} - y)^p \colon \mathbf{w} \in S, \mathbf{w}' \in S'\}$ for some S, S'. Assume that the condition $|\mathbf{w} \cdot \mathbf{x} - y| \leq M$ holds for all $(\mathbf{x}, y) \in \mathcal{Z}$ and all $\mathbf{w}$ such that $\|\mathbf{w}\|_{\mathcal{H}} \leq \Lambda$. Then, the following inequalities hold:*

$$\mathfrak{R}_T^{\text{seq}}(\mathcal{F}) \leq pM^{p-1}C_T(\mathfrak{R}_T^{\text{seq}}(H_1) + \mathfrak{R}_T^{\text{seq}}(H_1)), \tag{16}$$

where $H_1 = \{\mathbf{x} \to \mathbf{w} \cdot \Psi(\mathbf{x}) \colon \mathbf{w} \in S\}$, $H_2 = \{\mathbf{s} \to \mathbf{w}' \cdot \mathbf{s} \colon \mathbf{w}' \in S'\}$ and $C_T = 8(1 + 4\sqrt{2}\log^{3/2}(eT^2))$.

Proof. Since $x \to |x|^p$ is pM^{p-1}-Lipschitz over $[-M, M]$, by Lemma 13 in [Rakhlin et al., 2015a], the following bound holds:

$$\mathfrak{R}_T^{\text{seq}}(L \circ \mathcal{F}) \leq pM^{p-1}C_T\mathfrak{R}_T^{\text{seq}}(H'),$$

where $H' = \{(\mathbf{x}, \mathbf{s}, y) \to \mathbf{w} \cdot \Psi(\mathbf{x}) + \mathbf{w}' \cdot \mathbf{s} - y : \mathbf{w} \in S, \mathbf{w} \in S'\}$. Note that Lemma 13 requires that $\mathfrak{R}_T^{\text{seq}}(H') > 1/T$ which is guaranteed by Khintchine's inequality. By definition of the sequential Rademacher complexity

$$
\begin{aligned}
\mathfrak{R}_T^{\text{seq}}(H') &= \sup_{(\mathbf{x},y)} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}} \frac{1}{T}\sum_{t=1}^T \sigma_t(\mathbf{w}\cdot\Psi(\mathbf{x}_t(\boldsymbol{\sigma})) - y(\boldsymbol{\sigma}))\right] \\
&= \sup_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}\in S} \frac{1}{T}\sum_{t=1}^T \sigma_t \mathbf{w}\cdot\Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right] + \sup_{\mathbf{s}} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}'\in S'} \frac{1}{T}\sum_{t=1}^T \sigma_t \mathbf{w}'\cdot\mathbf{s}_t(\boldsymbol{\sigma})\right] \\
&\quad + \sup_{y} \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{T}\sum_{t=1}^T \sigma_t y(\boldsymbol{\sigma})\right] = \mathfrak{R}_T^{\text{seq}}(H_1) + \mathfrak{R}_T^{\text{seq}}(H_2),
\end{aligned}
$$

where, for the last equality, we used the fact that σ_ts are mean zero random variables and σ_t is independent of $y(\boldsymbol{\sigma}) = y(\sigma_1, \sigma_2, \ldots, \sigma_{t-1})$. This completes the proof. $\square$

Observe that, for example, if $H_1 = \{\mathbf{x} \to \mathbf{w}\cdot\Psi(\mathbf{x}): \|\mathbf{w}\|_H \le \Lambda\}$ where H is a Hilbert space with a corresponding feature map $\Psi\colon \mathcal{X} \to \mathcal{H}$ and PDS kernel K, then, by Lemma 6 in [Kuznetsov and Mohri, 2015], we have the inequality $\mathfrak{R}_T^{\text{seq}}(H_1) \le \frac{\Lambda r}{\sqrt{T}}$, where $r = \sup_x K(x,x)$. Similarly, if $H_2 = \{\mathbf{s} \to \mathbf{w}\cdot\mathbf{s}: \|\mathbf{w}\|_2 \le \Lambda_0\}$ then we have $\mathfrak{R}_T^{\text{seq}}(H_2) \le \frac{\Lambda_0 r_0}{\sqrt{T}}$, with $r_0 = \sup_{\mathbf{s}} \|\mathbf{s}\|_2$. The next result gives a bound on the sequential complexity of a weighted combination of N binary-valued functions $\mathcal{H}_1$. A common choice of such binary-valued functions are decision trees. Note that result is logarithmic in the number of such functions, which suggests using a large set of functions.

Lemma 5. *Let* $H = \{\mathbf{x} \to (\mathbf{w}\cdot\Psi(\mathbf{x}): \|\mathbf{w}\|_1 \le \Lambda\}$ *where, for each* $j \in [1, N]$, Ψ_j *is a binary-valued function. Then, the following inequalities hold:*

$$
\mathfrak{R}_T^{seq}(H) \le \sqrt{\frac{2\Lambda^2 \log 2N}{T}}. \tag{17}
$$

Proof. We observe that, by the definition of the dual now, the following equalities hold:

$$
\begin{aligned}
\sup_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_1 \le \Lambda} \sum_{t=1}^T \sigma_t \mathbf{w}\cdot\Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right] &= \Lambda \sup_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{t=1}^T \sigma_t \Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right\|_\infty \\
&= \Lambda \sup_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{j\in[1,N], s\in\{\pm 1\}} \sum_{t=1}^T \sigma_t \Psi_j(\mathbf{x}_t(\boldsymbol{\sigma}))\right].
\end{aligned}
$$

To bound the last quantity, we apply an argument similar to the one used in the proof of Massart's lemma. The key difference is that here $\mathbf{x}_t$ depends on σ, which requires a more careful analysis. Observe that, by the monotonicity of $\exp$, the following upper bound holds, for any $u > 0$:

$$
\exp\left(u\, \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{j\in[1,N], s\in\{\pm 1\}} \sum_{t=1}^T \sigma_t \Psi_j(\mathbf{x}_t(\boldsymbol{\sigma}))\right]\right) \le 2 \sum_{j=1}^N \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(u\sum_{t=1}^T \sigma_t \Psi_j(\mathbf{x}_t(\boldsymbol{\sigma}))\right)\right].
$$

Next, since σ_T and $\mathbf{x}_T(\boldsymbol{\sigma}) = \mathbf{x}_T(\sigma_1, \ldots, \sigma_{T-1})$ are independent, by Hoeffding's bound, we can write

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(u\sum_{t=1}^T \sigma_t \Psi_j(\mathbf{x}_t(\boldsymbol{\sigma}))\right)\right] &= \mathbb{E}_{\boldsymbol{\sigma}_1^{T-1}}\left[\exp\left(u\sum_{t=1}^{T-1} \sigma_t \Psi_j(\mathbf{x}_t(\boldsymbol{\sigma}))\right)\mathbb{E}_{\boldsymbol{\sigma}_1^{T-1}}\left[\exp(u\sigma_T \Psi_j(\mathbf{x}_T(\boldsymbol{\sigma})))\right]\right] \\
&\le e^{2u} \mathbb{E}_{\boldsymbol{\sigma}_1^{T-1}}\left[\exp\left(u\sum_{t=1}^{T-1} \sigma_t \Psi_j(\mathbf{x}_t(\boldsymbol{\sigma}))\right)\right].
\end{aligned}
$$

Iterating this result over t and optimizing over u yields the desired upper bound. $\square$

C State Discrepancy Estimation

We can decompose the discrepancy as follows:

$$\text{disc}(\mathbf{s}) \leq \sup_{h \in \mathcal{H}} \left(\frac{1}{\tau} \sum_{t=T-\tau+1}^{T} \mathcal{L}_t(h, g | \mathbf{Z}_1^{t-1}) - \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(h, g | \mathbf{Z}_1^{t-1}) \right) \tag{18}$$

$$+ \sup_{h \in \mathcal{H}} \left(\mathcal{L}_{T+1}(h, g | \mathbf{Z}_1^T) - \frac{1}{\tau} \sum_{t=T-\tau+1}^{T} \mathcal{L}_t(h, g | \mathbf{Z}_1^{t-1}) \right).$$

We will assume that the second term, which we will denote by disc_τ, is sufficiently small and will show that the first term can be estimated from data. Note that, in general, the requirement that disc_τ is small is necessary for learning Barve and Long [1996].

The following result shows that we can estimate the first term appearing in the upper bound (18) on $\text{disc}(\mathbf{s})$.

Theorem 5. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\alpha > 0$:*

$$\text{disc}(\mathbf{s}) - \widehat{\text{disc}}(\mathbf{s}) \leq 2\alpha + M \|\mathbf{u} - \mathbf{u}_\tau\|_2 \sqrt{2 \log \frac{\mathbb{E}_{\mathbf{z} \sim T(\mathbf{p})}[\mathcal{N}_1(\alpha, \mathcal{R}, \mathbf{z})]}{\delta}},$$

where $\mathbf{u}$ is the uniform distribution over the sample, $\mathbf{u}_\tau$ a uniform distribution over the last s points, and

$$\widehat{\text{disc}}(\mathbf{s}) = \sup_{h \in \mathcal{H}} \left(\frac{1}{\tau} \sum_{t=T-\tau+1}^{T} L(h, s_t, Z_t) - \frac{1}{T} \sum_{t=1}^{T} L(h, s_t, Z_t) \right). \tag{19}$$

Proof. We upper bound the difference of suprema by the supremum of the difference and then apply the same arguments as in the proof of Lemma 1, with the only difference that the examples are here weighted by $(1/T - 1/\tau)$. $\qquad\square$

D Experiments

In this section, we present the results of experiments with the BOOSTSM algorithm.

As subcomponents of our BOOSTSM algorithm, we used a linear trend model and an ensemble of decision trees as described in Section 5. As a baseline comparator, we used a regular STSM model with subcomponents that consists of a trend model and a linear model. Note that, since our comparator can only choose a single trend model, to make the comparison fair, we restricted our algorithm to choose a single trend model as well.

For our experiments we used web traffic data from Wikipedia articles. Each of these time series represents a number of daily views (integer) of a different Wikipedia article starting from January, 1st, 2012 up to March, 1st, 2017. This data is obtained via public pageviews API `https://wikitech.wikimedia.org/wiki/Analytics/PageviewAPI`. In our experiments, we used time series for nine articles that appear among top twenty articles in 2016 (in terms of number of total views): `Bernie Sanders`, `Batman vs Superman`, `Brownian motion`, `Donald Trump`, `Java`, `Isaac Newton`, `Janet Jackson`, `Merle Haggard`, `Main Page`, `Ted Cruz`.

The following experimental setup was used in our experiments. Each algorithm was trained on the first 1500 days, then used to predict the next 25 days. The average error was recorded and then the algorithm was retrained with 1525 days and used to predict the next 25 and so on. We report the average error across all these rounds in Table 2. We also report the running errors in Figure 3.

Since the time series used in our experiments have vastly different scales, both within time series and between different time series, RMSE is not an appropriate evaluation metric. Thus, for ease of comparison, we chose Symmetric Absolute Percentage Error (SMAPE) as our evaluation metric. This loss function is defined by $L(x, y) = |x - y| / (|x| + |y|)$ if at least one of x, y is not zero and $L(x, y) = 0$ otherwise. Note, however, that both algorithms were trained using the squared loss, as discussed in the previous sections. SMAPE is not a convex loss function and is generally hard to

Table 2: Mean SMAPE for BOOSTSMand STSM.

	BOOSTSM	STSM
Batman vs Superman	**0.6060 ± 0.0958**	0.6489 ± 0.1056
Bernie Sanders	0.7247 ± 0.0805	**0.7118 ± 0.0829**
Brownian motion	**0.4226 ± 0.0499**	0.6078 ± 0.1118
Donald Trump	**0.7124 ± 0.0727**	0.7304 ± 0.0603
Java	**0.5800 ± 0.1497**	0.7982 ± 0.0451
Isaac Newton	**0.4896 ± 0.0911**	0.5988 ± 0.0823
Janet Jackson	**0.7003 ± 0.0671**	0.7150 ± 0.0565
Merle Haggard	0.7400 ± 0.0703	**0.7205 ± 0.0728**
Main Page	**0.5639 ± 0.1041**	0.5809 ± 0.0958
Ted Cruz	**0.6910 ± 0.0972**	0.7272 ± 0.1067

optimize, which is why we used MSE as our optimization objective. The ranking of the models in our experiments in terms of RMSE is the same as the obtained using SMAPE.

In our experiments, BOOSTSM outperformed STSM on eight out of ten datasets, in some cases by a substantial margin. This suggests that BOOSTSM may be able to better adapt to different types of non-stationarities present in the data.
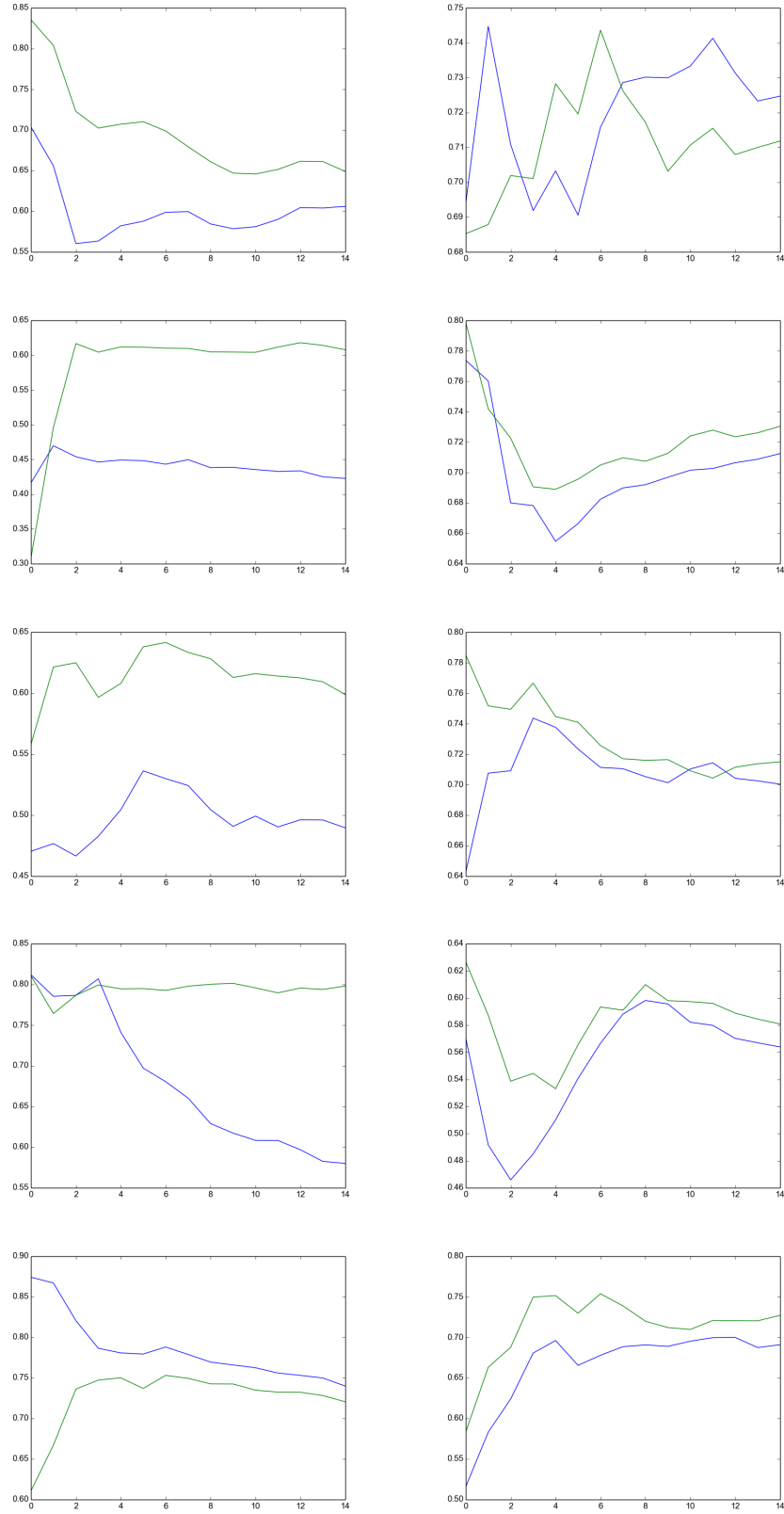
Table 3: Running SMAPE for BOOSTSM (blue) and STSM (green). Datasets appear in alphabetic order: left to right, top to bottom.