HUBER ADDITIVE MODELS FOR NON-STATIONARY TIME SERIES ANALYSIS

Yingjie Wang[1,2], **Xianrui Zhong**[3], **Fengxiang He**[2,*], **Hong Chen**[4,*] **& Dacheng Tao**[2]
[1]College of Informatics, Huazhong Agricultural University, China
[2]JD Explore Academy, JD.com Inc, China
[3]Department of Computer Science, University of Illinoist at Urbana-Champaign, USA
[4]College of Science, Huazhong Agricultural University, China
yingjiewang1201@gmail.com, xzhong23@illinois.edu,
fengxiang.f.he@gmail.com, chenh@mail.hzau.edu.cn,
dacheng.tao@gmail.com

ABSTRACT

Sparse additive models have shown promising flexibility and interpretability in processing time series data. However, existing methods usually assume the time series data to be stationary and the innovation is sampled from a Gaussian distribution. Both assumptions are too stringent for heavy-tailed and non-stationary time series data that frequently arise in practice, such as finance and medical fields. To address these problems, we propose an adaptive sparse Huber additive model for robust forecasting in both non-Gaussian data and (non)stationary data. In theory, the generalization bounds of our estimator are established for both stationary and non-stationary time series data, which are independent of the widely used mixing conditions in learning theory of dependent observations. Moreover, the error bound for non-stationary time series contains a discrepancy measure for the shifts of the data distributions over time. Such a discrepancy measure can be estimated empirically and used as a penalty in our method. Experimental results on both synthetic and real-world benchmark datasets validate the effectiveness of the proposed method. The code is available at https://github.com/xianruizhong/SpHAM.

1 INTRODUCTION

Additive model has become one of the most powerful tools for time series analysis due to the exemplary monograph (Stone, 1985; Hastie & Tibshirani, 1990) and companion software (Chambers & Hastie, 1992). For the past two decades, the growing importance of algorithmic flexibility and interpretability motivates the development of various additive models along with theoretical explorations (Huang & Yang, 2004; Wang & Yang, 2007; Chu & Glymour, 2008; Song & Yang, 2010; Yang et al., 2018; Chen et al., 2017; Liu et al., 2020; Chen et al., 2021a) and practical applications (Dominici et al., 2002; Wang & Brown, 2011; Ravindra et al., 2019; Bussmann et al., 2020; Wang et al., 2020). Although these aforementioned works have shown promising behaviours, the proposed methods in these works require some stringent assumptions on the stochastic process, e.g., various mixing conditions (Doukhan, 1994), stationary distribution and Gaussian innovation.

A number of attempts have been made to relax such stringent assumptions. For the purpose of dealing with non-Gaussian innovation, Qiu et al. (2015) develop an elliptical vector autoregressive model for estimating heavy-tailed stationary processes with parametric convergence analysis that reduces the influence of non-Gaussian innovation. Moreover, under stationarity and decaying β-mixing condition, Wong et al. (2020) derive nonasymptotic estimation error of lasso without assuming special parametric form of the data generating process. Stationarity and various mixing conditions are commonly adopted in many previous studies, see, e.g., (Mohri & Rostamizadeh, 2009; 2010; Kock & Callot, 2015; Wong et al., 2020). In practice, the mixing and stationary conditions are too stringent and not always valid (Baillie, 1996; Kuznetsov & Mohri, 2020a). To relax the mixing and stationary conditions, Adams & Nobel (2010) prove asymptotic guarantees for stationary ergodic

*Corresponding authors.

sequences. Agarwal & Duchi (2013) establish generalization bounds for asymptotically stationary (mixing) processes in the case of stable on-line learning algorithms. Kuznetsov & Mohri (2014) establish learning guarantees for fully non-stationary and mixing processes. Recently, Kuznetsov & Mohri (2020a) provide data-dependent generalization risk bounds for non-stationary and non-mixing stochastic processes.

However, the exploration of additive models for both robust and non-stationary time series forecasting is still very limited. In this paper, we propose a class of sparse Huber additive models with theoretical guarantees. Our main contributions are summarized as follows:

- *Algorithm design and theoretical guarantees*: In Section 3.1, we first propose a novel sparse Huber additive model (SpHAM) with Huber loss, sparsity-inducing $\ell_{2,1}$-norm regularizer and additive data-dependent hypothesis space. This proposed method works for stationary time series and can achieve robust forecasting and satisfactory inference (e.g., Granger causal discovery) simultaneously. In theory, Section 3.2 establishes the upper bound of the function approximation error of SpHAM by developing error decomposition technique (Wu et al., 2006; Chen et al., 2021b) and employing sequential Rademacher complexity (Rakhlin et al., 2010; 2015; Kuznetsov & Mohri, 2020a). With properly selected scale parameters, the theoretical findings indicate that: a) For stationary time series, the function approximation consistency with convergence rate $O(n^{-\frac{1}{2}})$ can be pursued, even if the innovation is non-Gaussian distribution (see Theorem 1 and Corollary 1 for more details). Moreover, this consistency analysis appears to be novel because no explicit data dependence assumptions are not imposed here, e.g., various mixing conditions used in (Doukhan, 1994; Mohri & Rostamizadeh, 2010; Zou et al., 2009; Wong et al., 2020); b) For non-stationary time series, the function approximation error is bounded by a discrepancy measure, which characterizes the drifts of the data distributions along with time (see Theorem 2 for more details). By penalizing such a discrepancy measure, Section 3.3 further proposes an adaptive SpHAM for non-stationary time series and provides its theoretical upper bound correspondingly (see Theorem 3 for more details).
- *Optimization and empirical evaluations*: The proposed SpHAM and adaptive SpHAM can be implemented efficiently by Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009). Experimental results on both synthetic and real-world benchmark CauseMe (Runge et al., 2019) validate the effectiveness of the proposed method.

Related works: There exist some interesting studies towards sparse additive model for time series analysis from both theoretical and practical viewpoints; see, e.g., Chu & Glymour (2008); Song & Yang (2010); Yang et al. (2018). Although they show promising interpretability and modeling capacity, all of them require Gaussian innovation, stationarity and various mixing conditions. Moreover, these assumptions are not always valid in heavy-tailed and non-stationary time series data. In contrast to these previous additive models, we are seeking to formulate our method and investigate its asymptotic properties without resorting to such strict assumptions.

Robust forecasting is a vibrant area of research in time series analysis (Qiu et al., 2015; Wong et al., 2020). As one of the triumphs and milestones of robust statistics, the success stories of Huber algorithms are mainly on robust prediction tasks with i.i.d datasets (Huber, 1964; Huber & Ronchetti, 2009; Loh, 2017; Feng & Wu, 2020), and their extensions to time series prediction are fairly sparse. To our best knowledge, this is the first work that considers Huber additive models for non-stationary and dependent time series data.

Our theoretical analyses are inspired from the successful usage of sequential Rademacher complexity in Kuznetsov & Mohri (2020b;a). The generalization risk bounds are established in their works for a general scenario of non-stationary and non-mixing stochastic processes. However, it should be mentioned that we are chiefly concerned about the function approximation analysis, which is essentially different from theirs and is very crucial for Huber algorithms, since the convergence of generalization risk cannot imply the convergence of function approximation (Sun et al., 2019; Feng & Wu, 2020; He & Tao, 2020). Moreover, the development of analysis techniques (e.g., error decomposition and sequential Rademacher complexity) for assessing our method may shed light on other robust models for nonstationary time series analysis.

To better highlight the novelty of our method, Table 1 summarizes the algorithmic properties of our method and other related works, e.g., Sparse additive models for time series (TS-SpAM) (Yang et al.,

Table 1: The properties of related methods

	TS-SpAM	DBF	R-Dantzig	Ours
Hypothesis space	Additive Spline-based	Kernel-based	Linear	Additive Kernel-based
Loss	Squared	Squared	Quantile-based	Huber
Mixing condition	Yes	No	Yes	No
Stationarity	Yes	No	Yes	No
Robustness	No	No	Yes	Yes
Sparsity	Yes	No	Yes	Yes

2018), Discrepancy-based Forecasting (DBF) (Kuznetsov & Mohri, 2020a) and Robust Dantzig-selector-type estimator (R-Dantzig) (Qiu et al., 2015).

The remainder of this paper is organized as follows. Section 2 recalls the background of additive models. Section 3 mainly provides our methods and theoretical guarantees. We provide empirical evaluations in Section 4. Finally, Section 5 concludes this paper. The source code package is available at `https://github.com/xianruizhong/SpHAM`.

2 Preliminary

Let $\{Z^t\}_{t=-\infty}^{\infty}$ be a stochastic time series with time index t, where variable $Z^t = (X^t, Y^t)$ takes values in the compact input space $\mathcal{X} \subset \mathbb{R}^p$ and the output space $\mathcal{Y} \subset \mathbb{R}$. We consider a common nonparametric model

$$Y^t = f^*(X^t) + \varepsilon^t, \;\; \mathbb{E}(\varepsilon^t) = 0, \tag{1}$$

where $f^*(\cdot)$ is the ground truth function, and the innovation ε^t is i.i.d. across time $t \in \mathbb{Z}$. For the sake of simplicity, we denote ρ^t and $\rho^t_{\mathcal{X}}$ as the jointed distribution of (X^t, Y^t) and the corresponding marginal distribution with respect to X^t, respectively. This setup actually covers a large number of scenarios commonly used in practice. For instance, the case X^t that contains p lagged values of Y^t (e.g., $X^t = Y^{(t-p)} \times \cdots \times Y^{(t-1)}$) corresponds to the p-order autoregressive models. Moreover, this case can be viewed as a vector autoregressive model in the sense that input X^t includes the historical information of multiple variables.

Although such a nonparametric model (1) makes very few assumptions on data generation, the related nonparametric algorithms suffer so-called "curse of dimensionality", see Fan & Gijbels (1996) for further discussion. An effective strategy for solving this problem is additive model. Usually, the additive structure is obtained by decomposing the input space $\mathcal{X} \in \mathbb{R}^p$ into $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_p$. Under the assumption that the ground truth admits an additive structure $f^* = \sum_{j=1}^p f_j^*$, the additive model can be defined by

$$Y^t = f_1^*(X_1^1) + \cdots + f_p^*(X_p^t) + \varepsilon^t, \tag{2}$$

where each component $f_j^* : \mathcal{X}_j \to \mathbb{R}$ is a smooth function. In linear time series analysis, a weak stationarity condition (i.e., the first two moments of time series are time invariant) is preferred (Han et al., 2015; Qiu et al., 2015). In contrast, strict stationarity is primarily used to analyse the nonlinear time series (Fan & Yao, 2005).

Definition 1. *A stochastic process $\{Z^t\}_{t=-\infty}^{\infty}$ is strictly stationary if $(Z^1, ..., Z^t)$ and $(Z^{1+k}, ..., Z^{t+k})$ have the same joint distributions for any $t \in \mathbb{Z}$ and $k \in \mathbb{Z}$.*

Note that, if not otherwise stated, the stationarity in this paper refers to strict stationarity. Suppose that we are given T size time series data $\{(x^t, y^t)\}_{t=1}^T \in \mathcal{Z}^T$ which are drawn from an additive data-generating model (2). Under stationarity condition and zero-mean Gaussian innovation with finite variance, the widely used methods that learn the ground truth f^* usually integrate squared loss and a smoothness- or sparsity-inducing regularizer $\Omega(\cdot)$ into a structural risk minimization scheme:

$$\min_{f \in \mathcal{H}} \sum_{t=1}^{T} (y^t - \sum_{j=1}^{p} f_j(x_j^t))^2 + \Omega(f),$$

where $\mathcal{H} := \{f_1 + ... + f_p : f_j \in \mathcal{H}_j, j = 1, ..., p\}$ is an additive hypothesis space. Commonly, each subspace $\mathcal{H}_j$ could be reproducing kernel Hilbert space (Chen et al., 2017; Kandasamy & Yu, 2016; Raskutti et al., 2012), the orthogonal basis inducing space (Ravikumar et al., 2009; Meier et al., 2009; Yang et al., 2018), or the composite function space with the neural network as a typical (Agarwal et al., 2020; Bussmann et al., 2020).

However, when facing heavy-tailed innovation, these methods may have degraded performance due to the amplification of the squared loss to large residuals. As one commonly used statistic in robust learning community (Peng et al., 2019; Wang et al., 2017b; Feng & Wu, 2020), Huber loss is defined as

$$\ell_\sigma(f(x^t) - y^t) = \begin{cases} (f(x^t) - y^t)^2, & \text{if } |f(x^t) - y^t| < \sigma \\ 2\sigma|f(x^t) - y^t| - \sigma^2, & \text{if } |f(x^t) - y^t| \geq \sigma, \end{cases} \tag{3}$$

where σ is a positive hyper-parameter. Note that in the previous studies (Huber & Ronchetti, 2009; Loh, 2017), the hyper-parameter σ is set to be fixed according to the 95% asymptotic efficiency rule. However, Huber regression with a fixed scale parameter may not be able to learn the ground truth when the noise is asymmetric, as argued recently in Feng & Wu (2020); Sun et al. (2019). In this paper, we choose the scale parameter σ by relating it to the moment condition of the noise distribution and the sample size so that the resulting regression estimator can asymptotically converge to the ground truth function.

3 METHOD

3.1 SPARSE HUBER ADDITIVE MODELS

In this paper, we choose reproducing kernel Hilbert space (RKHS) $\mathcal{H}_{K_j}, j = 1, .., p$, to form the the additive hypothesis space $\mathcal{H}$, where each $\mathcal{H}_{K_j}$ is associated with a symmetric and positive semi-definite Mercer kernel $K_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$. An additive RKHS is defined as

$$\mathcal{H}_K = \{f_1 + ... + f_p : f_j \in \mathcal{H}_{K_j}, j = 1, ..., p\} \tag{4}$$

with kernel norm $\|f\|_K^2 = \inf\{\|f\|_{K_1}^2 + ... + \|f\|_{K_p}^2\}$. By integrating the Huber loss (3), additive RKHS and kernel norm inducing regularizer into a Tikhonov regularization scheme, the regularized Huber additive model with kernel-norm can be formulated as

$$\hat{f}_\eta = \sum_{j=1}^p \hat{f}_{\eta,j} = \underset{f=\sum_{j=1}^p f_j, f_j \in \mathcal{H}_{K_j}}{\arg\min} \{\sum_{t=1}^T \ell_\sigma(y^t - \sum_{j=1}^p f_j(x_j^t)) + \eta \sum_{j=1}^p \tau_j \|f_j\|_{K_j}^2\}, \tag{5}$$

where η is positive regularization parameter and τ_j is the weight for j-th kernel norm. The representer theorem (Wahba, 1990) ensures that $\hat{f}_\eta$ can be represented as

$$\hat{f}_\eta = \sum_{j=1}^p \sum_{t=1}^T \alpha_{tj}^\eta K_j(x_j^t, \cdot),\ \alpha_{tj}^\eta \in \mathbb{R}.$$

To offer the method with sparsity, we consider the following sparsity-inducing penalty

$$\Omega(f) := \inf\{\sum_{j=1}^p \tau_j \|\alpha_j\|_2 : f = \sum_{j=1}^p \sum_{t=1}^T \alpha_{tj} K_j(x_j^t, \cdot)\}.$$

Let $\mathcal{H}_Z$ be an additive data dependent hypothesis space defined by

$$\mathcal{H}_Z = \{f = \sum_{j=1}^p \sum_{t=1}^T \alpha_{tj} K_j(x_j^t, \cdot) : \alpha_{tj} \in \mathbb{R}\}.$$

Then the SpHAM can be formulated as

$$\hat{f} = \underset{f \in \mathcal{H}_Z}{\arg\min} \{\frac{1}{T} \sum_{t=1}^T \ell_\sigma(y^t - \sum_{j=1}^p f_j(x_j^t)) + \lambda \Omega(f)\}.$$

Denote $\mathbf{K}_j^t = (K_j(x_j^1, x_j^t), ..., K_j(x_j^T, x_j^t))' \in \mathbb{R}^T$, $\alpha = (\alpha_1', ..., \alpha_p')' \in \mathbb{R}^{Tp}$ and $\alpha_j = (\alpha_{1j}, ..., \alpha_{Tj})' \in \mathbb{R}^T$, where α_j' here refers to the transpose of α_j for avoiding the confusion. The SpHAM can be represented as

$$\hat{f} = \sum_{j=1}^{p} \sum_{t=1}^{T} \alpha_{tj}^{\lambda} K_j(x_j^t, \cdot) \tag{6}$$

with

$$\alpha^{\lambda} = \underset{\alpha_j \in \mathbb{R}^T, j=1,...,p}{\arg\min} \{\frac{1}{T} \sum_{t=1}^{T} \ell_{\sigma}(y^t - \sum_{j=1}^{p} (\mathbf{K}_j^t)' \alpha_j) + \lambda \sum_{j=1}^{p} \tau_j \|\alpha_j\|_2\}. \tag{7}$$

The optimization problem (7) can be solved by Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009). We provide the detailed optimization procedure in Appendix E.

Remark 1. *Inspired by the studies on group additive models (Yin et al., 2012; Pan & Zhu, 2017; Chen et al., 2017), the modeling capacity of our proposed method can be improved by embedding the (Hierarchical) group structure. For instance, group SpHAM can be easily formulated by replacing the direct decomposition* $\{\mathcal{X}_j\}_{j=1}^p, \mathcal{X}_j \in \mathbb{R}$ *with the subgroups decomposition* $\{\mathcal{X}_d\}_{d=1}^D$*, where each* $\mathcal{X}_d \subset \mathcal{X}$ *may cover more than one variable.*

3.2 Asymptotic theory analysis

In this section, we provide the theoretical analyses of SpHAM, including the following:

- The function approximation errors of SpHAM for stationary time series data (Theorem 1; Section 3.2.1) and non-stationary time series data (Theorem 2; Section 3.2.2);
- An adaptive SpHAM (inspired by above theoretical findings) and its function approximation error for non-stationary time series (See Theorem 3; Section 3.2.3)

Due to space limitation, the high-level outline and detailed proofs are provided in the Appendix A-D.

3.2.1 Function approximation analysis for stationary time series

Through this paper, the marginal distribution w.r.t Y^t is assumed to be almost everywhere supported on $[-M, M]$ for some $M \geq 0$.

Assumption 1. *Assume that* $|Y^t|$*,* $\forall t \in \mathbb{Z}$*, is bounded and there exists a constant* $c > 0$ *such that* $\mathbb{E}|Y^t|^{1+c} < \infty$*,* $\forall t \in \mathbb{Z}$*.*

The moment condition in Assumption 1 is rather weak in the sense that the response variable Y^t possesses infinite variance. The same condition also applies to the distributions of the innovation ε^t, implying that heavy-tailed innovation is allowed.

Assumption 2. *Let* $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K_j(x, x)} < \infty, \forall j = 1, ..., p$*.*

Assumption 2 only requires that the kernel is bounded under compact $\mathcal{X}$, which holds for all Mercer kernels (e.g., Gaussian kernel) and has been used in many learning theory literatures; see, e.g., (Wu et al., 2006; Steinwart & Christmann, 2008; Wu & Zhou, 2008).

The following definitions are needed for Assumption 3. For any $j = 1, ..., p$, we define a kernel integral operator $L_{K_j, T+1} : L_2(\rho_{\mathcal{X}_j}^{T+1}) \to L_2(\rho_{\mathcal{X}_j}^{T+1})$ associated with the kernel K_j by

$$L_{K_j, T+1}(f)(x_j^{T+1}) = \int_{\mathcal{X}_j} K_j(x_j^{T+1}, u_j) f(u_j) d\rho_{\mathcal{X}_j}^{T+1}(u_j).$$

Note that $L_{K_j, T+1}$ is a compact and positive operator on $L_2(\rho_{\mathcal{X}_j}^{T+1})$. According to Mercer theorem, we can find the corresponding normalized eigenpairs $\{(\zeta_i^j, \psi_i^j)\}_{i \geq 1}$ such that $\{\psi_i^j\}_{i \geq 1}$ is an orthonormal basis of $L_2(\rho_{\mathcal{X}}^{T+1})$ and $\zeta_i^j \to 0$ as $i \to \infty$. Then for given $r > 0$, we defined the r-th power $L_{K_j, T+1}^r$ by

$$L_{K_j, T+1}^r (\sum_{i \geq 1} \beta_i^j \psi_i^j) = \sum_{i \geq 1} \beta_i^j (\zeta_i^j)^r \psi_i^j.$$

Assumption 3. *We assume that* $f_j^* : \mathcal{X}_j \to \mathbb{R}, \forall j = 1, ..., p$ *is a function of the form* $f_j^* = L^r_{K_j,T+1}(g_j^*), \forall r \in (0, \frac{1}{2}]$ *with some* $g_j^* \in L_2(\rho_{\mathcal{X}_j}^{T+1}), \forall h \in \mathbb{Z}$.

Assumption 3 is a natural extension from i.i.d setting (Assumption 1 in (Wu et al., 2006; Christmann & Zhou, 2016; Chen & Wang, 2018)) to non-i.i.d time series. Indeed, this assumption stands in most practical cases. For instance, if $r = 0.5$, the ground truth function f^* needs to be a real-valued function in the RKHS $\mathcal{H}_K$. The RKHS admits a large class of bounded and continuous real-valued functions that are in Hilbert space. Moreover, this assumption has been widely used in learning theory; please refer to (Smale & Zhou, 2003; Cucker & Zhou, 2007) for more discussions.

Theorem 1. *Suppose that the process* $\{Z^t\}_{t=-\infty}^{\infty}$ *is stationary. Let Assumptions 1-3 be true. By taking* $\sigma = T^m, \eta = T^{-\frac{1}{4r}}$ *and* $\lambda = T^{-\frac{1}{4r}-m}$*, we have for any* $0 < \delta < 1$*,*

$$\|\hat{f} - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \leq \widetilde{C} \log(1/\delta) T^{\Psi(m,c,r)}$$

with confidence at least $1 - \delta$*, where* $\widetilde{C}$ *is a positive constant independently of* T, λ, η, δ *and* σ *and*

$$\Psi(m, c, r) = \begin{cases} \max\{-\frac{1}{2}, m-1, -cm + \frac{1}{4r} + m - 1\}, & \text{if } m \leq 1 - \frac{1}{4r} \\ \max\{-\frac{1}{2}, m-1, -cm + \frac{1}{2r} + 2m - 2\}, & \text{if } m > 1 - \frac{1}{4r}. \end{cases}$$

Remark 2. *In the stationary case, our bound appears to be novel for the following reasons: a) the result is completely independent of the various mixing conditions which are widely used in the theory analysis of non-i.i.d dependent time series (Mohri & Rostamizadeh, 2009; Yang et al., 2018; Mohri & Rostamizadeh, 2010; Guo & Shi, 2011; Qiu et al., 2015); b) compared with Yang et al. (2018), the innovation assumption is rather weak in the sense that the innovation possesses infinite variance and thus admits a heavy-tailed distribution.*

Corollary 1. *Suppose that the process* $\{Z^t\}_{t=-\infty}^{\infty}$ *is stationary. Let all the conditions in Theorem 1 be true. We then have for any* $0 < \delta < 1$

$$\|\hat{f} - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \leq \widetilde{C} \log(1/\delta) T^{\Psi(m,c)}$$

with confidence at least $1 - \delta$*, where*

$$\Psi(m, c) = \begin{cases} \max\{-\frac{1}{2}, -cm + m - \frac{1}{2}\}, & \text{if } m \leq \frac{1}{2} \\ \max\{m-1, -cm + +2m - 1\}, & \text{if } m > \frac{1}{2}. \end{cases}$$

Figure 1 summaries the convergence rates in Corollary 1 by taking different σ and c. If the innovation is Gaussian distribution with finite variance (i.e., Assumption 1 holds for any $c > 1$), one can arbitrarily select a $\sigma = T^m$ $(0 < m < \frac{1}{2})$ to obtain convergence rates $O(n^{-\frac{1}{2}})$. Moreover, we can see that the convergence rate will decrease as m increases. Combined with the conclusion in Lemma 1 (i.e., the equivalence relation between Huber loss based empirical risk and MSE as $\sigma \to \infty$), it indicates that σ indeed plays a trade-off role between algorithmic robustness and variance-reduction. For the weak moment condition (e.g., $0 < c < 1$), one may get slower convergence rates (e.g., $O(n^{(1-c)m-\frac{1}{2}})$ or $O(n^{(2-c)m-1})$), which also coincides with our intuitive understanding that small σ may be conducive to robust forecasting. Note that our method will not converge when σ and c are both located in the white area in Figure 1.

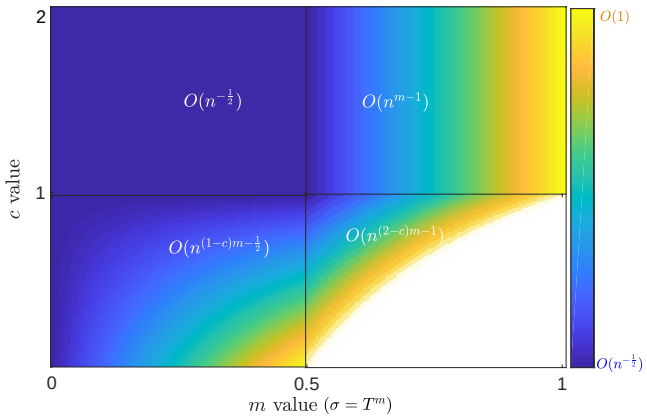

Figure 1: The convergence rates under different σ and c.

As a comparison, due to the non-robustness of the squared loss, the most existing convergence rates are established under the assumption that the innovation is Gaussian distribution with finite variance (see, e.g., Yang et al. (2018); Han et al. (2015); Wang & Yang (2007); Kock & Callot (2015)). However, from Theorem 1 and Corollary 1, the asymptotic convergence of SpHAM can be obtained under weaker moment condition, which verifies the robustness of our method. Recall the learning rate $O(n^{\frac{2d}{2d+1}})$ derived in Yang et al. (2018), where d is the order of smoothness of the component function $f_j, j = 1, ..., p$. Relatively slow convergence rate $O(n^{-\frac{1}{2}})$ we obtained indicates the sacrifice for the absence of mixing condition.

3.2.2 FUNCTION APPROXIMATION ANALYSIS FOR NON-STATIONARY TIME SERIES

In nonstationary time series setting, different Z^ts may follow different distributions. Thus we introduce the weights $\mathbf{s} = \{s_1, ..., s_T\}$ and assign them to the losses made on different sample points in terms of their relevance to forecasting the future Z^{T+1}. For given weights $\{s_t\}_{t=1}^T$, we define the estimator $\hat{f}^{\mathbf{s}} = \sum_{j=1}^p \hat{f}_j^{\mathbf{s}}$ as the minimizer of the following weighted objective

$$\mathcal{E}_\lambda^{\mathbf{s}}(\hat{f}^{\mathbf{s}}) := \min_{f \in \mathcal{H}_Z} \{\sum_{t=1}^T s_t \ell_\sigma(y^t - \sum_{j=1}^p f_j(x_j^t)) + \lambda \Omega(f)\}. \tag{8}$$

Next, we introduce a discrepancy measure to characterize the discrepancy between the target distribution and the distributions of observations (Kuznetsov & Mohri, 2020a).

Definition 2. *For any $f \in \mathcal{H}_Z$, the discrepancy measure with respect to Huber loss is defined as*

$$\text{disc}(\mathbf{s}) := \sup_{f \in \mathcal{H}_Z} \Big\{ \mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \sum_{t=1}^T s_t \mathbb{E}\ell_\sigma(f(x^t) - y^t) \Big\}.$$

The discrepancy measures the non-stationarity of the stochastic process $\{Z^t\}_{t=-\infty}^{\infty}$ with respect to both the loss function ℓ_σ and the hypothesis set $\mathcal{H}_Z$.

Theorem 2. *Let Assumptions 1-3 be true. We assume that $f_j^* \in \mathcal{H}_{K_j}, \forall j = 1, ..., p$. By taking $\sigma = T^{\frac{1}{2c}}$, $\lambda = T^{-1}$ and $\eta = T^{-\frac{1}{2}}$, we have for any $0 < \delta < 1$*

$$\|\hat{f}^{\mathbf{s}} - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \leq \text{disc}(\mathbf{s}) + \mathcal{E}_\lambda^{\mathbf{s}}(\hat{f}^{\mathbf{s}}) + \widetilde{C}_1 \|\mathbf{s}\|_2 T^{\frac{1}{2c}} + \widetilde{C}_2 \log(1/\delta) T^{-\frac{1}{2}}$$

with confidence at least $1 - \delta$, where $\widetilde{C}_1, \widetilde{C}_2$ are two positive constants independently of T, λ, η, δ and σ.

Note that there are several existing studies towards analyzing nonstationary and non-mixing time series, see, e.g., Kuznetsov & Mohri (2020a). Different from them, the error bound we derived is with respect to the function approximation rather than generalization risk, which is very crucial for Huber regression problem, since the convergence of generalization risk cannot imply the convergence of function approximation (Sun et al., 2019; Feng & Wu, 2020).

3.2.3 ADAPTIVE SPARSE HUBER ADDITIVE MODEL FOR NONSTATIONARY TIME SERIES

Theorem 2 illustrates that we shall minimize the following optimization problem for non-stationary time series forecasting:

$$\min_{f \in \mathcal{H}_Z} \{\sum_{t=1}^T s_t \ell_\sigma(y_t - \sum_{j=1}^p f_j(x_j^t)) + \text{disc}(\mathbf{s}) + \lambda_1 \Omega(f) + \lambda_2 T^{\frac{1}{2c}} \|\mathbf{s}\|_2\},$$

where λ_1 and λ_2 are two positive regularization parameters, and c is a positive constant introduced in Assumption 1. Although the discrepancy measure $\text{disc}(\mathbf{s})$ is crucial for such an optimization problem, we cannot obtain its exact value since we do not have access to the distributions of $Z^t, t \in \mathbb{Z}$. Hence, we need to estimate the approximated discrepancy from given data. Inspired by Kuznetsov & Mohri (2020a), one natural and necessary assumption is that there exists an underlying representation relationship between distribution ρ^{T+1} and distributions $\rho^t, t = 1, ..., T$.

Assumption 4. *Denote by a probability set* $\mathbf{q}^* = \{q_t^*\}_{t=1}^T$ *with* $\sum_{t=1}^T q_t^* = 1$. *We assume that the following term is sufficiently small:*

$$\mathrm{disc}(\mathbf{q}^*) := \sup_{f \in \mathcal{H}_Z} [\mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \sum_{t=1}^T q_t \mathbb{E}\ell_\sigma(f(x^t) - y^t)].$$

Note that the priori $\mathbf{q}^*$ can be any distribution, which shall be given empirically according to the trend of the time series. For instance, in a particular scenario of the distribution ρ^{T+1} does not change drastically compared with the distributions $\rho^t, t = 1, ..., T$, Kuznetsov & Mohri (2020a) have proven that the Assumption 4 holds if the $\mathbf{s}^*$ is an uniform distribution over last $l > 0$ observations, where l is a hyper-parameter that can be tuned in practical applications.

Theorem 3. *Let Assumption 4 and the conditions in Theorem 2 be true. We have for any* $0 < \delta < 1$

$$\begin{aligned}\|\hat{f}^{\mathbf{s}} - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \leq \quad & \sup_{f \in \mathcal{H}_Z} \sum_{t=1}^T (q_t^* - s_t)\ell_\sigma(f(x^t) - y^t) + \mathcal{E}_\lambda^{\mathbf{s}}(\hat{f}^{\mathbf{s}}) \\ & + \widetilde{C}_1 \log(1/\delta)(\|\mathbf{q}^* - \mathbf{s}\|_2 + \|\mathbf{s}\|_2 T^{\frac{1}{2c}}) + \widetilde{C}_2 \log(1/\delta) T^{-\frac{1}{2}}\end{aligned}$$

with confidence at least $1 - \delta$, *where* $\widetilde{C}_1, \widetilde{C}_2$ *are two constants independently of* T, λ, η, δ *and* σ.

Finally, the optimization problem of adaptive SpHAM can be formulated as following two stages:

Step A: finding the weight $\hat{\mathbf{s}}$:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \{ \sup_{f \in \mathcal{H}_Z} \sum_{t=1}^T (q_t^* - s_t)\ell_\sigma(f(x^t) - y^t) + \lambda_1 \|\mathbf{q}^* - \mathbf{s}\|_2^2 + \lambda_2 T^{\frac{1}{2c}} \|\mathbf{s}\|_2^2 \} \tag{9}$$

Step B: forecasting:

$$\hat{f}^{\mathbf{s}}(x^{T+1}) = \sum_{j=1}^p \sum_{t=1}^T \alpha_{tj}^{\mathbf{s}} K(x_j^t, x_j^{T+1}),$$

where

$$\alpha^{\mathbf{s}} = \arg\min_{\alpha_j \in \mathbb{R}^T, j=1,...,p} \{ \sum_{t=1}^T \hat{s}_t \ell_\sigma(y^t - \sum_{j=1}^p (\mathbf{K}_j^t)' \alpha_j) + \lambda \sum_{j=1}^p \tau_j \|\alpha_j\|_2 \}. \tag{10}$$

The optimization problems (9) can be solved by standard gradient descent method. Similarly to the strategy for optimization problem (7), we use Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009) for **Step B**. We provide the detailed procedure in Appendix E.

4 EXPERIMENT

This section validates the effectiveness of SpHAM and adaptive SpHAM. In all experiments, the Gaussian kernel $K_j(u, v) = \exp(-\frac{\|u-v\|_2^2}{2d^2})$, where $j = 1, ..., p$ and bandwidth $d > 0$, is employed for constructing the additive data dependent hypothesis space. Due to limited space, the evaluations on real-world data are provided in Appendix F. We consider two synthetic examples as below:

Example A: Inspired by (Kuznetsov & Mohri, 2020a), a stationary time series is generated according to the non-linear additive autoregressive model:

$$Y^t = \frac{3}{2}\sin(\frac{\pi}{2}Y^{t-2}) - \sin(\frac{\pi}{2}Y^{t-3}) + 2\varepsilon^t,$$

where the innovation ε^ts are i.i.d. drawn from Gaussian distribution $N(0, 0.5)$ and Student distribution with degree of freedom 2, respectively.

Example B: Inspired by (Kuznetsov & Mohri, 2020a), a time series with smooth drift is generated by

$$Y^t = \frac{t}{400}\sin(Y^{t-1}) + \frac{1}{2}\varepsilon^t,$$

where the distributions of noise ε^t are the same as above.

Hyper-parameter selection and evaluation criterions: Recall that SpHAM algorithm requires three hyper-parameters: regularization parameter λ, bandwidth of kernel d and Huber parameter σ. We set these parameters according to the suggestions in our Theorems. Based on the suggestion in Theorem 1-3, the selection of Huber parameter σ is $\sigma = T^{\frac{1}{48}}$ and the regularization parameter is $\lambda = T^{-1}$. Moreover, we set the bandwidth $d = 0.5$ and tune $l \in \{100, 150, 200, ..., 350\}$. The evaluation criterions for forecasting used here contains Average Sample Error(ASE)= $\frac{1}{N}\sqrt{\sum_{t=1}^{N}(\hat{f}(x^t) - y^t)^2}$ and True Deviation (TD)= $\frac{1}{N}\sqrt{\sum_{t=1}^{N}(\hat{f}(x^t) - f^*(x^t))^2}$, where N is the number of test samples.

For each example, we generate time series with 4000 sample points. The samples at time $t = \{1500, 1501, ..., 1899\}$ are used as a training set, and the samples at next time $t = \{1900, ..., 1999\}$ are considered as the test data. The competitors include Simple Exponential Smoothing (SES) , TS-SpAM (Yang et al., 2018) and Vanilla Long Short-Term Memory (LSTM). The tuning parameters of other competing methods such as LSTM and SES, are chosen according to their original python packages. All the evaluations are repeated for 50 times. The average results for *Examples A-B* are presented in Table 2.

From Table 2, the results on Example A verify that our SpHAM is competitive with other approaches based on square loss under Gaussian noise, and performs better in presence of heavy-tailed t noise. Moreover, the results on Example B shows the promising performance of our adaptive SpHAM for non-stationary time series forecasting.

Table 2: The results on synthetic data.

		Gaussian noise		Student noise	
	Methods	ASE (Std)	TD (Std)	ASE (Std)	TD (Std)
Example A	SES	0.114(±.008)	0.098(±.005)	0.493(±.313)	0.105(±.009)
	LSTM	**0.063**(±.005)	**0.022**(±.005)	0.538(±.342)	0.219(±.089)
	TS-SpAM	0.070(±.007)	0.035(±.008)	0.505(±.332)	0.131(±.042)
	SpHAM (ours)	0.076(±.005)	0.042(±.008)	**0.492**(±.333)	**0.095**(±.006)
Example B	SES	0.435(±.015)	0.424(±.015)	0.402(±.080)	0.366(±.017)
	LSTM	0.120(±.125)	0.114(±.127)	0.247(±.125)	0.181(±.088)
	TS-SpAM	0.114(±.022)	0.105(±.021)	0.173(±.125)	0.104(±.051)
	SpHAM (ours)	0.115(±.028)	0.112(±.029)	0.179(±.116)	0.096(±.026)
	Adaptive SpHAM (ours)	**0.102**(±.026)	**0.096**(±.026)	**0.172**(±.118)	**0.085**(±.024)

5 Conclusion

We propose an adaptive sparse Huber additive model by integrating Huber loss and $\ell_{2,1}$-norm regularizer into an additive data dependent hypothesis space. We theoretically explore the asymptotic properties of our method for both non-Gaussian and (non)stationary time series. Experimental results on both synthetic and real-world data validate the effectiveness of the proposed method.

Acknowledgments

This work is supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700) and the National Natural Science Foundation of China under Grant No. 12071166. We sincerely appreciate the anonymous ICLR reviewers for their helpful comments.

References

Terrence M. Adams and Andrew B. Nobel. Uniform convergence of vapnik-chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367, 2010.

Alekh Agarwal and John C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *arXiv:2004.13912v1*, 2020.

Richard T. Baillie. Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73(1):5–59, 1996.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bart Bussmann, Jannes Nys, and Steven Latre. Neural additive vector autoregression models for causal discovery in time series data. *arXiv:2010.09429v1*, 2020.

John M. Chambers and Trevor Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

Hong Chen and Yulong Wang. Kernel-based sparse regression with the correntropy-induced loss. *Applied and Computational Harmonic Analysis*, 44(1):144–164, 2018.

Hong Chen, Xiaoqian Wang, Cheng Deng, and Heng Huang. Group sparse additive machine. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 198–208. 2017.

Hong Chen, Changying Guo, Yingjie Wang, and Huijuan Xiong. Sparse additive machine with ramp loss. *Analysis and Applications*, 19(3):509 – 528, 2021a.

Hong Chen, Yingjie Wang, Feng Zheng, Cheng Deng, and Heng Huang. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2373 – 2387, 2021b.

Andreas Christmann and Ding-Xuan Zhou. Learning rates for the risk of kernel based quantile regression estimators in additive models. *Analysis and Applications*, 14(3):449–477, 2016.

Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9:967–991, 2008.

Benoit Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Francesca Dominici, Aidan McDermott, Scott L. Zeger, and Jonathan M. Samet. On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, 156(3):193–203, 2002.

Paul Doukhan. *Mixing: Properties and Examples. Lecture Notes in Statistics*. Springer, 1994.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, 1996.

Jianqing Fan and Qiwei Yao. *Nonlinear time series: Nonparametric and parametric methods*. Springer, 2005.

Yunlong Feng and Qiang Wu. A statistical learning assessment of huber regression. *arXiv:2009.12755v1*, 2020.

Zheng-Chu Guo and Lei Shi. Classification with non-i.i.d. sampling. *Mathematical and Computer Modelling*, 54(5):1347–1364, 2011.

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

Trevor J. Hastie and Robert J. Tibshirani. *Generalized additive models*. London: Chapman and Hall, 1990.

Fengxiang He and Dacheng Tao. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020.

Jianhua Z. Huang and Lijian Yang. Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(2):463–477, 2004.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2009.

Egide Kalisa, Sulaiman Fadlallah, Mabano Amani, Lamek Nahayo, and Gabriel Habiyaremye. Temperature and air pollution relationship during heatwaves in birmingham, uk. *Sustainable Cities and Society*, 43:111–120, 2018.

Kirthevasan Kandasamy and Yaoliang Yu. Additive approximations in high dimensional nonparametric regression via the SALSA. In *International Conference on Machine Learning (ICML)*, pp. 69–78, 2016.

Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for time series prediction with non-stationary processes. In *International Conference on Algorithmic Learning Theory*, 2014.

Vitaly Kuznetsov and Mehryar Mohri. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Annals of Mathematics and Artificial Intelligence*, 88:367–399, 2020a.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems (NIPS)*, 2020b.

Guodong Liu, Hong Chen, and Heng Huang. Sparse shrunk additive models. In *International Conference on Machine Learning (ICML)*, 2020.

Jiajia Liu, Yudong Ye, Chenglong Shen, Yuming Wang, and R Erdelyi. A new tool for cme arrival time prediction using machine learning algorithms: Cat-puma. *The Astrophysical Journal*, 855(2): 109–118, 2018.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust mestimators. *The Annals of Statistics*, 45(2):866–896, 2017.

Lukas Meier, Sara Van De Geer, and Peter Buhlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

George Ruchathi Mwaniki, Chelsea Rosenkrance, H. Will Wallace, B. Tom Jobson, Mathew H. Erickson, Brian K. Lamb, Rick J. Hardy, Rasa Zalakeviciute, and Timothy M. VanReken. Factors contributing to elevated concentrations of pm2.5 during wintertime near boise, idaho. *Atmospheric Pollution Research*, 5(1):96–103, 2014.

Chao Pan and Michael Zhu. Group additive structure identification for kernel nonparametric regression. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4907–4916. 2017.

Jiangtao Peng, Luoqing Li, and Yuan Yan Tang. Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 30(6):1790–1802, 2019.

Huitong Qiu, Sheng Xu, Fang Han, Han Liu, and Brian Caffo. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1843–1851, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: random averages, combinatorial parameters, and learnability. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1984–1992, 2010.

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

Garvesh Raskutti, Martin J. Wainwright, and B Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(2): 389–427, 2012.

Pradeep Ravikumar, Han Liu, John Lafferty, and Larry Wasserman. SpAM: sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71:1009–1030, 2009.

Khaiwal Ravindra, Preety Rattan, Suman Mor, and Ashutosh Nath Aggarwal. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International*, 132:104987, 2019.

Meir Ron. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, 2000.

Jakob Runge, Sebastian Bathiany, Erik M. Bollt, Gustau Camps-Valls, Dim Coumou, Ethan R Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter L. Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10, 2019.

Stephen Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 01:17–41, 2003.

Qiongxia Song and Lijian Yang. Oracally efficient spline smoothing of nonlinear additive autoregression models with simultaneous confidence band. *Journal of Multivariate Analysis*, 101(9): 2008–2025, 2010.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science and Business Media, 2008.

Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13 (2):689–705, 1985.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2019.

Pham Dinh Tao and Le Thi Hoai An. A d.c. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8:476–505, 1998.

Grace Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics, 1990.

Li Wang and Lijian Yang. Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics*, 35(6):2474 – 2503, 2007.

Xiaofeng Wang and Donald E. Brown. The spatio-temporal generalized additive model for criminal incidents. In *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*, pp. 42–47, 2011.

Xiaoqian Wang, Hong Chen, Weidong Cai, Dinggang Shen, and Heng Huang. Regularized modal regression with applications in cognitive impairment prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2017a.

Yimin Wang, Jiajia Liu, Ye Jiang, and Robert Erdélyi. Cme arrival time prediction using convolutional neural network. *The Astrophysical Journal*, 881(11):15, 2019.

Yingjie Wang, Hong Chen, Feng Zheng, Chen Xu, Tieliang Gong, and Yanhong Chen. Multi-task additive models for robust estimation and automatic structure discovery. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

Yulong Wang, Yuan Yan Tang, and Luoqing Li. Correntropy matching pursuit with application to robust digit and face recognition. *IEEE Transactions on Cybernetics*, 47(6):1354–1366, 2017b.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

Qiang Wu and Ding Xuan Zhou. Learning with sample dependent hypothesis spaces. *Computers and Mathematics with Applications*, 56(11):2896–2907, 2008.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Qiang Wu, Yiming Yin, and Ding-Xuan Zhou. Multi-kernel regularized classfiers. *Biometrika*, 23: 108–134, 2007.

Wenjun Yan, Marcus A. Worsley, Thang Pham, Alex Zettl, Carlo Carraro, and Roya Maboudian. Effects of ambient humidity and temperature on the no2 sensing characteristics of ws2/graphene aerogel. *Applied Surface Science*, 450:372–379, 2018.

Yingxiang Yang, Adams Wei Yu, Zhaoran Wang, and Tuo Zhao. Detecting nonlinear causality in multivariate time series with sparse additive models. *arXiv:1803.03919v2*, 2018.

Junming Yin, Xi Chen, and Eric P Xing. Group sparse additive models. In *International Conference on Machine Learning (ICML)*, pp. 871–878, 2012.

Bin Zou, Luoqing Li, and Zongben Xu. The generalization performance of erm algorithm with strongly mixing observations. *Machine Learning*, 75(3):275–295, 2009.

A PROOF SKETCH

This section sketches the proof focusing on the conceptual aspects. The full proofs are provided in Appendix B-D.

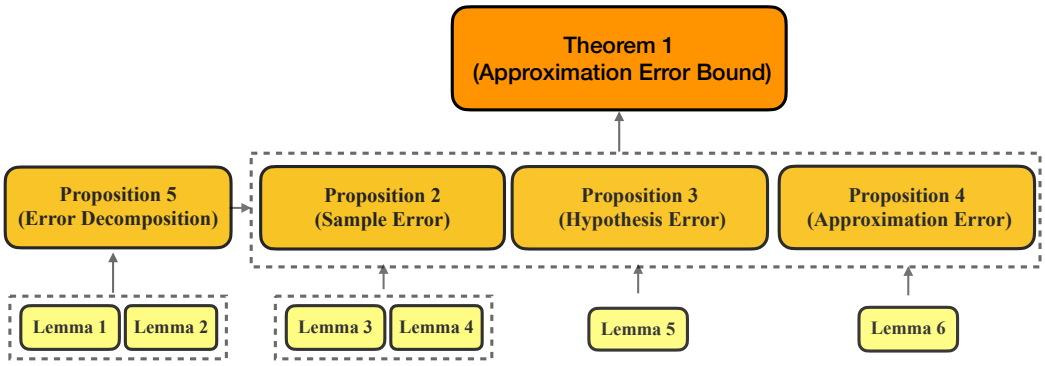

Figure 2: The important ingredients for Theorem 1.

Figure 2 summaries the important ingredients for the proof of Theorem 1. Lemma 1 describes the properties of $\hat{f}$ and Lemma 2 illustrates the relation between huber-based risk and MSE. Following these two lemmas and previous works (Wang et al., 2017a; Feng & Wu, 2020), we then decompose the function error into three important parts: sample error, hypothesis error and approximation error. To bound the sample error, we need to employ the sequential Rademacher complexity (Kuznetsov & Mohri, 2020a), which is the key to measuring the capacity of the data-dependent hypothesis space for non-i.i.d data. Lemmas 3-4 give an important concentration inequality and the upper bound of sequential Rademacher complexity, respectively. Finally, Theorem 1 is obtained by combining Propositions 2-4. The proof of Theorems 2 is similar to the above process, except that we develop a different error decomposition in Proposition 5. On the basis of Theorem 2, by applying an assumption on data distribution, we derive Theorem 3.

The full proofs of Theorems 1-3 are provided in Sections C-E, respectively.

B PROOF OF THEOREM 1

We first illustrate two key properties of $\hat{f}$.

Lemma 1. *Let Assumptions 1-2 be true. From the definition of $\hat{f}$ in Eq. (6), there hold*

$$\|\hat{f}\|_\infty \leq M_{\hat{f}} := \frac{2\kappa^2 M^2}{\lambda T \min_{j=1,\ldots,p} \tau_j}.$$

and

$$\Omega(\hat{f}) = \sum_{j=1}^{p} \tau_j \|\alpha_j^\lambda\|_2 \leq \kappa^{-2} M_{\hat{f}} = \frac{2M^2}{\lambda T \min_{j=1,\ldots,p} \tau_j},$$

where M is a positive constant such that $|Y^t| \leq M, \forall t \in \mathbb{Z}$.

Proof. Denote by $\mathcal{E}_T^\sigma(f) := \sum_{t=1}^T \ell_\sigma(f(x^{(t)}) - y^{(t)})$ for notational convenience. From the definition of $\hat{f}$, we know that

$$\mathcal{E}_T^\sigma(\hat{f}) + \lambda\Omega(\hat{f}) \leq \mathcal{E}_T^\sigma(0).$$

It means that

$$\Omega(\hat{f}) = \sum_{j=1}^{p} \tau_j \|\alpha_j^\lambda\|_2 \leq \lambda^{-1} \mathcal{E}_T^\sigma(0).$$

From the definition of Huber loss (3), we obtain

$$\mathcal{E}_T^{\sigma}(0) = \frac{1}{T}\sum_{t=1}^{T}\ell_\sigma(y^{(t)}) \leq \sigma^2, \text{ if } |y^{(t)}| \leq \sigma, \forall t = 1, ..., T,$$

and

$$\mathcal{E}_T^{\sigma}(0) = \frac{1}{T}\sum_{t=1}^{T}\ell_\sigma(y^{(t)}) = 2\sigma|y^{(t)}| - \sigma^2 \leq 2\sigma M - \sigma^2, \text{ if } |y^{(t)}| > \sigma, \ \forall t = 1, ..., T.$$

Therefore,

$$\sum_{j=1}^{p}\tau_j\|\alpha_j^{\lambda}\|_2 \leq \frac{2M^2}{\lambda T \min_{j=1,...,p}\tau_j}.$$

According to the property of RKHS, we conclude that

$$\|\hat{f}\|_\infty \leq \kappa\|\hat{f}\|_K \leq \kappa\sqrt{\sum_{j=1}^{p}\|\hat{f}_j\|_{K_j}^2} \leq \kappa^2\sum_{j=1}^{p}\|\alpha_j^{\lambda}\|_2 \leq \frac{2\kappa^2 M^2}{\lambda T \min_{j=1,...,p}\tau_j}.$$

This completes the proof by denoting $M_{\hat{f}} = \frac{2\kappa^2 M^2}{\lambda T \min_{j=1,...,p}\tau_j}$. □

Let $\bar{\mathcal{H}}_Z$ be a data-dependent hypothesis space constrained by sparsity-inducing regularizer $\Omega(\cdot)$, i.e.,

$$\bar{\mathcal{H}}_Z = \{f(x) = \sum_{j=1}^{p}\sum_{i=1}^{T}\alpha_{ij}K_j(x_j^{(t)}, x_j) : \sum_{j=1}^{p}\tau_j\|\alpha_j\|_2 \leq \frac{2M^2}{\lambda T \min_{j=1,..,p}\tau_j}\}.$$

For any $h \in \mathbb{Z}$, denote by $\mathcal{E}_{T+1}^{\sigma}(f) := \mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1})$ and $\mathcal{E}_{T+1}(f) := \mathbb{E}(f(x^{T+1}) - y^{T+1})^2$, respectively. In the following, we describe the relationship between $\mathcal{E}_{T+1}^{\sigma}(f)$ and $\mathcal{E}_{T+1}(f), \forall f \in \bar{\mathcal{H}}_Z$. Note that a similar proof is given in (Feng & Wu, 2020) for the i.i.d case.

Lemma 2. *Let Assumptions 1-2 be true. For any $h \in \mathbb{Z}$ and $f \in \bar{\mathcal{H}}_Z$, there holds*

$$\Big|[\mathcal{E}_{T+1}^{\sigma}(f) - \mathcal{E}_{T+1}^{\sigma}(f^*)] - [\mathcal{E}_{T+1}(f) - \mathcal{E}_{T+1}(f^*)]\Big| \leq \frac{M_c}{\sigma^c},$$

where both c and $M_c = 2^{3+c}(M_{f^*} + M_{\hat{f}} + 1)^2\mathbb{E}(|Y^{T+1}|^{1+c})$ are positive constants.

Proof. For any $\sigma > \max\{M + M_{\hat{f}}, 1\}$, we denote two events I_Y and II_Y as follows

$$I_{Y^{T+1}} = \{Y^{T+1} : |Y^{T+1}| \geq \sigma/2\}$$

and

$$II_{Y^{T+1}} = \{Y^{T+1} : |Y^{T+1}| < \sigma/2\}.$$

Then, for any $f \in \bar{\mathcal{H}}_Z$ and $\|f^*\|_\infty = M_{f^*} < \infty$, we have

$$\begin{aligned}
&\Big|[\mathcal{E}_{T+1}^{\sigma}(f) - \mathcal{E}_{T+1}^{\sigma}(f^*)] - [\mathcal{E}_{T+1}(f) - \mathcal{E}_{T+1}(f^*)]\Big| \\
= &\Big|\Big[\int_{\mathcal{Z}}\ell_\sigma(f(x^{T+1}) - y^{T+1})d\rho^{T+1} - \int_{\mathcal{Z}}\ell_\sigma(f^*(x^{T+1}) - y^{T+1})d\rho^{T+1}\Big] - \|f - f^*\|_{L_2(\rho_{\mathcal{X}}^{T+1})}^2\Big| \\
= &\Big|\Big[\int_{\mathcal{X}}\int_{\mathcal{Y}}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \ell_\sigma(f^*(x^{T+1}) - y^{T+1})d\rho_{\mathcal{Y}|\mathcal{X}}^{T+1}d\rho_{\mathcal{X}}^{T+1}\Big] - \|f - f^*\|_{L_2(\rho_{\mathcal{X}}^{T+1})}^2\Big| \\
\leq &\Big|\int_{\mathcal{X}}\int_{I_{Y^{T+1}}\cup II_{Y^{T+1}}}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \ell_\sigma(f^*(x^{T+1}) - y^{T+1})d\rho_{\mathcal{Y}|\mathcal{X}}^{T+1}d\rho_{\mathcal{X}}^{T+1}\Big| \\
&+\Big|\int_{\mathcal{X}}\int_{I_{Y^{T+1}}\cup II_{Y^{T+1}}}(f(x^{T+1}) - y^{T+1})^2 - (f^*(x^{T+1}) - y^{T+1})^2 d\rho_{\mathcal{Y}|\mathcal{X}}^{T+1}d\rho_{\mathcal{X}}^{T+1}\Big| \\
\leq &\Big|\int_{\mathcal{X}}\int_{I_{Y^{T+1}}}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \ell_\sigma(f^*(x^{T+1}) - y^{T+1})d\rho_{\mathcal{Y}|\mathcal{X}}^{T+1}d\rho_{\mathcal{X}}^{T+1}\Big| \\
&+\Big|\int_{\mathcal{X}}\int_{I_{Y^{T+1}}}(f(x^{T+1}) - y^{T+1})^2 - (f^*(x^{T+1}) - y^{T+1})^2 d\rho_{\mathcal{Y}|\mathcal{X}}^{T+1}d\rho_{\mathcal{X}}^{T+1}\Big|.
\end{aligned}$$

The last inequality is based on the fact that

$$|y^{T+1} - f(x^{T+1})| < |y^{T+1}| + \|f\|_\infty < \sigma,\ \forall (x^{T+1}, y^{T+1}) \in \mathcal{X} \times II_{Y^{T+1}}$$

and hence

$$\ell_\sigma(f(x^{T+1}) - y^{T+1}) = (f(x^{T+1}) - y^{T+1})^2.$$

Recalling that the Huber loss is 2σ-Lipschitz continuous, then we get

$$\begin{aligned}
&\Big| \int_{\mathcal{X}} \int_{I_{Y^{T+1}}} \ell_\sigma(f(x^{T+1}) - y^{T+1}) - \ell_\sigma(f^*(x^{T+1}) - y^{T+1}) d\rho^{T+1}_{\mathcal{Y}|\mathcal{X}} d\rho^{T+1}_{\mathcal{X}} \Big| \\
&\leq 2\sigma \Big| \int_{\mathcal{X}} \int_{I_{Y^{T+1}}} |f(x) - f^*(x)| d\rho^{T+1}_{\mathcal{Y}|\mathcal{X}} d\rho^{T+1}_{\mathcal{X}} \Big| \\
&\leq 2\sigma \|f - f^*\|_\infty \mathrm{Prob}\{I_Y\}.
\end{aligned}$$

According to Markov's inequality, for any constant $c > 0$, we further have

$$\mathrm{Prob}\{I_{Y^{T+1}}\} = \mathrm{Prob}\{|Y^{T+1}| \geq \sigma/2\} = \mathrm{Prob}\{|Y^{T+1}|^{1+c} \geq (\sigma/2)^{1+c}\} \leq \frac{2^{1+c}\mathbb{E}(|Y^{T+1}|^{1+c})}{\sigma^{1+c}}.$$

Then we have

$$\begin{aligned}
&\Big| \int_{\mathcal{X}} \int_{I_{Y^{T+1}}} \ell_\sigma(f(x^{T+1}) - y^{T+1}) - \ell_\sigma(f^*(x^{T+1}) - y^{T+1}) d\rho^{T+1}_{\mathcal{Y}|\mathcal{X}} d\rho^{T+1}_{\mathcal{X}} \Big| \\
&\leq \frac{2^{2+c}\|f - f^*\|_\infty \mathbb{E}(|Y^{T+1}|^{1+c})}{\sigma^c}.
\end{aligned}$$

Moreover, the second term in the right-hand side can be bounded by

$$\begin{aligned}
&\Big| \int_{\mathcal{X}} \int_{I_{Y^{T+1}}} (f(x^{T+1}) - y^{T+1})^2 - (f^*(x^{T+1}) - y^{T+1})^2 d\rho^{T+1}_{\mathcal{Y}|\mathcal{X}} d\rho^{T+1}_{\mathcal{X}} \Big| \\
&\leq \|f - f^*\|_\infty \int_{\mathcal{X}} \int_{I_{Y^{T+1}}} |2y^{T+1} - f(x^{T+1}) - f^*(x^{T+1})| d\rho^{T+1}_{\mathcal{Y}|\mathcal{X}} d\rho^{T+1}_{\mathcal{X}} \\
&\leq \|f - f^*\|_\infty \Big[\int_{I_{Y^{T+1}}} 2|y^{T+1}| d\rho^{T+1}_{\mathcal{Y}} + (\|f\|_\infty + \|f^*\|_\infty) \mathrm{Prob}(I_{Y^{T+1}}) \Big].
\end{aligned}$$

According to Holder inequality, we have

$$\int_{I_{Y^{T+1}}} |Y^{T+1}| d\rho^{T+1}_{\mathcal{Y}} \leq (\mathrm{Prob}(I_{Y^{T+1}}))^{\frac{c}{1+c}} (\mathbb{E}|Y^{T+1}|^{1+c})^{\frac{1}{1+c}} \leq \frac{2^c \mathbb{E}(|Y^{T+1}|^{1+c})}{\sigma^c}.$$

Then we can deduce that

$$\begin{aligned}
&\|f - f^*\|_\infty \Big[\int_{I_{Y^{T+1}}} 2|y^{T+1}| d\rho^{T+1}_{\mathcal{Y}} + (\|f\|_\infty + \|f^*\|_\infty) \mathrm{Prob}(I_{Y^{T+1}}) \Big] \\
&\leq \frac{(2^{2+c} + 2^{1+c})\|f - f^*\|_\infty \mathbb{E}(|Y^{T+1}|^{1+c})}{\sigma^c} + \frac{2^{1+c}(\|f\|_\infty + \|f^*\|_\infty)^2 \mathbb{E}(|Y^{T+1}|^{1+c})}{\sigma^{1+c}}
\end{aligned}$$

Finally, there holds

$$\Big| \Big[\int_{\mathcal{Z}} \ell_\sigma(f(x^{T+1}) - Y^{T+1}) - \int_{\mathcal{Z}} \ell_\sigma(f^*(x^{T+1}) - Y^{T+1}) \Big] - \|f - f^*\|^2_{L_2(\rho^{T+1}_{\mathcal{X}})} \Big| \lesssim \frac{M_c}{\sigma^c}$$

with

$$M_c = 2^{3+c}(M_{f^*} + M_{\hat{f}} + 1)^2 \mathbb{E}(|Y^{T+1}|^{1+c}).$$

This completes the proof. □

Define a stepping-stone function with respect to the distribution ρ^{T+1}:

$$f^*_{\eta,T+1} = \sum_{j=1}^p f^*_{\eta,T+1,j} = \arg\min_{f \in \mathcal{H}_K} \mathbb{E}\ell_\sigma(f(x^{T+1}) - Y^{T+1}) + \eta \sum_{j=1}^p \tau_j \|f_j\|^2_{K_j}.$$

To establish the bound of function approximation for stationary time series setting, we make the following error decomposition.

Proposition 1. *Let Assumptions 1-2 be true. For any $f \in \bar{\mathcal{H}}_Z$, there holds*

$$\|f - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \leq E_1 + E_2 + E_3 + \frac{2M_c}{\sigma^c},$$

where

$$E_1 = \{\mathcal{E}^\sigma_{T+1}(f) - \mathcal{E}^\sigma_T(f)\} + \{\mathcal{E}^\sigma_T(f^*_{\eta,T+1}) - \mathcal{E}^\sigma_{T+1}(f^*_{\eta,T+1})\}$$

$$E_2 = \mathcal{E}_{T+1}(f^*_{\eta,T+1}) - \mathcal{E}_{T+1}(f^*) + \eta \sum_{j=1}^p \tau_j \|f^*_{\eta,T+1,j}\|^2_{K_j}$$

and

$$E_3 = \{\mathcal{E}^\sigma_T(\hat{f}) + \lambda\Omega(\hat{f}) - \mathcal{E}^\sigma_T(\hat{f}_\eta) - \eta \sum_{j=1}^p \tau_j \|\hat{f}_{\eta,j}\|^2_{K_j}\}.$$

Proof. According to Lemma 2, for any $f \in \bar{\mathcal{H}}_Z$, we can make following error decomposition

$$\begin{aligned}
& \|f - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} = \mathcal{E}_{T+1}(f) - \mathcal{E}_{T+1}(f^*) \leq \mathcal{E}^\sigma_{T+1}(f) - \mathcal{E}^\sigma_{T+1}(f^*) + \frac{M_c}{\sigma^c} \\
= \; & \underbrace{\{\mathcal{E}^\sigma_{T+1}(f) - \mathcal{E}^\sigma_T(f)\} + \{\mathcal{E}^\sigma_T(f^*_{\eta,T+1}) - \mathcal{E}^\sigma_{T+1}(f^*_{\eta,T+1})\}}_{E_1} + \{\mathcal{E}^\sigma_T(f) - \mathcal{E}^\sigma_T(f^*_{\eta,T+1})\} \\
& \qquad + \mathcal{E}^\sigma_{T+1}(f^*_{\eta,T+1}) - \mathcal{E}^\sigma_{T+1}(f^*) + \frac{M_c}{\sigma^c} \\
\leq \; & E_1 + \underbrace{\mathcal{E}_{T+1}(f^*_{\eta,T+1}) - \mathcal{E}_{T+1}(f^*) + \eta \sum_{j=1}^p \tau_j \|f^*_{\eta,T+1,j}\|^2_{K_j}}_{E_2} + \{\mathcal{E}^\sigma_T(f) + \lambda\Omega(f) - \mathcal{E}^\sigma_T(f^*_{\eta,T+1}) \\
& \qquad - \eta \sum_{j=1}^p \tau_j \|f^*_{\eta,T+1,j}\|^2_{K_j}\} + \frac{2M_c}{\sigma^c} \\
\leq \; & E_1 + E_2 + \underbrace{\{\mathcal{E}^\sigma_T(f) + \lambda\Omega(f) - \mathcal{E}^\sigma_T(\hat{f}_\eta) - \eta \sum_{j=1}^p \tau_j \|\hat{f}_{\eta,j}\|^2_{K_j}\}}_{E_3} + \frac{2M_c}{\sigma^c} \\
\leq \; & E_1 + E_2 + E_3 + \frac{2M_c}{\sigma^c}.
\end{aligned}$$

□

In statistical machine learning community, we call E_1, E_2 and E_3 sample error, approximation error and hypothesis error, respectively. The sample error E_1 describes the divergence between the empirical risk $\mathcal{E}^\sigma_T(f)$ and the expected risk $\mathcal{E}^\sigma(f)$. The hypothesis error E_2 characterizes the difference between the empirical regularized risks with $\mathcal{H}_K$ and $\mathcal{H}_Z$. The approximation error measures the approximation ability of RKHS $\mathcal{H}_K$ to $\mathcal{H}$.

B.1 Sample Error

In this section, we focus on providing the bound of sample error E_1. Unfortunately, the traditional tools for complexity analysis such as covering number (Ron, 2000; Guo & Shi, 2011), Rademacher complexity (Mohri & Rostamizadeh, 2009)) and concentration inequalities ((Wu et al., 2007)), cannot be applied into this non-i.i.d. setting directly. To solve this problem, we employ the sequential Rademacher complexity developed in (Rakhlin et al., 2010; Kuznetsov & Mohri, 2020a). We define a function-based random variable as

$$\xi_f(z) = \ell_\sigma(f(x) - y) - \ell_\sigma(f^*_{\eta,T+1}(x) - y), f \in \bar{\mathcal{H}}_Z.$$

Now, we turn to establish the bound of

$$\mathbb{E}\xi_f(z^{T+1}) - \frac{1}{T}\sum_{t=1}^T \xi_f(z^t), \;\; \forall f \in \bar{\mathcal{H}}_Z.$$

A necessary ingredient needed for our analysis is data-dependent sequential Rademacher complexity Rakhlin et al. (2010), which we review in the following. We adopt the following definition of a complete binary tree: a $\mathcal{Z}$-valued complete binary tree $\mathbf{v}$ is a sequence $(v_1, ..., v_T)$ of T mappings $v_t : \{\pm 1\}^{t-1} \to \mathcal{Z}, t \in [1, T]$. A path in the tree is $\gamma = (\gamma_1, ..., \gamma_{T-1}) \in \{\pm 1\}^{T-1}$. To simplify the notation, we will write $v_t(\boldsymbol{\gamma})$ instead $v_t(\gamma_1, ..., \gamma_{t-1})$, even though v_t depends only on the first $t-1$ elements of $\boldsymbol{\gamma}$. The following definition generalizes the classical notion of Rademacher complexity to sequential setting.

Definition 3. *The sequential Rademacher complexity $\mathcal{R}_T(G)$ of a function class $\mathcal{G}$ is defined by*

$$\mathcal{R}_T(\mathcal{G}) = \sup_{\mathbf{v}} \mathbb{E}[\sup_{g \in \mathcal{G}} \frac{1}{T} \sum_{t=1}^{T} \gamma_t g(v_t(\boldsymbol{\gamma}))],$$

where the supremum is taken over all complete binary trees of depth T with values in $\mathcal{Z}$ and $\boldsymbol{\gamma}$ is a sequence of Rademacher random variables.

Based on the definition of sequential Rademacher complexity, we have the following concentration inequality:

Lemma 3. *Suppose that the time series $\{Z^t\}_{t=-\infty}^{\infty}$ is strictly stationary. For any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in \bar{\mathcal{H}}_Z$ and all $\alpha > 0$*

$$\mathbb{E}\xi_f(z^{T+1}) \leq \frac{1}{T} \sum_{t=1}^{T} \xi_f(z^t) + \frac{1}{\sqrt{T}} + 6M_{\ell_\sigma} \sqrt{4\pi \log T} \mathcal{R}_T(\mathcal{G}) + \frac{M_{\ell_\sigma} \sqrt{8 \log \frac{1}{\delta}}}{\sqrt{T}},$$

where $\mathcal{G} := \{\ell_\sigma(f, z) + \ell_\sigma(f_\eta, z) : f \in \bar{\mathcal{H}}_Z\}$ and M_{ℓ_σ} is a positive constant such that $\ell_\sigma(f(x^{(t)}) - y^{(t)}) \leq M_{\ell_\sigma}$ for any $f \in \bar{\mathcal{H}}_Z$ and $t \in \mathbb{Z}$.

According to the property of sequential Rademacher complexity (see Proposition 14 in (Rakhlin et al., 2015)), we have

$$\mathcal{R}_T(\mathcal{G}) = \mathcal{R}_T(\ell_\sigma \circ \bar{\mathcal{H}}_Z, z) \leq 16\sigma(1 + 4\sqrt{2} \log^{3/2}(eT^2)) \mathcal{R}_T(\bar{\mathcal{H}}_Z) \tag{11}$$

with the 2σ-Lipschitz constant of $\ell_\sigma(\cdot)$.

Lemma 4. *Under Assumptions 1-2, there holds*

$$\mathcal{R}_T(\mathcal{G}) \leq \frac{32\sigma[1 + 4\sqrt{2} \log^{3/2}(eT^2)] M^2 p^{\frac{1}{2}} \kappa}{T}.$$

Proof. In fact, to every kernel K_j, we can associate a feature map ϕ with inner product $< \phi(x_j), \phi(x_j') >= K(x_j, x_j')$ for any $x_j, x_j' \in \mathcal{X}_j$ and $j = 1, ..., p$. Then the constrained data-dependent hypothesis space $\bar{\mathcal{H}}_Z$ can be rewritten as

$$\bar{\mathcal{H}}_Z = \{f(x) = \sum_{j=1}^{p} \sum_{t=1}^{T} \alpha_{ij} \phi(x_j^{(t)}) \phi(x_j) : \sum_{j=1}^{p} \tau_j \|\alpha_j\|_2 \leq \frac{2M^2}{\lambda T \min_{j=1,..,p} \tau_j}\}.$$

By direct computation, we have

$$
\begin{aligned}
\mathcal{R}_T(\bar{\mathcal{H}}_Z) &= \sup_{\{\bar{x}_t,\bar{y}_t\}_{t=1}^T} \mathbb{E}_\sigma[\sup_{\alpha\in\mathbb{R}^{Tp}} \frac{1}{T}\sum_{t=1}^T \sigma_t(\sum_{j=1}^p\sum_{i=1}^T \alpha_{ij}\phi(x_j^{(i)})\phi(\bar{x}_j^{(t)}(\sigma)) - y^{(t)})] \\
&= \frac{1}{T}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \mathbb{E}_\sigma \sup_{\alpha\in\mathbb{R}^{Tp}} \sum_{t=1}^T \sigma_t \sum_{j=1}^p\sum_{i=1}^T \alpha_{ij}\phi(x_j^{(i)})\phi(\bar{x}_j^{(t)}(\sigma)) - \frac{1}{T}\sup_{\{\bar{y}^{(t)}\}_{t=1}^T} \mathbb{E}_\sigma \sum_{t=1}^T \sigma_t y^{(t)}(\sigma) \\
&= \frac{1}{T}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \mathbb{E}_\sigma \sup_{\alpha\in\mathbb{R}^{Tp}} \sum_{t=1}^T \sigma_t \sum_{j=1}^p\sum_{i=1}^T \alpha_{ij}\phi(x_j^{(i)})\phi(\bar{x}_j^{(t)}(\sigma)) \\
&= \frac{1}{T}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \mathbb{E}_\sigma \sup_{\alpha\in\mathbb{R}^{Tp}} \sum_{t=1}^T \sigma_t \sum_{j=1}^p\sum_{i=1}^T \alpha_{ij}K(x_j^{(i)},\bar{x}_j^{(t)}(\sigma)) \\
&= \frac{1}{T}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \mathbb{E}_\sigma \sup_{\alpha\in\mathbb{R}^{Tp}} \alpha'\sum_{t=1}^T \sigma_t\mathbf{K}_t(\sigma),\ \ \mathbf{K}_t(\sigma) = (K(x_1^{(1)},\bar{x}_1^{(t)}(\sigma)),...,K(x_p^{(T)},\bar{x}_p^{(t)}(\sigma))' \in \mathbb{R}^{Tp} \\
&\le \frac{1}{T}\sup_{\alpha\in\mathbb{R}^{Tp}} \|\alpha'\| \sup_{\{\bar{x}_t\}_{t=1}^T} \mathbb{E}_\sigma\|\sum_{t=1}^T \sigma_t\mathbf{K}_t(\sigma)\|_2 \\
&\le \frac{2M^2}{T^2}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \mathbb{E}_\sigma\|\sum_{t=1}^T \sigma_t\mathbf{K}_t(\sigma)\|_2 \\
&= \frac{2M^2}{T^2}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \sqrt{\mathbb{E}_\sigma[\sum_{s,t=1}^T \sigma_t\sigma_s\mathbf{K}_t'(\sigma)\mathbf{K}_s(\sigma)]} \\
&\le \frac{2M^2}{T^2}\sup_{\{\bar{x}^{(t)}\}_{t=1}^T} \sqrt{\mathbb{E}_\sigma[\sum_{t=1}^T \mathbf{K}_t'(\sigma)\mathbf{K}_t(\sigma)]} \\
&\le \frac{2M^2p^{\frac{1}{2}}\kappa}{T}
\end{aligned}
$$

Thus, combining above result with Eq. (11). We get the desirable result. □

Combining the results in Lemma 4 with the inequality in Lemma 3, we then obtain the bound of sample error E_1 for the strictly stationary time series setting.

Proposition 2. *Under Assumptions 1-2, for any $\delta > 0$, the following inequality holds for all $f \in \bar{\mathcal{H}}_Z$ and all $\alpha > 0$*

$$E_1 \le \frac{1}{\sqrt{T}} + 6M_{\ell_\sigma}\sqrt{4\pi\log T}R + \frac{M_{\ell_\sigma}\sqrt{8\log\frac{1}{\delta}}}{\sqrt{T}}$$

with probability at least $1-\delta$, where

$$R = \frac{32\sigma[1+4\sqrt{2}\log^{3/2}(eT^2)]M^2p^{\frac{1}{2}}\kappa}{T}.$$

B.2 Approximation Error

Since the output function $f^*_{\eta,T+1}$ is the optimal estimator in RKHS $\mathcal{H}_K$ for time $T+1$, the learning rates of the learning algorithm indeed depend on the approximation ability of the hypothesis space $\mathcal{H}_K$ with respect to the optimal risk $\mathcal{E}(f^*)$ measured by the approximation error E_2. For any $j = 1,...,p$ and $h \in \mathbb{Z}$, we first define the kernel integral operator $L_{K_j,T+1} : L_2(\rho_{\mathcal{X}_j}^{T+1}) \to L_2(\rho_{\mathcal{X}_j}^{T+1})$ associated with the kernel K_j by

$$L_{K_j,T+1}(f)(x_j^{T+1}) = \int_{\mathcal{X}_j} K_j(x_j^{T+1},u_j)f(u_j)d\rho_{\mathcal{X}_j}^{T+1}(u_j), \forall h \in \mathbb{Z}.$$

Note that $L_{K_j,T+1}$ is a compact and positive operator on $L_2(\rho_{\mathcal{X}_j}^{T+1})$. According to Mercer theorem, we can find the corresponding normalized eigenpairs $\{(\zeta_{h,i}^j, \psi_{h,i}^j)\}_{i\geq 1}$ such that $\{\psi_{h,i}^j\}_{i\geq 1}$ is an orthonormal basis of $L_2(\rho_{\mathcal{X}}^{T+1})$ and $\zeta_{h,i}^j \to 0$ as $i \to \infty$. Then for given $r > 0$, we defined the r-th power $L_{K_j,T+1}^r$ by

$$L_{K_j,T+1}^r(\sum_{i\geq 1} \beta_{h,i}^j \psi_{h,i}^j) = \sum_{i\geq 1} \beta_{h,i}^j (\zeta_{h,i}^j)^r \psi_{h,i}^j.$$

We introduce an intermediate function as follows:

$$\tilde{f}_{\eta,T+1,j} = (L_{K_j,T+1} + \eta\tau_j I)^{-1} L_{K_j,T+1} f_j^*, \forall j = 1, ..., p.$$

Lemma 5. *Under Assumption 3, for the intermediate function $\tilde{f}_{\eta,T+1,j}$ defined above, there holds*

$$\mathbb{E}\|\tilde{f}_{\eta,T+1,j}(x_j^{T+1}) - f_j^*(x_j^{T+1})\| + \eta\tau_j \|\tilde{f}_{\eta,T+1,j}(x_j^{T+1})\|_{K_j}^2 \leq 2(\eta\tau_j)^{2r} \|L_{K_j,T+1}^{-r} f_j^*\|_2^2.$$

Proof. Under Assumption 3, for any $h \in \mathbb{Z}$, we know that $f_j^* = L_{K_j,h}^r(g_{j,h}^*)$ for some $g_{h,j}^* = \sum_{i\geq 1} \beta_{h,i}\psi_{h,i}^j \in L_2(\rho_{\mathcal{X}_j}^{T+1})$. Then we have

$$f_j^* = L_{K_j,T+1}^r(\sum_{i\geq 1} \beta_{h,i}\psi_{h,i}^j) = \sum_{i\geq 1} (\zeta_{h,i}^j)^r \beta_{h,i}\psi_{h,i}^j.$$

The case $r = \frac{1}{2}$ means each f_j^* lies in the RKHS $\mathcal{H}_{K_j}$. Then we have

$$\tilde{f}_{\eta,h,j} - f_j^* = (L_{K_j,h} + \eta\tau_j I)^{-1} L_{K_j,h} f_j^* - f_j^* = \sum_{i\geq 1} \frac{\eta\tau_j (\zeta_{h,i}^j)^r \beta_{h,i}^j \psi_{h,i}^j}{\zeta_i^j + \eta\tau_j}.$$

Then we have

$$\begin{aligned}
\|\tilde{f}_{\eta,T+1,j} - f_j^*\|_2^2 &= (\sum_{i\geq 1} \frac{\eta\tau_j (\zeta_{h,i}^j)^r \beta_{h,i}^j}{\zeta_{h,i}^j + \eta\tau_j})^2 = (\eta\tau_j)^{2r} \sum_{i\geq 1} (\frac{\eta\tau_j}{\zeta_{h,i}^j + \eta\tau_j})^{2-2r} (\frac{\zeta_{h,i}^j}{\zeta_{h,i}^j + \eta\tau_j})^{2r} (\beta_{h,i}^j)^2 \\
&\leq (\eta\tau_j)^{2r} \sum_{i\geq 1} (\beta_{h,i}^j)^2 = (\eta\tau_j)^{2r} \|L_{K_j,T+1}^{-r} f_j^*\|_2^2.
\end{aligned}$$

Similarly, we also have

$$\begin{aligned}
\eta\tau_j \|\tilde{f}_{\eta,T+1,j}\|_2^2 &= (\eta\tau_j)^2 \sum_{i\geq 1} (\frac{(\zeta_{h,i}^j)^{1+r}}{\zeta_{h,i}^j + \eta\tau_j})^2 (\beta_{h,i}^j)^2 = (\eta\tau_j)^{2r} \sum_{i\geq 1} (\frac{\zeta_{h,i}^j}{\zeta_{h,i}^j + \eta\tau_j})^{2+2r} (\frac{\eta\tau_j}{\zeta_{h,i}^j + \eta\tau_j})^{2-2r} (\beta_{h,i}^j)^2 \\
&\leq (\eta\tau_j)^{2r} \|L_{K_j,T+1}^{-r} f_j^*\|_2^2.
\end{aligned}$$

□

Proposition 3. *Under Assumptions 3, there holds*

$$E_2 = \mathcal{E}(f_{\eta,T+1}^*) - \mathcal{E}(f^*) + \eta \sum_{j=1}^p \tau_j \|f_{\eta,T+1,j}^*\|_K^2 \leq (p+1) \sum_{j=1}^p (\eta\tau_j)^{2r} \|L_{K_j,T+1}^{-r} f_j^*\|_2^2.$$

Proof. According to Lemma 1, we have

$$\begin{aligned}
& \mathcal{E}_{T+1}(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}(f^*) + \eta \sum_{j=1}^p \tau_j \|f_{\eta,T+1,j}^*\|_K^2 \\
\leq\ & \mathcal{E}_{T+1}^\sigma(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}^\sigma(f^*) + \eta \sum_{j=1}^p \tau_j \|f_{\eta,T+1,j}^*\|_K^2 + \frac{M_c}{\sigma^c} \\
\leq\ & \mathcal{E}_{T+1}^\sigma(\tilde{f}_{\eta,T+1}) - \mathcal{E}_{T+1}^\sigma(f^*) + \eta \sum_{j=1}^p \tau_j \|\tilde{f}_{\eta,T+1,j}\|_K^2 + \frac{M_c}{\sigma^c} \\
\leq\ & \mathcal{E}_{T+1}(\tilde{f}_{\eta,T+1}) - \mathcal{E}_{T+1}(f^*) + \eta \sum_{j=1}^p \tau_j \|\tilde{f}_{\eta,T+1,j}\|_K^2 + \frac{2M_c}{\sigma^c}
\end{aligned}$$

Then we have

$$\begin{aligned}\mathcal{E}_{T+1}(\tilde{f}_{\eta,T+1}) - \mathcal{E}_{T+1}(f^*) + \eta\sum_{j=1}^{p}\tau_j\|\tilde{f}_{\eta,T+1,j}\|_K^2 &= \|\tilde{f}_{\eta,T+1} - f^*\|_2^2 + \eta\sum_{j=1}^{p}\tau_j\|\tilde{f}_{\eta,T+1,j}\|_K^2 \\ &\leq \|\sum_{j=1}^{p}[\tilde{f}_{\eta,h,j} - f_j^*]\|_2^2 + \eta\sum_{j=1}^{p}\tau_j\|\tilde{f}_{\eta,h,j}\|_K^2 \\ &\leq \sum_{j=1}^{p}p\|\tilde{f}_{\eta,T+1,j} - f_j^*\|_2^2 + \eta\sum_{j=1}^{p}\tau_j\|\tilde{f}_{\eta,T+1,j}\|_K^2 \\ &\leq (p+1)\sum_{j=1}^{p}(\eta\tau_j)^{2r}\|L_{K_j,T+1}^{-r}f_j^*\|_2^2.\end{aligned}$$

Combing above results, we get the desired result. □

B.3 Hypothesis Error

This section focuses on bounding the hypothesis error

$$E_3 = \{\mathcal{E}_T^\sigma(\hat{f}) + \lambda\Omega(\hat{f}) - \mathcal{E}_T^\sigma(\hat{f}_\eta) - \eta\sum_{j=1}^{p}\|\hat{f}_{\eta,j}\|_K^2\}.$$

We first give the properties of $\hat{f}$ and then use them to bridge $\hat{f}$ and f_η.

Lemma 6. *For all $j = 1, ..., p$, there holds*

$$\tau_j\|\alpha_j^\eta\|_2 = \frac{1}{\eta T}\sqrt{\sum_{t=1}^{T}(\ell'_\sigma(y^t - \sum_{j=1}^{p}\boldsymbol{K}'_{jt}\alpha_j^\eta))^2} \leq \frac{2\sigma}{\eta T^{\frac{1}{2}}}$$

Proof. Recall the represent theorem which ensures that

$$\hat{f}_\eta = \sum_{j=1}^{p}\sum_{t=1}^{T}\alpha_{tj}^\eta K_j(x_j^t, \cdot),\ \alpha_{tj}^\eta \in \mathbb{R}.$$

For notation simplicity, denote $\alpha_j^\eta = (\alpha_{1j}^\eta, ..., \alpha_{Tj}^\eta)' \in \mathbb{R}^T$ and $\boldsymbol{\alpha}^\eta = ((\alpha_1^\eta)', ..., (\alpha_p^\eta)')' \in \mathbb{R}^{Tp}$. From the definition of Eq. (7), we deduce that

$$\boldsymbol{\alpha}^\eta = \underset{\alpha\in\mathbb{R}^{Tp}}{\arg\min}\{\frac{1}{T}\sum_{t=1}^{T}\ell_\sigma(y^t - \sum_{j=1}^{p}\boldsymbol{K}'_{jt}\alpha_j) + \eta\sum_{j=1}^{p}\tau_j(\alpha_j^\eta)'\boldsymbol{K}_j\alpha_j\},$$

where $\boldsymbol{K}_j = \{K(x_j^s, x_j^i)\}_{i,s=1}^n \in \mathbb{R}^{T\times T}$. we then have

$$\frac{1}{T}\sum_{t=1}^{T}\ell'_\sigma(y^t - \sum_{j=1}^{p}\boldsymbol{K}'_{jt}\alpha_j^\eta)\boldsymbol{K}'_{ji} = \eta\tau_j\boldsymbol{K}_j\alpha_j^\eta.$$

It is easy to deduce that

$$\alpha_j^\eta = \frac{1}{\eta\tau_j T}(\ell'_\sigma(y^1 - \sum_{j=1}^{p}\boldsymbol{K}'_{j1}\alpha_j^\eta), ..., \ell'_\sigma(y^T - \sum_{j=1}^{p}\boldsymbol{K}'_{jT}\alpha_j^\eta))'.$$

Then we obtain that, for any $j = 1, ..., p$,

$$\tau_j\|\alpha_j^\eta\|_2 = \frac{1}{\eta T}\sqrt{\sum_{t=1}^{T}(\ell'_\sigma(y^t - \sum_{j=1}^{p}\boldsymbol{K}'_{jt}\alpha_j^\eta))^2} \leq \frac{2\sigma}{\eta T^{\frac{1}{2}}}$$

This completes the proof. □

Proposition 4. *Under Assumptions 3, there holds*

$$E_3 \leq \lambda\Omega(\hat{f}_\eta) = \lambda \sum_{j=1}^{p} \tau_j \|\alpha_j^\eta\|_2 \leq \frac{2\lambda p\sigma}{\eta T^{\frac{1}{2}}}.$$

Proof.

$$\begin{aligned} E_3 &= \{\mathcal{E}_T^\sigma(\hat{f}) + \lambda\Omega(\hat{f}) - \mathcal{E}_T^\sigma(\hat{f}_\eta) - \eta \sum_{j=1}^{p} \|\hat{f}_{\eta,j}\|_K^2\} \\ &= \mathcal{E}_T^\sigma(\hat{f}) + \lambda\Omega(\hat{f}) - \mathcal{E}_T^\sigma(\hat{f}_\eta) - \lambda\Omega(\hat{f}_\eta) - \eta \sum_{j=1}^{p} \|\hat{f}_{\eta,j}\|_K^2 + \lambda\Omega(\hat{f}_\eta) \\ &\leq \lambda\Omega(\hat{f}_\eta) \end{aligned}$$

Combining the above inequality with Lemma 6, we get that

$$E_3 \leq \lambda\Omega(\hat{f}_\eta) = \lambda \sum_{j=1}^{p} \tau_j \|\alpha_j^\eta\|_2 \leq \frac{2\lambda p\sigma}{\eta T^{\frac{1}{2}}}.$$

□

Proof of Theorem 1: Combining Propositions 1-4, for $T\eta \leq 1$, we have with confidence $1-\delta$

$$\|f - f^*\|_{L_2(\rho_{\mathcal{X}}^{T+1})}^2 \leq \widetilde{C}\log(1/\delta)(T^{-\frac{1}{2}} + T^{-1}\sigma + \eta^{2r} + \lambda\eta^{-1}T^{-\frac{1}{2}}\sigma + \sigma^{-c}\lambda^{-2}T^{-2} + \sigma^{-c}\lambda^{-1}T^{-1}),$$

where $\widetilde{C}$ is a positive constant independently of $T, \lambda, \eta, \delta, \sigma$. By taking $\sigma = T^m, \eta = T^\beta$ and $\lambda = T^\gamma$, we then have with confidence $1-\delta$

$$\|f - f^*\|_{L_2(\rho_{\mathcal{X}}^{T+1})}^2 \leq \widetilde{C}\log(1/\delta)T^{\Psi(m,\beta,\gamma,c,r)},$$

where

$$\Psi(m,\beta,\gamma,c,r) = \max\{-\frac{1}{2}, m-1, 2r\beta, \gamma - \frac{1}{2} - \beta + m, -cm - 2\gamma - 2, -cm - \gamma - 1\}.$$

By taking $\sigma = \frac{1}{2c}$, $\beta = -\frac{1}{4r}$ and $\gamma = -\frac{1}{4r} - \frac{1}{2c}$ Direct computation shows that

$$\Psi(m,c,r) = \begin{cases} \max\{-\frac{1}{2}, \frac{1}{2c} - 1\}, & \text{if } m \leq 1 - \frac{1}{4r} \\ \max\{-\frac{1}{2}, m-1, -cm + \frac{1}{2r} + 2m - 2\}, & \text{if } m > 1 - \frac{1}{4r}. \end{cases}$$

This completes the proof.

C Proof of Theorem 2

Denote by $\mathbf{s} = \{s_t\}_{t=1}^T$ a probability set with $\sum_{t=1}^T s_t = 1$. We define following functions associated with the probability set $\mathbf{s}$:

$$f_{T+1}^* = \sum_{j=1}^{p} f_{T+1,j}^* = \arg\min_{f \in \mathcal{H}_K} \mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1}),$$

$$f_T^{*,\mathbf{s}} = \sum_{j=1}^{p} f_{T,j}^{*,\mathbf{s}} = \arg\min_{f \in \mathcal{H}_K} \sum_{t=1}^{T} s_t \ell_\sigma(f(x^t) - y^t),$$

$$\hat{f}_\eta^{\mathbf{s}} = \sum_{j=1}^{p} \hat{f}_{\eta,j}^{\mathbf{s}} = \underset{f = \sum_{j=1}^p f_j, f_j \in \mathcal{H}_{K_j}}{\arg\min} \{\sum_{t=1}^{T} s_t \ell_\sigma(y^t - \sum_{j=1}^{p} f_j(x_j^t)) + \eta \sum_{j=1}^{p} \tau_j \|f_j\|_{K_j}^2\},$$

and

$$\hat{f}^{\mathbf{s}} = \sum_{j=1}^{p} \hat{f}_j^{\mathbf{s}} = \underset{f = \sum_{j=1}^p f_j, f_j \in \mathcal{H}_{K_j}}{\arg\min} \{\sum_{t=1}^{T} s_t \ell_\sigma(y^t - \sum_{j=1}^{p} f_j(x_j^t)) + \lambda\Omega(f)\}.$$

Correspondingly, we denote by $\mathcal{E}_T^{\sigma,\mathbf{s}}(f) = \sum_{t=1}^T s_t \ell_\sigma(f(x^t) - y^t)$.

Proposition 5. *Let Assumptions 1-2 be true. For any $f \in \bar{\mathcal{H}}_Z$, there holds*

$$\|f - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \leq E_1 + E_2 + E_3 + \mathcal{E}_T^{\sigma}(f_T^*) + \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^*\|^2_{K_j} + \frac{2M_c}{\sigma^c},$$

where

$$E_1 = \mathcal{E}_{T+1}^{\sigma}(\hat{f}^{\mathbf{s}}) - \mathcal{E}_T^{\sigma,\mathbf{s}}(\hat{f}^{\mathbf{s}})$$

$$E_2 = \mathcal{E}_T^{\sigma,\mathbf{s}}(\hat{f}^{\mathbf{s}}) + \lambda\Omega(\hat{f}^{\mathbf{s}}) - \mathcal{E}_T^{\sigma,\mathbf{s}}(\hat{f}_{\eta}^{\mathbf{s}}) - \eta \sum_{j=1}^{p} \tau_j \|\hat{f}_{\eta,j}^{\mathbf{s}}\|^2_{K_j}$$

and

$$E_3 = \mathcal{E}_{T+1}(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}(f^*) + \eta \sum_{j=1}^{p} \tau_j \|f_{\eta,T+1,j}^*\|^2_{K_j}.$$

Proof. According to Lemma 2, for any $f \in \bar{\mathcal{H}}_Z$, we can make following error decomposition

$$\begin{aligned}
& \|f - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \\
= \;& \mathcal{E}_{T+1}(f) - \mathcal{E}_{T+1}(f^*) \\
\leq \;& \mathcal{E}_{T+1}^{\sigma}(f) - \mathcal{E}_{T+1}^{\sigma}(f^*) + \frac{M_c}{\sigma^c} \\
\leq \;& \underbrace{\{\mathcal{E}_{T+1}^{\sigma}(f) - \mathcal{E}_T^{\sigma,\mathbf{s}}(f)\}}_{E_1} + \{\mathcal{E}_T^{\sigma,\mathbf{s}}(f) - \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}})\} + \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}}) + \mathcal{E}_{T+1}^{\sigma}(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}^{\sigma}(f^*) \\
& \qquad + \eta \sum_{j=1}^{p} \tau_j \|f_{\eta,T+1,j}^*\|^2_{K_j}\} + \frac{M_c}{\sigma^c} \\
\leq \;& E_1 + \{\mathcal{E}_T^{\sigma,\mathbf{s}}(f) + \lambda\Omega(f) - \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}}) - \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^{*,\mathbf{s}}\|^2_{K_j}\} + \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}}) + \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^{*,\mathbf{s}}\|^2_{K_j} \\
& \qquad + \mathcal{E}_{T+1}^{\sigma}(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}^{\sigma}(f^*) + \eta \sum_{j=1}^{p} \tau_j \|f_{\eta,T+1,j}^*\|^2_{K_j}\} + \frac{M_c}{\sigma^c} \\
\leq \;& E_1 + \underbrace{\{\mathcal{E}_T^{\sigma,\mathbf{s}}(f) + \lambda\Omega(f) - \mathcal{E}_T^{\sigma,\mathbf{s}}(\hat{f}_{\eta}^{\mathbf{s}}) - \eta \sum_{j=1}^{p} \tau_j \|\hat{f}_{\eta,j}^{\mathbf{s}}\|^2_{K_j}\}}_{E_2} + \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}}) + \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^{*,\mathbf{s}}\|^2_{K_j} \\
& \qquad + \mathcal{E}_{T+1}^{\sigma}(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}^{\sigma}(f^*) + \eta \sum_{j=1}^{p} \tau_j \|f_{\eta,T+1,j}^*\|^2_{K_j}\} + \frac{M_c}{\sigma^c} \\
\leq \;& E_1 + E_2 + \underbrace{\mathcal{E}_{T+1}(f_{\eta,T+1}^*) - \mathcal{E}_{T+1}(f^*) + \eta \sum_{j=1}^{p} \tau_j \|f_{\eta,T+1,j}^*\|^2_{K_j}}_{E_3} + \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}}) + \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^{*,\mathbf{s}}\|^2_{K_j} + \frac{2M_c}{\sigma^c} \\
\leq \;& E_1 + E_2 + E_3 + \mathcal{E}_T^{\sigma,\mathbf{s}}(f_T^{*,\mathbf{s}}) + \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^{*,\mathbf{s}}\|^2_{K_j} + \frac{2M_c}{\sigma^c} \\
\leq \;& E_1 + E_2 + E_3 + \min_{f \in \mathcal{H}_Z} \{\mathcal{E}_T^{\sigma,\mathbf{s}}(f) + \lambda\Omega(f)\} + \eta \sum_{j=1}^{p} \tau_j \|f_{T,j}^{*,\mathbf{s}}\|^2_{K_j} + \frac{2M_c}{\sigma^c}.
\end{aligned}$$

□

Following the generalization analysis in (Kuznetsov & Mohri, 2020a), we need to discrepancy measure to measure the discrepancy of the non-stationarity of the stochastic process $\{Z_t\}_{t=-\infty}^{\infty}$ with respect to both the loss function ℓ_σ and the hypothesis set $\bar{\mathcal{H}}_Z$.

Definition 4. *The discrepancy describes the discrepancy between target distribution and the distribution of the sample. For any $f \in \bar{\mathcal{H}}_Z$, the discrepancy measure with respect to Huber loss is defined as*

$$\mathrm{disc}(\mathbf{s}) := \sup_{f \in \mathcal{H}_Z} \Big\{ \mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \sum_{t=1}^{T} \mathbb{E} s_t \ell_\sigma(f(x^t) - y^t) \Big\}.$$

Lemma 7. *For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $f \in \bar{\mathcal{H}}_Z$ and all $\alpha > 0$*

$$\mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1}) \leq \sum_{t=1}^{T} s_t \ell_\sigma(f(x^t) - y^t) + \mathrm{disc}(\mathbf{s}) + \|\mathbf{s}\|_2 + 6M_{\ell_\sigma}\sqrt{4\pi \log T}\mathcal{R}_T(\mathcal{G}) + M_{\ell_\sigma}\|\mathbf{s}\|_2\sqrt{8\log\frac{1}{\delta}},$$

where $\mathcal{G} := \{\ell_\sigma(f, z) : f \in \bar{\mathcal{H}}_Z\}$ and M_{ℓ_σ} is a positive constant such that $\ell_\sigma(f(x^t) - y^t) \leq M_{\ell_\sigma}$ for any $f \in \bar{\mathcal{H}}_Z$ and $t \in \mathbb{Z}$.

Proof of Theorem 2 The proofs of bounding errors E_1, E_2 and E_3 proceeds similarly to the proof of Theorem 1 and are omitted for brevity.

According to Lemma 3, we have

$$E_1 \leq \mathrm{disc}(\mathbf{s}) + \|\mathbf{s}\|_2 + 6M_{\ell_\sigma}\sqrt{4\pi \log T}\mathcal{R} + M_{\ell_\sigma}\|\mathbf{s}\|_2\sqrt{8\log\frac{1}{\delta}},$$

where

$$R = \|\mathbf{s}\|_2[32\sigma[1 + 4\sqrt{2}\log^{3/2}(eT^2)]M^2 p^{\frac{1}{2}}\kappa].$$

Moreover, from Proposition 4, we can get

$$E_3 \leq (p+1)\sum_{j=1}^{p}(\eta\tau_j)^{2r}\|L_{K_j,T+1}^{-r} f_j^*\|_2^2.$$

Similarly, to bound the hypothesis error

$$E_2 = \{\mathcal{E}_T^{\sigma,\mathbf{s}}(\hat{f}^{\mathbf{s}}) + \lambda\Omega(\hat{f}^{\mathbf{s}}) - \mathcal{E}_T^{\sigma,\mathbf{s}}(\hat{f}_\eta^{\mathbf{s}}) - \eta\sum_{j=1}^{p}\|\hat{f}_{\eta,j}^{\mathbf{s}}\|_K^2\},$$

we first give the properties of $\hat{f}$ and then use them to bridge $\hat{f}$ and f_η.

Lemma 8. *For all $j = 1, ..., p$, there holds*

$$E_2 \leq \frac{2\lambda p\sigma\|\mathbf{s}\|_2}{\eta}.$$

Proof. Recall the represent theorem which ensures that

$$\hat{f}_\eta^{\mathbf{s}} = \sum_{j=1}^{p}\sum_{t=1}^{T}\alpha_{tj}^{\eta,\mathbf{s}} K_j(x_j^t, \cdot),\ \alpha_{tj}^{\eta,\mathbf{s}} \in \mathbb{R}.$$

For notation simplicity, denote $\alpha_j^{\eta,\mathbf{s}} = (\alpha_{1j}^{\eta,\mathbf{s}}, ..., \alpha_{Tj}^{\eta,\mathbf{s}})' \in \mathbb{R}^T$ and $\boldsymbol{\alpha}^{\eta,\mathbf{s}} = ((\alpha_1^{\eta,\mathbf{s}})', ..., (\alpha_p^{\eta,\mathbf{s}})')' \in \mathbb{R}^{Tp}$. We then deduce that

$$\boldsymbol{\alpha}^{\eta,\mathbf{s}} = \arg\min_{\alpha \in \mathbb{R}^{Tp}}\{\sum_{t=1}^{T} s_t \ell_\sigma(y^t - \sum_{j=1}^{p}\boldsymbol{K}_{jt}'\alpha_j) + \eta\sum_{j=1}^{p}\tau_j(\alpha_j^\eta)'\boldsymbol{K}_j\alpha_j\},$$

where $\boldsymbol{K}_j = \{K(x_j^s, x_j^i)\}_{i,s=1}^n \in \mathbb{R}^{T\times T}$. We get

$$\sum_{t=1}^{T} s_t \ell_\sigma'(y^t - \sum_{j=1}^{p}\boldsymbol{K}_{jt}'\alpha_j^{\eta,\mathbf{s}})\boldsymbol{K}_{ji}' = \eta\tau_j\boldsymbol{K}_j\alpha_j^{\eta,\mathbf{s}}.$$

It is easy to deduce that

$$\alpha_j^{\eta,\mathbf{s}} = \frac{1}{\eta\tau_j}(s_1\ell'_\sigma(y^1 - \sum_{j=1}^p \boldsymbol{K}'_{j1}\alpha_j^\eta), ..., s_T\ell'_\sigma(y^T - \sum_{j=1}^p \boldsymbol{K}'_{jT}\alpha_j^\eta))'.$$

Then we obtain that, for any $j = 1, ..., p$,

$$\tau_j\|\alpha_j^\eta\|_2 = \frac{1}{\eta}\sqrt{\sum_{t=1}^T s_t^2(\ell'_\sigma(y^t - \sum_{j=1}^p \boldsymbol{K}'_{jt}\alpha_j^\eta))^2} \leq \frac{2\sigma\|\mathbf{s}\|_2}{\eta}$$

This completes the proof. □

Under Assumptions 1-2, there holds

$$\begin{aligned}
E_2 &= \{\mathcal{E}_T^\sigma(f) + \lambda\Omega(\hat{f}^{\mathbf{s}}) - \mathcal{E}_T^\sigma(\hat{f}_\eta^{\mathbf{s}}) - \eta\sum_{j=1}^p \|\hat{f}_{\eta,j}^{\mathbf{s}}\|_K^2\} \\
&= \mathcal{E}_T^\sigma(\hat{f}^{\mathbf{s}}) + \lambda\Omega(\hat{f}^{\mathbf{s}}) - \mathcal{E}_T^\sigma(\hat{f}_\eta^{\mathbf{s}}) - \lambda\Omega(\hat{f}_\eta^{\mathbf{s}}) - \eta\sum_{j=1}^p \|\hat{f}_{\eta,j}^{\mathbf{s}}\|_K^2 + \lambda\Omega(\hat{f}_\eta^{\mathbf{s}}) \\
&\leq \lambda\Omega(\hat{f}_\eta^{\mathbf{s}}) = \lambda\sum_{j=1}^p \tau_j\|\alpha_j^{\eta,\mathbf{s}}\|_2 \\
&\leq \frac{2\lambda p\sigma\|\mathbf{s}\|_2}{\eta}.
\end{aligned}$$

By combining the above results, we have with confidence $1-\delta$

$$\begin{aligned}
&\|\hat{f}^{\mathbf{s}} - f^*\|_{L_2(\rho_{\mathcal{X}}^{T+1})}^2 \\
\leq\ & \mathrm{disc}(\mathbf{s}) + \min_{f\in\mathcal{H}_Z}\{\mathcal{E}_T^{\sigma,\mathbf{s}}(f) + \lambda\Omega(f)\} \\
&+\widetilde{C}\log(1/\delta)(\|\mathbf{s}\|_2 + \|\mathbf{s}\|_2\sigma + \eta^{2r} + \lambda\eta^{-1}\|\mathbf{s}\|_2\sigma + \sigma^{-c}\lambda^{-2}T^{-2} + \sigma^{-c}\lambda^{-1}T^{-1} + \eta),
\end{aligned}$$

where $\widetilde{C}$ is a positive constant independently of $T, \lambda, \eta, \delta, \sigma$ and p. By taking $\sigma = T^{\frac{1}{2c}}$, $\lambda = T^{-1}$ and $\eta = T^{-\frac{1}{2}}$, we completes the proof.

D Proof of Theorem 3

According to the definition of $\mathrm{disc}(T+1)$, we have

$$\begin{aligned}
\mathrm{disc}(\mathbf{s}) &= \sup_{f\in\mathcal{H}_Z}\Big\{\mathbb{E}\ell_\sigma(f(x^{T+1}) - y^{T+1}) - \sum_{t=1}^T s_t\mathbb{E}\ell_\sigma(f(x^t) - y^t)\Big\} \\
&\leq \sup_{f\in\mathcal{H}_Z}[\sum_{t=1}^T q_t^*\mathbb{E}\ell_\sigma(f(x^t) - y^t) - \sum_{t=1}^T s_t\mathbb{E}\ell_\sigma(f(x^t) - y^t)] + \mathrm{disc}(\mathbf{q}^*).
\end{aligned}$$

Furthermore, it is easy to deduce that

$$\begin{aligned}
&\sup_{f\in\mathcal{H}_Z}\sum_{t=1}^T(q_t^* - s_t)\mathbb{E}\ell_\sigma(f(x^t) - y^t) - \sup_{f\in\mathcal{H}_Z}\sum_{t=1}^T(q_t^* - s_t)\ell_\sigma(f(x^t) - y^t) \\
\leq\ & \sup_{f\in\mathcal{H}_Z}\sum_{t=1}^T(q_t^* - s_t)[\mathbb{E}\ell_\sigma(f(x^t) - y^t) - \ell_\sigma(f(x^t) - y^t)].
\end{aligned}$$

Proof of Theorem 3 According to the proof of Theorem 1 in Kuznetsov & Mohri (2020a), we have

$$\sup_{f\in\mathcal{H}_Z}\sum_{t=1}^T(q_t^* - s_t)[\mathbb{E}\ell_\sigma(f(x^t) - y^t) - \ell_\sigma(f(x^t) - y^t)] \leq \frac{1}{\sqrt{T}} + 6M_{\ell_\sigma}\sqrt{4\pi\log T}\mathcal{R}_T(\mathcal{G}) + M_{\ell_\sigma}\|\mathbf{q}^* - \mathbf{s}\|_2\sqrt{8\log\frac{1}{\delta}}$$

By combining above results, we have with confidence $1-\delta$

$$\begin{aligned}
&\|\hat{f}^{\mathbf{s}} - f^*\|^2_{L_2(\rho_{\mathcal{X}}^{T+1})} \\
\leq\ & \text{disc}(\mathbf{q}^*) + \sup_{f\in\mathcal{H}_Z} \sum_{t=1}^{T} (q_t^* - s_t)\ell_\sigma(f(x^t) - y^t) + \min_{f\in\mathcal{H}_Z} \{\mathcal{E}_T^{\sigma,\mathbf{s}}(f) + \lambda\Omega(f)\} \\
&+\widetilde{C}\log(1/\delta)(\|\mathbf{q}^* - \mathbf{s}\|_2 + \|\mathbf{s}\|_2 + \|\mathbf{s}\|_2\sigma + \eta^{2r} + \lambda\eta^{-1}\|\mathbf{s}\|_2\sigma + \sigma^{-c}\lambda^{-2}T^{-2} + \sigma^{-c}\lambda^{-1}T^{-1} + \eta).
\end{aligned}$$

By taking $\sigma = T^{\frac{1}{2c}}$, $\lambda = T^{-1}$ and $\eta = T^{-\frac{1}{2}}$, we complete the proof.

Algorithm 1: Optimization procedure for adaptive SpHAM

Input: Data $\{(x^t, y^t)\}_{t=1}^T$, Max-Iter $Z \in \mathbb{Z}$, Mercer kernel $K_j, j = 1, ..., p$ with bandwidth d, Weights $\tau_l, l = 1, ..., p$, $\mathbf{q}^*$.
Initialization: Lipschitz constant L, $\mathbf{s}^0$.
Step A: Computing weights $\hat{\mathbf{s}}$:
for $z = 1, ..., Z$ **do**
  1. Compute A^{z-1} via DC-programming (or gradient descent method);
  2. Update $\mathbf{s}^z$ via (12).
Output: $\hat{\mathbf{s}} = \mathbf{s}^Z$.
Step B: Computing $\hat{f}^{\mathbf{s}}$:
for $z = 1, ..., Z$ **do**
  1): Compute $\alpha^z = p_L(\beta^z)$ via (14);
  2): $m_{z+1} = \frac{1+\sqrt{1+4m_z^2}}{2}$;
  3): $\beta^{z+1} = \alpha^z + \frac{m_z - 1}{m_{z+1}}(\alpha^z - \alpha^{z-1})$.
Output: $\alpha^{\hat{\mathbf{s}}} = \alpha^Z$;
Prediction function: $\hat{f}^{\mathbf{s}} = \sum_{j=1}^P \sum_{t=1}^T \alpha_{tj}^{\hat{\mathbf{s}}} K_j(x_j^t, \cdot)$;
Variable selection: $\{j : \|\alpha_j^{\mathbf{s}}\|_2 \geq v, j = 1, ..., p\}$.

E Experiment Optimization

The optimization problem (10) reduces to problem 7 when taking $\mathbf{s}^* = \frac{1}{T}\mathbb{I}_T$. Recall the optimization problem in **Step A**:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \{ \sup_{f\in\mathcal{H}_Z} \sum_{t=1}^{T} (q_t^* - s_t)\ell_\sigma(f(x^t) - y^t) + \frac{\lambda_1}{2}\|\mathbf{q}^* - \mathbf{s}\|_2^2 + \frac{\lambda_2}{2}\|\mathbf{s}\|_2^2 T^{\frac{1}{2c}} \}.$$

The above optimization problem can be equivalently rewritten as a common type of bilevel optimization problem (Colson et al., 2007), i.e,

Outer problem: $\min_{\mathbf{s}} \sum_{t=1}^T (q_t^* - s_t)\ell_\sigma(f_{\mathbf{s}}(x_t) - y_t) + \lambda_1\|\mathbf{q}^* - \mathbf{s}\|_2^2 + \lambda_2 T^{1/2c}\|\mathbf{s}\|_2^2$,

Inner problem: $f_{\mathbf{s}} = \arg\max_{f\in\mathcal{H}_{\mathbf{z}}} \sum_{t=1}^T (q_t^* - s_t)\ell_\sigma(f(x_t) - y_t)$,

where the outer (min) problem parameterized by $\mathbf{s}$, is nested within the inner (max) problem. The outer problem can be solved by standard gradient descent, where in each step, we need to optimize the inner (max) problem with last updated $\mathbf{s}$.

We denote by k the iteration time. Then for $k+1$-th iteration, we have the following gradient update rule

$$\mathbf{s}^{k+1} = s^k - \gamma(A^k - \lambda_1(\mathbf{q}^* - \mathbf{s}^k) + \lambda_2 T^{\frac{1}{2c}}\mathbf{s}^k), \tag{12}$$

where γ is learning rate,

$$A^k = -\Big(\ell_\sigma(\sum_{t=1}^T\sum_{j=1}^p \alpha_{tj}^{\mathbf{s}^k} K_j(x_j^t, x_j^1) - y^1), ..., \ell_\sigma(\sum_{t=1}^T\sum_{j=1}^p \alpha_{tj}^{\mathbf{s}^k} K_j(x_j^t, x_j^T) - y^T)\Big)' \in \mathbb{R}^T$$

and

$$\alpha^{\mathbf{s}^k} = (\alpha_{11}^{\mathbf{s}^k}, ..., \alpha_{T1}^{\mathbf{s}^k}, ..., \alpha_{1p}^{\mathbf{s}^k}, ..., \alpha_{Tp}^{\mathbf{s}^k})' \in \mathbb{R}^{Tp}$$

is obtained by the following weighted optimization problem

$$\alpha^{\mathbf{s}^k} = \arg\max_{\alpha} \sum_{t=1}^{T} (q_t^* - s_t^k) \ell_\sigma (\sum_{m=1}^{T} \sum_{j=1}^{p} \alpha_{mj} K_j(x_j^m, x_j^t) - y^t).$$

Note that the inner problem is subjected to $f \in \mathcal{H}_K$, i.e., $\|\alpha\|_2 = \sum_{j=1}^{p} \tau_j \|\alpha_j\|_2 \leq \frac{2M^2}{\lambda T \min_{j=1,...,p} \tau_j}$. A widely-used method for solving this constrained inner problem is DC programming (Tao & An, 1998). For simplicity, we here transform this inner problem into a regularized problem with ℓ_2-norm regularizer, and solve this regularized problem by standard gradient method. After obtaining the solution $\hat{\mathbf{s}}$, we turn to solve the following weighted optimization problem in **Step B**:

$$\alpha^{\hat{\mathbf{s}}} = \arg\min_{\alpha_j \in \mathbb{R}^T, j=1,...,p} \{\sum_{t=1}^{T} \hat{s}_t \ell_\sigma (y^t - \sum_{j=1}^{p} (\mathbf{K}_j^t)' \alpha_j) + \lambda \sum_{j=1}^{p} \tau_j ||\alpha_j||_2\}. \tag{13}$$

We can see that this problem contains non-smooth function $\lambda \sum_{j=1}^{p} \tau_j ||\alpha_j||_2$ which makes the standard gradient descent method inapplicable. To conquer this challenge, we leverage fast iterative shrinkage-thresholding algorithm (FISTA)(Beck & Teboulle, 2009). Our optimization problem becomes

$$\begin{aligned} \alpha^k = p_L(\alpha^{k-1}) &:= \arg\min_{\alpha} \Big\{ \lambda \sum_{j}^{p} \tau_j ||\alpha_j||_2 + \frac{L}{2} ||\alpha - (\alpha^{k-1} - \frac{1}{L} \sum_{t=1}^{T} \hat{s}_t \nabla \ell_\sigma (y^t - (K^t)' \alpha^{k-1})||^2 \Big\} \\ &= \left(\left(1 - \frac{\lambda}{||v_j||_2} \right)_+ v_j \right)_{1 \leq j \leq p}, \end{aligned} \tag{14}$$

where

$$v_j = \alpha_j^{k-1} - \frac{1}{L} \sum_{t=1}^{T} \hat{s}_t \nabla_{\alpha_j} \ell_\sigma (y^t - (K^t)' \alpha^{k-1}))$$

and

$$\nabla_{\alpha_j} \ell_\sigma (y^t - (K^t)' \alpha^{k-1})) = \begin{cases} 2((K^t)'\alpha - y^t)(K_j^t)' & |(K^t)'\alpha - y^t| < \sigma \\ 2\sigma & (K^t)'\alpha - y^t \geq \sigma \\ -2\sigma & (K^t)'\alpha - y^t \leq -\sigma \end{cases}$$

and $L = \max \frac{1}{T} \sum_{t=1}^{T} \|(K^t)' K^t\|$ is the lipschitz constant $\frac{1}{T} \sum_{t=1}^{T} \nabla \ell_\sigma (y^t - (K^t)' \alpha^{k-1})$. Denote Proj(s) as a projection of s and $v > 0$ as the threshold value. To focus on the weights of samples which are really useful for forecasting, we consider Proj(s) as a box projection such that $s_t > q_t^*, \forall t = 1, ..., p$). Finally, the optimization procedure for adaptive SpHAM can be summarized in Algorithm 1. Note that if we only run **Step B** with $\hat{s}_t = \frac{1}{T}, \forall t = 1, ..., T$, we can further obtain the optimization procedure for SpHAM.

Remark 3. *The computational complexity of Algorithm 1 depends on the optimization strategy for DC programming, the training size T, dimension p and the iteration times Z. We denote by the computational complexity of DC programming $O(DC(T, p))$. Then, the computational complexity of Step A is $O(ZDC(T, p) + ZT)$ and the computational complexity of Step B is $O(ZT^3 p^3)$. Thus, the total computational complexity of Algorithm 1 is $O(ZO(DC) + ZT^3 p^3)$. For large scale data, we can further speed up Algorithm 1 by random Fourier features technique [Rahimi and Recht 2007], which is leaved for future work. This has been carefully discussed in the revised manuscript.*

F Additional Experiment

F.1 Evaluation on benchmark data

We test our algorithm on nonlinear dataset from CauseMe. The hyper-parameter selection is the same as the one in Section 4. The results in Table 3 verify the effectiveness of our method.

Table 3: The ASE on CauseMe data (p refers to the dimension of features).

Methods	$(p = 3, T = 300)$	$(p = 5, T = 300)$
LSTM	0.7681	**0.9230**
TS−SpAM	0.7782	0.9485
SpHAM	**0.7548**	0.9484

F.2 EXPERIMENTS ON AIR QUALITY DATASET

We use the Air Quality dataset obtained from UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/Air+quality`) to test our model's ability to detect the Granger causality. This dataset includes 9358 hourly air quality data in an Italian city, collected from March 2004 to February 2005. The details of the dataset can be obtained on the UCI website. The Granger causal network we detect is shown in Figure 3. Given the network, we observe that the temperature (T) influences ozone (O_3) and nitrogen dioxide (NO_2) which is validated in Kalisa et al. (2018). Mwaniki et al. (2014) confirm the relationship between relative humidity (RH) and nitrogen dioxide (NO_2). Yan et al. (2018) verify the relationship between humidity and nitrogen dioxide (NO_2).

F.3 EXPERIMENTS ON CORONAL MASS EJECTIONS DATASET

Coronal Mass Ejections (CMEs) are the most violent eruptions in the Solar System. Despite machine learning approaches have been applied to these tasks recently Wang et al. (2019); Liu et al. (2018), there is no any work for interpretable prediction with Granger causal network. CMEs data are provided in The Richardson and Cane List (`http://www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm`). From this link, we collect 152 ICMEs observations from 1996 to 2016. The features of CMEs are provided in SOHO LASCO CME Catalog (`https://cdaw.gsfc.nasa.gov/CME_list/`). In-situ solar wind parameters can be downloaded from OMNIWeb Plus (`https://omniweb.gsfc.nasa.gov/`). A total of 9 features are chosen as input, including: (1) Central PA (CPA), (2) Angular Width, (3) three approximated speeds (Linear Speed, 2nd-order Speed at final height,and 2nd-order Speed at 20 Rs), (4) Mass, (5) Kinetic Energy, (6) MPA and (7)CMEs arrival time. Figure 4 shows that Granger causal network when the output is CMEs Arrival time. Some interesting findings are concluded from this Granger causal network. For instance, Speed and Mass, as the significant variables causing CMEs arrival time, have been also screened out in Liu et al. (2018). Morover, the CMEs Angular Width does not cause the CMEs arrival time forecasting, while Liu et al. (2018) state that they have a significant correlation. This indicates that the CMEs arrival time is affected by the Angular Width at current time, but not by the historical Angular Width.

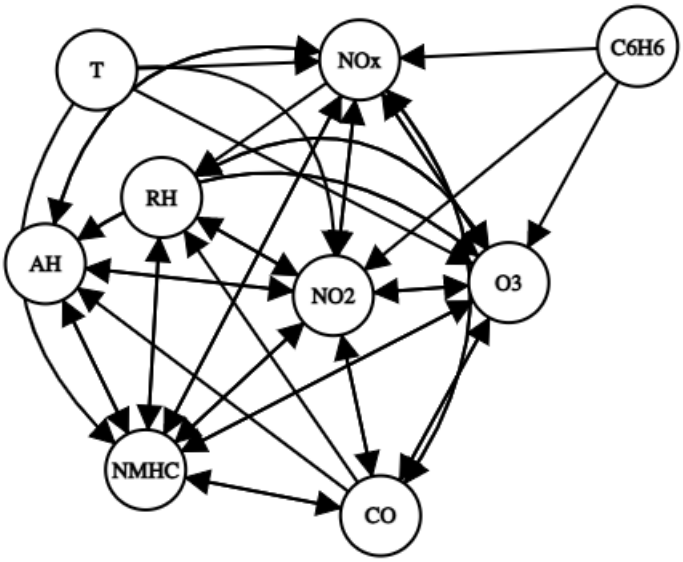

Figure 3: The Granger causal network on Air Quality dataset.

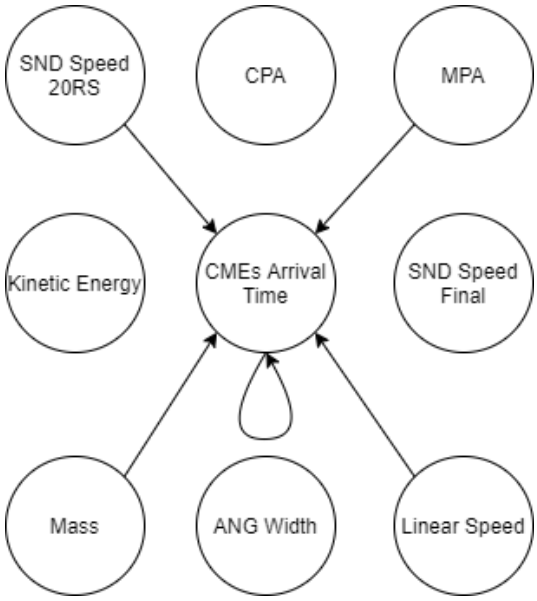

Figure 4: The Granger causal network on CMEs dataset.