HUBER ADDITIVE MODELS FOR NON-STATIONARY TIME SERIES ANALYSIS

Anonymous authors
Paper under double-blind review

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 and inference (e.g., Granger causal discovery) 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.

1 INTRODUCTION

Additive models have 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 theory exploration (Huang & Yang, 2004; Wang & Yang, 2007; Chu & Glymour, 2008; Song & Yang, 2010; Yang et al., 2018) and practical applications (Dominici et al., 2002; Wang & Brown, 2011; Ravindra et al., 2019; Bussmann et al., 2020). Although the aforementioned works have shown promising behaviours, the methods proposed in these works require some strict assumptions on the stochastic process, e.g., Gaussian innovation, mixing dependency or stationary distribution. However, these formerly mentioned approaches may have degraded performance when facing heavy-tailed and non-stationary time series (Raya & Murad, 1998; Qiu et al., 2015).

A number of attempts have been made to relax the strict assumptions on the stochastic process. Qiu et al. (2015) develops 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 mixing conditions are commonly adopted in many previous studies of non-i.i.d settings (Doukhan, 1994), which ensure the use of traditional complexity tools (e.g., Rademacher complexity (Bartlett & Mendelson, 2002; Mohri & Rostamizadeh, 2009), covering number (Ron, 2000; Zhou, 2002; Guo & Shi, 2011) and stability (Mohri & Rostamizadeh, 2010)) for theoretical exploration. However, the mixing and stationary conditions are too strict and not always valid. For instance, long memory models (e.g., ARFIMA) may not be mixing (Baillie, 1996; Kuznetsov & Mohri, 2020). To relax the mixing and stationary conditions, Adams & Nobel (2010) prove asymptotic guarantees for stationary ergodic sequences. Alekh & C. (2013) give 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 (2020) provide data-dependent generalization bounds for the non-stationary and non-mixing stochastic processes.

However, both theoretically and practically, the exploration on robust additive models for (non)stationary time series analysis is still limited. In this paper, we propose a class of sparse Huber additive models with statistical guarantees. Our main contributions are summarized as below:

- *Algorithm design*: We first propose a novel sparse Huber additive model (SpHAM) by integrating Huber loss and sparsity-inducing $\ell_{2,1}$-norm regularizer into an additive data-dependent hypothesis space. For stationary time series, this proposed algorithm can achieve robust forecasting and satisfactory inference (e.g., variable selection and Granger causal discovery) simultaneously, even under non-Gaussian innovation, e.g., heavy-tailed innovation and outliers.
- *Theoretical guarantees*: We establish the function approximation error bounds of SpHAM by developing error decomposition technique and employing sequential Rademacher complexity (Rakhlin et al., 2010; 2015; Kuznetsov & Mohri, 2020). With properly selected scale parameter, the theoretical findings indicate that: a) for stationary time series, the convergence rate $O(n^{-\frac{1}{2}})$ can be achieved, even when the innovation is non-Gaussian distribution. In fact, this error bound appears to be novel even in the stationary case because the mixing conditions are not imposed here; b) for non-stationary time series, the approximation error of our method is bounded by a discrepancy measure for the shifts of the data distributions along with time (see Theorems 2-3 for more details), which inspires us to further propose an adaptive SpHAM by penalizing such a discrepancy measure.
- *Optimization and empirical evaluations*: The proposed SpHAM and adptive SpHAM can be implemented efficiently by Difference of Convex programming (DC programming) (Tao & An, 1998) and 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: We compare our method with related works and show improvement from the following aspects (the characteristics of our method are in bold):

a) ***SpHAM for non-i.i.d data*** *vs. Huber algorithm for i.i.d data:* Huber algorithms play an important role in robust estimation and inference (Huber, 1964; Huber & Ronchetti, 2009; Loh, 2017; Feng & Wu, 2020). However, the most existing methods concern i.i.d data. Our method fills the gap in the investigation of whether Huber additive models work when data are non-i.i.d with non-Gaussian innovation.

b) ***Robust forecasting*** *vs. non-robust forecasting:* Under Gaussian innovation, there has been a lot of studies on sparse additive models (Wang & Yang, 2007; Chu & Glymour, 2008; Yang et al., 2018; Bussmann et al., 2020). Differently, our method benefits from the advantageous properties of Huber algorithms in robust forecasting, and the corresponding theoretical findings in Theorem 1 verify the convergence of our method in weak moment condition.

c) ***Nonstationary data*** *vs. stationary data:* In order to develop widely used complexity tools (e.g., Rademacher complexity and covering number) for non-i.i.d case, early proposed methods assume that the data satisfies (strict) stationary or various mixing conditions (Doukhan, 1994; Qiu et al., 2015; Yang et al., 2018). To overcome these limitations, our method employs sequential Rademacher complexity in theoretical analysis, and the corresponding theoretical results inspire us to modify our method to deal with non-stationary data.

d) ***Function approximation analysis*** *vs. generalization analysis:* Recently, Kuznetsov & Mohri (2020) support a class of methods for nonstationary time series with generalization error analysis. In contrast, we are chiefly concerned about the function approximation error analysis, since the convergence of generalization cannot imply the convergence of function approximation for Huber loss based algorithms (Sun et al., 2019; Feng & Wu, 2020).

e) ***Interpretable*** *vs. non-interpretable:* Compared with the smoothness-inducing regularized methods Kuznetsov & Mohri (2020); Feng & Wu (2020), the sparsity-inducing regularizer ensures that our method can screen out informative variables or detect the Granger causal among variables.

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-SpA) (Yang et al.,

Table 1: The properties of related methods

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

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

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 larger number of scenarios commonly used in practice. For instance, the case X^t 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 information of multiple variables at different lagged times.

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 as

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

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 if our focus is on nonlinear relationships (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 most existing methods to learn f^* are usually integrating 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. Some existing works construct the additive hypothesis space $\mathcal{H}$ via spline-based basis (Yang et al., 2018) and composite neural net (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. In statistic learning community, as one commonly used robust statistic, 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

In this section, we first propose a sparse Huber additive model (SpHAM). Secondly, we conduct some assessments of SpHAM from a statistical learning viewpoint. Finally, with the help of our theoretical findings, we give an adaptive SpHAM for non-stationary time series forecasting.

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 $K_j:\mathcal{X}_j\times\mathcal{X}_j\to\mathbb{R}$. Then 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) and kernel norm into an additive RKHS, the kernel-based Huber additive model 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 model with sparsity, we consider the following sparsity-inducing penalty

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

in an additive data dependent hypothesis space

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

Clearly, this data dependent hypothesis spaces is the additive RKHS depending on the observations $\{(x^t,y^t)\}_{t=1}^T$. 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 transposition of α_j for avoiding confusion with time T. 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 due to the space limitation.

Remark 1. *The proposed method can be easily extended to the group sparse setting by replacing the direct decomposition $\{\mathcal{X}_j\}_{j=1}^p$ with the subgroups decomposition $\{\mathcal{X}_j\}_{j=1}^d$, where each $\mathcal{X}_j$ refers to the component input space concerning the interactions among variables ((Yin et al., 2012; Lin & Zhang, 2007)) or lagged times ((Nicholson et al., 2020)).*

3.2 Asymptotic theory analysis

In this section, we provide the theoretical analysis of SpHAM. It is natural and necessary to investigate the following three problems:

- The problem of the function estimation calibration and convergence rates for stationary time series data (See Theorem 1 in Section 3.2.1);
- The problem of the function estimation calibration for non-stationary time series data (See Theorem 2 in Section 3.2.2);
- Whether the theoretical findings can inspire us to modify the SpHAM for non-stationary time series forecasting (See Theorem 3 in Section 3.3).

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

3.2.1 Function approximation analysis for stationary time series

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

Some familiar kernels satisfy the requirements of Assumption 2, e.g., Gaussian kernel, Sigmoid kernel, and Logistic kernel.

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* $g_j^* \in L_2(\rho_{\mathcal{X}_j}^{T+1}), \forall h \in \mathbb{Z}$.

The Assumption 3 is necessary to bridge the relation between the ground truth f^* and the functions in $\mathcal{H}_K$. Indeed, this assumption is a natural extension from i.i.d setting (Assumption 1 in Christmann & Zhou (2016)) to non-i.i.d time series.

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.

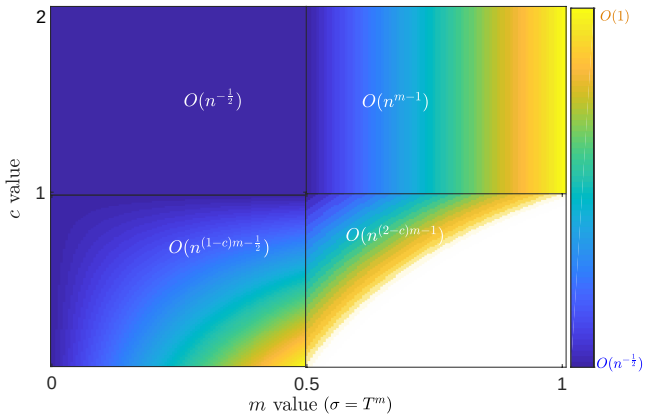

Figure 1: The convergence rates under different σ and c.

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, we see that the convergence of SpHAM can be proven 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 non-stationary 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

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

Furthermore, we introduce a discrepancy measure developed describes the discrepancy between target distribution and the distribution of the sample (Kuznetsov & Mohri, 2020).

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 measure is a natural measure 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 $\mathcal{H}_Z$.

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 non-stationary and non-mixing time series, see, e.g., Kuznetsov & Mohri (2020). Different from them, the error bound we derived is with respect to the function approximation error rather than generalization error, which is very crucial for Huber regression problem, since the convergence of generalization cannot imply the convergence of function approximation when the innovation is non-Gaussian (Sun et al., 2019; Feng & Wu, 2020).

3.3 ADAPTIVE SPARSE HUBER ADDITIVE MODEL FOR NON-STATIONARY 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, 2020), one natural and necessary assumption is that there is 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:*

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

This assumption is necessary for learning non-stationary time series, see Kuznetsov & Mohri (2020) for further discussion.

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

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

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

Note that the priori $\mathbf{q}$ can be any distribution. Also, some optimization strategies for hyper-parameter selection can also be used for finding the underlying probability set $\mathbf{q}^*$, e.g., bilevel optimization scheme (Frecon et al., 2018; Franceschi et al., 2018). We leave these extensions to future work. For simplicity, this paper assume that the distributions ρ^{T+1} does not change drastically compared with the distributions $\rho^t, t = 1, ..., T$. Thus, we consider the underlying probability $\mathbf{q}^*$ as an uniform distribution over last l observations in this paper, where $l > 0$ is a hyper-parameter that can be tuned via cross-validation. 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, where, for each step, we employ DC programming to evaluate the supremum over $f \in \mathcal{H}_Z$. 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 $K_j(u, v) = \exp(-\frac{\|u-v\|_2^2}{2d^2}), j = 1, ..., p$, is employed for constructing the additive data dependent hypothesis space, where $d > 0$ is the bandwidth. Due to limited space, we only represent the evaluations on the synthetic time series. The real-world experiments can be found in Appendix F.

4.1 Evaluation on synthetic stationary data

We consider two synthetic examples: stationary autoregressive model and nonstationary autoregressive model:

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

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

where the innovation ε^ts are i.i.d. drawn from Gaussian distribution $N(0, 1)$ and Student distribution with freedom 2, respectively.

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

$$Y^t = \frac{t}{400}\sin(Y^{t-1}) + \epsilon^t.$$

Hyperparameter selection and evaluation criterions: Recall that SpHAM algorithm requires three hyper-parameters: regularization parameter λ, bandwidth of kernel d and Huber parameter σ. For Huber parameter σ selection, we set $\sigma = T^{\frac{1}{4}}$ based on the suggestion in Theorem 1 and Corollary 1. Moreover, we make no attempt to optimize the other hyper-parameters in our method. Thus, we set the $d = 1$, $\lambda = 0.001$, $\lambda_1 = 1000$, $\lambda_2 = 30$ and $l = 60$. Suppose that there are N-size test samples. The evaluation criterions for forecasting used here contains Average Sample Error(ASE)= $\frac{1}{N}\sum_{t=1}^{N}(\hat{f}(x^t) - y^t)_2^2$ and True Deviation (TD)= $\frac{1}{N}\sum_{t=1}^{N}(\hat{f}(x^t) - f^*(x^t))_2^2$.

For each example, we generate time series with 4000 sample points. The samples at time $t = \{1500, 1501, ..., 1900\}$ are used as a training set, and the samples at time $\{1901, 1902, 1903\}$ are considered as the test set. The competing methods include standard Autoregression (AR) and TS-SpAM (Yang et al., 2018). All the evaluations are repeated for 10 times. The average results for *Example A* and *Example B* are presented in Table 2.

From the results on Example A, we see that SpHAM enjoys robust forecasting in presence of heavy-tailed t noise. Moreover, the RKHS offer our method with nonlinear forecasting, which makes our method always outperforms linear AR. The results in Example B verify the effectiveness of our method for nonstationary time series forecasting.

Table 2: The results on synthetic stationary data.

		Gaussian noise		Student noise	
	Methods	ASE	TD	ASE	TD
Example A (stationary)	AR	3.4832	3.5021	21.372	3.5426
	TS-SpAM	0.6144	**0.5845**	11.242	2.3499
	SpHAM	**0.6121**	0.5932	**10.941**	**1.7112**
Example B (nonstationary)	AR	11.624	10.915	13.334	13.012
	TS-SpAM	7.1213	5.8141	7.8315	7.1575
	SpHAM	6.8677	5.5576	9.3382	7.8111
	Adaptive SpHAM	**2.1643**	**1.4933**	**4.7256**	**1.2986**

4.2 Evaluation on benchmark data

We test our algorithm on nonlinear Nonlinear-VAR dataset from CauseMe. The hyper-parameters are the same as we use in Section 4.1, except for d which is fine-tuned by cross validation in $\{0.1, 0.5, 1, 2\}$. 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	$(p = 3, T = 300)$	$(p = 5, T = 300)$
VAR	4.1678	5.4330
TS$-$SpAM	3.8908	**4.7424**
SpHAM	**3.8240**	4.7473

5 Conclusion

We propose an adaptive sparse Huber additive model by integrating Huber loss and $\ell_{2,1}$-norm regularizer into a additive data dependent hypothesis space. We explore, both theoretically and empirically, the ability of our method for robust estimation and inference in both non-Gaussian data and(non)stationary time series data. To our best knowledge, this is the first work on sparse Huber additive models for time series analysis. Experimental results on both synthetic and real-world data validate the effectiveness of the proposed method.

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.

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

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

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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.

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.

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.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine learning (ICML)*, pp. 1563–1572, 2018.

Jordan Frecon, Saverio Salzo, and Massimiliano Pontil. Bilevel learning of the group lasso structure. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 8301–8311, 2018.

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.

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.

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, 2020.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2007.

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

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.

William B. Nicholson, Ines Wilms, Jacob Bien, and David S. Matteson. High dimensional forecasting via interpretable vector autoregression. *arXiv:1412.5250v4*, 2020.

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.

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.

Feldman Raya and Taqqu Murad. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Springer, 1998.

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.

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.

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.

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

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.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

A PROOF SKETCH

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

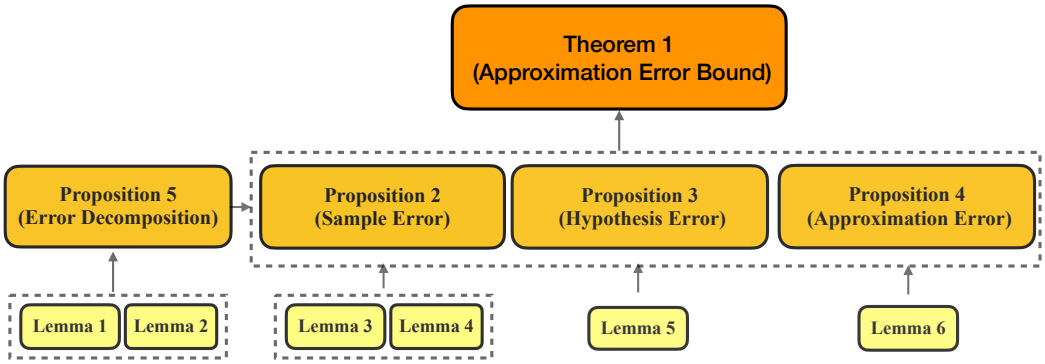

Figure 2: The important ingredients for Theorem 1.

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, 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, 2020), which is the key to measuring the ability of models for non-i.i.d data. Lemmas 3-4 give a 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*

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

and

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

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 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 of equation (1) can be bounded as

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

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 maker 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

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

□

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, 2020). 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 , 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 it to bridge $\hat{f}$ and f_η.

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

Proof. Recall the represent theorem 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. (), we deduce that

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

where $\boldsymbol{K}_j = \{K(x_j^s, x_j^i)\}_{i,s=1}^n \in \mathbb{R}^{T\times T}$. From Eq. (), we cam deduce that

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

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 , 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 1, 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 $\mathbf{s}$:

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

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

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

and

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

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, 2020), 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 1, 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 1, 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 it 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 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 , there holds

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

Combining the above results with Proposition 1, 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 , we have

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

Then combining above results with Lemma and Theorem , 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 completes 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;
  2. Update $\mathbf{s}^z$ via (12).
  3. projection: $\mathbf{s}^z = \text{Proj}(\mathbf{s}^z)$
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**:

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

This problem can be solved by standard gradient descent. In the following, 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

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

The above problem then can be solved by DC-programming .

After obtaining the solution $\hat{\mathbf{s}}$, we turn to solve the following weighted optimization problem in **Step B**:

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

We can see that this problems 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

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

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

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

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.

F EXPERIMENT

F.1 EXPERIMENTS ON REAL-WORLD TIME SERIES DATA

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 connections of the features of the data our method detect are shown in Figure 3.

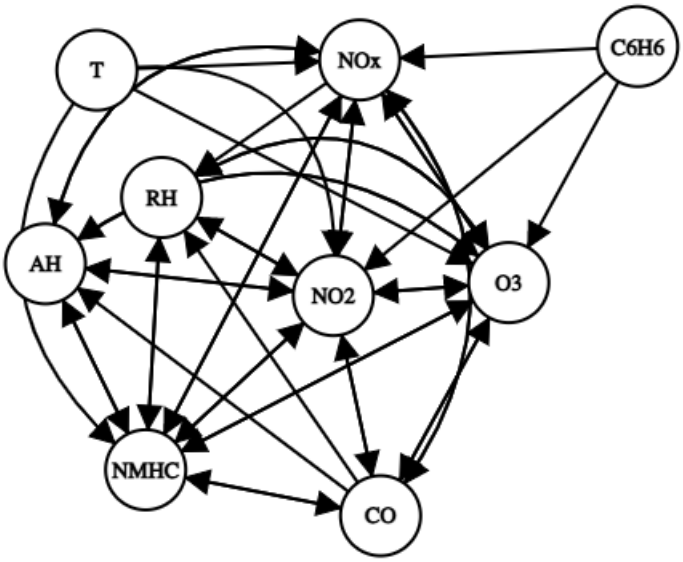

Figure 3: The result of Granger causality.