Tilted Sparse Additive Models

Yingjie Wang [1 2] **Hong Chen** [2 3] **Weifeng Liu** [1] **Fengxiang He** [4 5] **Tieliang Gong** [6] **Youcheng Fu** [2] **Dacheng Tao** [7]

Abstract

Additive models have been burgeoning in data analysis due to their flexible representation and desirable interpretability. However, most existing approaches are constructed under empirical risk minimization (ERM), and thus perform poorly in situations where average performance is not a suitable criterion for the problems of interest, e.g., data with complex non-Gaussian noise, imbalanced labels or both of them. In this paper, a novel class of sparse additive models is proposed under tilted empirical risk minimization (TERM), which addresses the deficiencies in ERM by imposing tilted impact on individual losses, and is flexibly capable of achieving a variety of learning objectives, e.g., variable selection, robust estimation, imbalanced classification and multiobjective learning. On the theoretical side, a learning theory analysis which is centered around the generalization bound and function approximation error bound (under some specific data distributions) is conducted rigorously. On the practical side, an accelerated optimization algorithm is designed by integrating Prox-SVRG and random Fourier acceleration technique. The empirical assessments verify the competitive performance of our approach on both synthetic and real data.

1. Introduction

Additive models (Hastie & Tibshirani, 1990; Stone, 1985), as natural extensions of linear models, have drawn much attention in high-dimensional data analysis due to their flexible representation and desirable interpretability. Let X be the explanatory variable that takes values in a p-dimensional metric space $\mathcal{X}$ and let Y be the corresponding real-valued response in an output set $\mathcal{Y}$. The most typical additive models were formulated by dividing the input space into p parts directly, i.e., $\mathcal{X} = (\mathcal{X}_1, ..., \mathcal{X}_p), \mathcal{X}_j \in \mathbb{R}, j \in \{1, ..., p\}$. In this setting, the hypothesis space can be stated as

$$\mathcal{H} = \Big\{ f : f(X) = \sum_{j=1}^{p} f_j(X_j), f_j \in \mathcal{H}_j \Big\},$$

where $X = (X_1, ..., X_p)^T \in \mathbb{R}^p, X_j \in \mathcal{X}_j$ and $\mathcal{H}_j$ is the component function space on $\mathcal{X}_j$. Usually, the component hypothesis spaces $\mathcal{H}_j, j = 1, ..., p$, can be reproducing kernel Hilbert space (RKHS) (Chen et al., 2017; Kandasamy & Yu, 2016; Raskutti et al., 2012; Christmann & Zhou, 2016), the space spanned by the orthogonal basis (Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Yin et al., 2012), and the space formed by neural networks (Agarwal et al., 2020; Yang et al., 2020). Over the past decades, many studies concerning the theoretical as well as practical investigations of additive models have been conducted under the empirical risk minimization (ERM) $\min_{f\in\mathcal{H}} \sum_{i=1}^{n} \ell(f(x_i), y_i)$ (Hastie & Tibshirani, 1990; Stone, 1985; Ravikumar et al., 2009; Kandasamy & Yu, 2016; Raskutti et al., 2012), where $\ell(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is the loss function. The additive models are versatile by assigning different loss functions, where the several usual ones include the squared loss $\ell(f(x), y) = (f(x) - y)^2$ for regression (Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Raskutti et al., 2012; Yin et al., 2012; Tan & Zhang, 2019), and the logistic loss (or hinge loss) for classification (Ravikumar et al., 2009; Chen et al., 2017).

However, most existing approaches are constructed under empirical risk minimization (ERM), and perform poorly in situations where average performance is not a suitable criterion for the problems of interest, e.g., data with complex non-Gaussian noise, imbalanced labels or both of them. Although several works in relation to additive models have been proposed to address the these problems by introducing

[1] College of Control Science and Engineering, China University of Petroleum (East China),Qingdao,China [2] College of Informatics, Huazhong Agricultural University, China [3] Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China [4] JD Explore Academy, JD.com, Inc., Beijing, China [5] Artificial Intelligence and its Applications Institute, School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom [6] School of Computer Science and Technology, Xi'an Jiaotong University, China [7] The University of Sydney, Sydney, Australia. Correspondence to: Hong Chen <chenh@mail.hzau.edu.cn>, Weifeng Liu <liuwf@upc.edu.cn>, Fengxiang He <F.He@ed.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

some robust metrics, e.g., quantile loss (Lv et al., 2018), modal risk (Chen et al., 2020) and huber loss (Wang et al., 2022), these methods are specially designed and can only solve one of the above problems independently.

Recently, (Li et al., 2021a) introduced a unified learning framework called tilted empirical risk minimization (TERM) as a flexible alternative to ERM, which have the ability to handle various known issues occurred in ERM, such as robustness to noise in regression/classification and class imbalance. In fact, the flexibly of TERM comes from the fact that we can tune the impact of individual losses using a scale parameter called the tilt, and thus increase or decrease the influence of outliers (or minor class), respectively. For instance, negative scale parameter is used for robust estimation/classification and positive scale parameter is used for imbalanced classification. However, in the face of some common data distributions (e.g., Gaussian distribution, skewed distribution or heavy-tailed distribution), the theoretical explorations on the approximation ability of TERM-based estimator with proper scale parameter selection strategies to the ground truth function are still insufficient.

In this paper, we try to pave a way for sparse additive models under TERM with computational feasibility and theoretical guarantees. To the best of our knowledge, this is the first work that investigates the sparse additive models under TERM. The main contributions made in this study can be summarized as follows:

- *Algorithm design:* Following the principle of TERM (Li et al., 2021a), we propose a class of new tilted sparse additive models (T-SpAM) based on Tikhonv regularization scheme with the data dependent hypothesis space and the sparsity-induced $\ell_{2,1}$-norm regularizer. The proposed method can address the deficiencies in additive models under ERM by imposing tilted impact on individual losses, and is capable of achieving a variety of learning tasks, e.g., variable selection, robust estimation, imbalanced classification or multi-objective learning (e.g., achieving robust estimation and imbalanced classification simultaneously).
- *Theoretical guarantees*. Theoretical foundations are established for T-SpAM from three aspects: 1) Generalization consistency verifies the T-SpAM can generalize well to unseen out-of-sample under some weak conditions; 2) Function approximation analysis demonstrates the estimator of T-SpAM can approach f^* with proper hyper-parameter settings under three specific data distributions (e.g., student's t distribution, skewed normal distribution and normal distribution); 3) Variable selection analysis supports that T-SpAM can identify truly informative variables under proper parametric conditions.
- *Computing Feasibility*. The proposed T-SpAM can be computationally effective by employing Prox-SVRG (Reddi et al., 2016) for non-convex objective and utilizing random Fourier features technique for speeding up the matrix calculation (Rahimi & Recht, 2007). Moreover, empirical evaluations on both synthetic and real-world data verify the effectiveness of our T-SpAM.

Related works: The principle of TERM was originally introduced in (Howard & Matheson, 1972) and revisited in reinforcement learning (Borkar, 2002; Osogami, 2012). Recently, in (Li et al., 2021a; Lee et al., 2020), TERM has shown its appealing robustness and scalability to a variety of learning tasks. Although rich empirical evaluations are stated respectively in (Li et al., 2021a), there is no enough attention on the statistical learning theory of TERM. This paper tries to establish the learning theory foundations of T-SpAM on generalization bound, function approximation and variable selection consistency. To better highlight the novelty of T-SpAM, its algorithmic properties are summarized in Table 1, where the competitors include SpAM (Sparse additive models (Ravikumar et al., 2009)), SAQR (Sparse Additive Quantile Models (Lv et al., 2018)), SpMAM (Sparse Modal Additive Models (Chen et al., 2020)) and TERM (Ravikumar et al., 2009).

2. Tilted Sparse Additive Models

This paper chooses a RKHS to form the additive hypothesis space. Let $K_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$ be a symmetric and positive definite kernel on $\mathcal{X}_j \times \mathcal{X}_j$, and $\mathcal{H}_{K_j}$ be the corresponding RKHS on $\mathcal{X}_j$ with norm $\|\cdot\|_{K_j}, j \in \{1, ..., p\}$. The RKHS with additive structure is given by

$$\mathcal{H}_K = \{\sum_{j=1}^{p} f_j : f_j \in \mathcal{H}_{K_j}, j = 1, ..., p\}$$

with

$$\|f\|_K^2 = \inf\{\sum_{j=1}^{p} \|f_j\|_{K_j}^2 : f = \sum_{j=1}^{p} f_j\}.$$

To formulate T-SpAM, we introduce the TERM scheme (Li et al., 2021a; Lee et al., 2020).

Definition 2.1. Given $\mathbf{z} := \{(x_i, y_i)\}_{i=1}^n$ drawn independently from an intrinsic distribution ρ, the TERM is defined as

$$\mathcal{R}_{\mathbf{z}}(t, f) := \frac{1}{t} \log(\frac{1}{n} \sum_{i=1}^{n} e^{t\ell(f(x_i), y_i)}), \quad (1)$$

and its population version is

$$\mathcal{R}(t, f) := \frac{1}{t} \log \int_{\mathcal{Z}} e^{t\ell(f(x), y)} d\rho(x, y),$$

where $t \in (-\infty, 0) \cup (0, +\infty)$ is the scale parameter.

Table 1. Algorithmic properties (✓-has the given information, ×-hasn't the given information)

	SpAM	SAQR	SpMAM	TERM	T-SpAM (ours)
Variable Selection	✓	✓	✓	×	✓
Robust Estimation	×	✓	✓	✓	✓
Imbalanced Classification	×	×	×	✓	✓
Multi-objective Learning	×	×	×	✓	✓
Generalization Analysis	✓	✓	×	×	✓
Variable Selection Analysis	✓	×	✓	×	✓
Function Approximation Analysis	×	✓	✓	×	✓
Optimization Acceleration	×	×	×	×	✓

It has been verified that TERM can be tuned to magnify (by taking $t \in (0, +\infty)$) or suppress (by taking $t \in (-\infty, 0)$) the influence of outliers (Li et al., 2021a). Then, the TERM-based regularized additive models can be formulated as

$$f_{\eta,t} = \underset{\substack{f_j \in \mathcal{H}_{K_j} \\ f=\sum_{j=1}^p f_j}}{\arg\min} \{\mathcal{R}_{\mathbf{z}}(t, f) + \eta \sum_{j=1}^p \tau_j \|f_j\|_{K_j}^2\}, \qquad (2)$$

where $\eta > 0$ is a regularization parameter and $\{\tau_j\}_{j=1}^p$ are positive weights. The properties of RKHS assure the minimizer of (2) can be represented as

$$f_{\eta,t} = \sum_{j=1}^p \sum_{i=1}^n \alpha_{ji}^\eta K_j(x_{ij}, \cdot), \alpha_{ji}^\eta \in \mathbb{R}.$$

In view of the above representation, we consider sparse learning in the data dependent hypothesis space

$$\mathcal{H}_{\mathbf{z}} = \{\sum_{j=1}^p \sum_{i=1}^n \alpha_{ji} K_j(x_{ij}, \cdot) : \alpha_{ji} \in \mathbb{R}\}$$

with a coefficient-induced penalty

$$\Omega_{\mathbf{z}}(f) = \inf\{\sum_{j=1}^p \tau_j \|\boldsymbol{\alpha}_j\|_2 : f = \sum_{j=1}^p \sum_{i=1}^n \alpha_{ji} K_j(x_{ij}, \cdot)\},$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, ..., \boldsymbol{\alpha}_p^T)^T \in \mathbb{R}^{np}$ and $\boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jn})^T \in \mathbb{R}^n$. The T-SpAM can be formulated as

$$f_{\mathbf{z},t} = \arg\min_{f \in \mathcal{H}_{\mathbf{z}}} \{\mathcal{R}_{\mathbf{z}}(t, f) + \lambda \Omega_{\mathbf{z}}(f)\}.$$

Denote $K_{ji} = (K_j(x_{ij}, x_{1j}), ..., K_j(x_{ij}, x_{nj}))^T \in \mathbb{R}^n$ and $\mathbf{K}_i = (K_{1i}^T, ..., K_{pi}^T)^T \in \mathbb{R}^{np}$. The estimator of T-SpAM can be rewritten as

$$f_{\mathbf{z},t} = \sum_{j=1}^p f_{\mathbf{z},t,j} = \sum_{j=1}^p \sum_{i=1}^n \alpha_{ji}^{\mathbf{z},t} K_j(x_{ij}, \cdot) \qquad (3)$$

with

$$\boldsymbol{\alpha}^{\mathbf{z},t} = \arg\min_{\alpha \in \mathbb{R}^{np}} \{\frac{1}{t} \log(\frac{1}{n} \sum_{i=1}^n e^{t\ell(\mathbf{K}_i^T \boldsymbol{\alpha}, y_i)}) + \lambda \sum_{j=1}^p \tau_j \|\boldsymbol{\alpha}_j\|_2\}. (4)$$

Remark 2.2. To circumvent the large-scale kernel matrix calculation in (4), we introduce the random Fourier features technique (Rahimi & Recht, 2007). Denote $\psi_j : \mathbb{R} \to \mathbb{R}^d$ as a random Fourier feature map associated with K_j. We know that, for any $j = 1, ..., p$,

$$K_j(x_{ij}, x_{tj}) \approx \psi_j(x_{ij})^T \psi_j(x_{tj}), \forall x_{ij}, x_{tj} \in \mathcal{X}_j.$$

Then, (3) can be approximated by

$$f_{\mathbf{z},\psi}(x_t) = \sum_{j=1}^p \hat{w}_j^T \psi_j(x_{tj}) \qquad (5)$$

with

$$\{\hat{w}_j\}_{j=1}^p = \underset{w_j \in \mathbb{R}^d, j=1,...,p}{\arg\min} \{\frac{1}{t} \log(\frac{1}{n} \sum_{i=1}^n e^{t\ell(\sum_{j=1}^p w_j^T \psi_j(x_{ij}), y_i)}) + \lambda \sum_{j=1}^p \tau_j \|w_j\|_2\}. \qquad (6)$$

Lemmas 3-4 in (Li et al., 2021a) show that the objective (13) is non-convex as $t \in (-\infty, 0)$ and smooth when $t \in (-\infty, 0) \cup (0, +\infty)$. As a result, we employ the prox-SVRG (Reddi et al., 2016) to compute (13) and provide the detailed steps in **Appendix F**.

3. Theoretical Analysis

This section provides some necessary assumptions firstly, and then gives the main theoretical results. All proofs are provided in **Appendixes B-E**.

3.1. Generalization Bound of T-SpAM

To conduct the theoretical analysis, we firstly introduce a projection operator.

Definition 3.1. The projection operator is defined by a hard threshold function $\mathcal{P}(f)(x) := \max\{-M, \min\{f(x), M\}\}$, $\forall f \in \mathcal{H}_K$ for regression estimation, and by the sigmoid function for binary classification task, i.e., $\mathcal{P}(f)(x) := \frac{1}{1+e^{-f(x)}}$, $\forall f \in \mathcal{H}_K$.

The projection operator can ensure better generalization, which has been used extensively in learning theory literatures, e.g., (Steinwart et al., 2009; Wu et al., 2007; Liu et al., 2020).

Assumption 3.2. The output Y is bounded by the interval $[-M, M]$ with a positive constant $M < +\infty$. The kernel is bounded, i.e., $\tilde{\kappa} = \sup_{j,u\in\mathcal{X}} \sqrt{K_j(u,u)} < \infty$.

The above assumptions have been used for theoretical analysis of additive models (Lv et al., 2018; Chen et al., 2020).

The generalization error bound measures the difference between $\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t}))$ and $\mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))$, which assesses the out-of-sample prediction ability of the estimator $\mathcal{P}(f_{\mathbf{z},t})$.

Theorem 3.3. *Let Assumption 3.2 be true and $K_j \in C^{\nu}$ for any $j = 1, ..., p$. For any loss function with $\|\ell(\mathcal{P}f(X), Y)\|_\infty \leq +\infty$ and its derivative $\|\ell'(\mathcal{P}f(X), Y)\|_\infty \leq +\infty$. By taking $\lambda = n^{-\zeta}$, for any fixed $t \in (-\infty, 0) \cup (0, +\infty)$ and $0 < \delta < 1$, there holds*

$$\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \lesssim n^{\Phi(s,\zeta)} \log(1/\delta)$$

with confidence at least $1 - \delta$, where $\Phi(s,\zeta) = \max\{\frac{s-2+2s\zeta}{4}, -\frac{1}{2}\}$ and

$$s = \begin{cases} \frac{2}{1+2\nu}, & \nu \in (0, 1]; \\ \frac{2}{1+\nu}, & \nu \in (1, 3/2]; \\ \frac{1}{\nu}, & \nu \in (3/2, \infty). \end{cases} \tag{7}$$

Remark 3.4. Figure 1 summaries the convergence rates of generalization error bound deduced in Theorem 3.3 by taking different s and ζ. Note that the generalization error of T-SpAM will not converge when ζ and v are both located in the white area. Moreover, the larger the value of v, the larger the selection range of regularization parameter λ results in faster convergence rate. For any $t \in (-\infty, 0) \cup (0, +\infty)$, $\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \to 0$ when $\zeta \in (-\infty, \frac{2-s}{2s})$ and $n \to +\infty$. By taking $\zeta \leq -\frac{1}{2}$, we can get the learning rate with polynomial decay $O(n^{-\frac{1}{2}})$.

3.2. Function Approximation Bounds under Specific Data Distribution Assumptions

The definition of $f_{\mathbf{z},t}$ in (3) indicates that the approximation ability of estimator $\mathcal{P}(f_{\mathbf{z},t})$ is closely related to the scale parameter t. Thus, in the presence of different data distributions, it is important to explore how to select t to achieve good approximation performance. In this section, we pursue the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $\mathcal{P}(f_{\mathbf{z},t})$ and f^* under three specific types of noise distributions ((Noise assumptions A-C)), where $\rho_\mathcal{X}$ is the marginal distribution of ρ on $\mathcal{X}$ and $L^2_{\rho_\mathcal{X}}$ is the square-integrable function space.

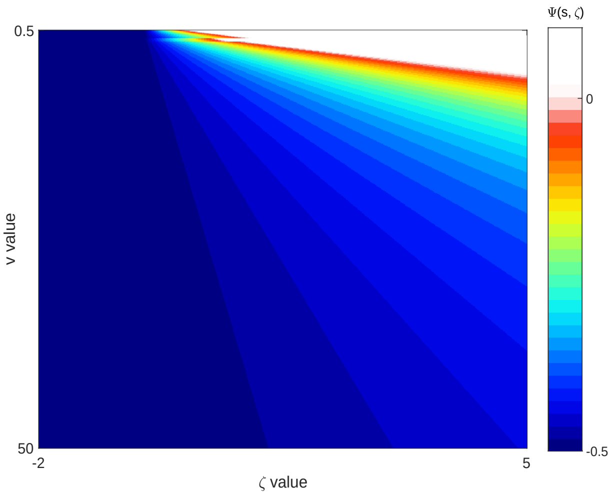


Figure 1. The convergence rates of generalization error by taking different ζ and v.

This section considers a common data-generating model as $Y = f^*(X) + \epsilon$, where $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, ϵ is a random noise and f^* is an unknown ground truth function. Moreover, the loss function is selected as the squared loss $\ell(f(x), y) = (f(x) - y)^2$. Here, we consider an commonly used case in learning theory (Guo & Zhou, 2013; Zou et al., 2009; Shi et al., 2011), where f^* is bounded and belongs to the RKHS, i.e., $f^* = \sum_{j=1}^p f_j^*$, $\|f^*\|_\infty < +\infty$ and $f_j^* \in \mathcal{H}_{K_j}, \forall j \in \{1, ..., p\}$.

Assumption 3.5. (Noise assumption A) The noise variable satisfies $\mathbb{E}(\epsilon | X = x) = 0, \forall x \in \mathcal{X}$.

Theorem 3.6. *Let Assumptions 3.2, 3.5 be true and $f_{\eta,t}$ in (2) be bounded. By taking $\lambda = n^{-\zeta}$ with $\zeta \in (-\frac{1}{2}, \frac{2-s}{2s})$ and $\eta = n^{-\frac{1+2\zeta}{4}}$, for any $0 < \delta < 1$, there holds*

$$\begin{aligned} &\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \\ &\quad \lesssim \left[|t|^{-1} n^{\max\{\frac{s-2+2s\zeta}{4}, \frac{-1}{2}\}} + n^{\frac{-1-2\zeta}{4}} + |t| \right] \log(1/\delta) \end{aligned}$$

with confidence at least $1 - \delta$.

Corollary 3.7. *Let the conditions of Theorem 3.6 be true. By taking $t = \pm n^{-\beta}$ with $\beta \in (0, \min\{\frac{1}{2}, \frac{2-s-2s\zeta}{4}\})$, the estimator $\mathcal{P}(f_{\mathbf{z},t})$ can approach f^* with an sample-dependent error of less than*

$$O(n^{\max\{\frac{4\beta+s-2+2s\zeta}{4}, \frac{2\beta-1}{2}, \frac{-1-2\zeta}{4}, -\beta\}}).$$

Assumption 3.5 admits a class of zero-mean noise distribution, e.g., student's t distribution, skewed Gaussian distribution and Gaussian distribution. Theorem 3.6 and Corollary 3.7 indicate that a proper t value ensures that our estimator approximate f^*, which verify the robustness of our T-SpAM. It should be noticed that the previous theoretical studies of additive models are usually limited to ERM (or SRM) under Gaussian noise, see e.g., (Huang et al., 2010; Meier et al., 2009; Ravikumar et al., 2009; Yuan & Zhou, 2016).

We turn now to analyze the approximation performance on a class of symmetric zero-mean noise distribution.

Assumption 3.8. (Noise assumption B) The noise ϵ satisfies $\mathbb{E}(\epsilon|X=x)=0$, $p_{\epsilon|X}(m|X=x)=p_{\epsilon|X}(-m|X=x)$, $\forall m\in\mathbb{R},\forall x\in\mathcal{X}$ and $\|p_{\epsilon|X}\|_\infty<+\infty$, where $p_{\epsilon|X}$ denotes the conditional density of noise ϵ.

Theorem 3.9. *Let Assumptions 3.2, 3.8 be true and $f_{\eta,t}$ be bounded. For fixed $t\in(-\frac{1}{2(M+\|f^*\|_\infty)^2},0)$ and any $0<\delta<1$, with confidence at least $1-\delta$, there holds*

$$\|\mathcal{P}(f_{\mathbf{z},t})-f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}\lesssim n^{\max\{\frac{2s\zeta+s-2}{4},-\frac{1}{2},\frac{-1-2\zeta}{4}\}}\log(1/\delta).$$

Corollary 3.10. *Let the conditions of Theorem 3.9 be true and $K_j\in C^\infty$ for any $j=1,...,p$. Setting $\zeta>\frac{1}{2}$, we get*

$$\|\mathcal{P}(f_{\mathbf{z},t})-f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}=O(n^{-\frac{1}{2}}).$$

To introduce the third type of noise assumption, we first define strongly m-concave function (Doss & Wellner, 2013; Feng et al., 2020).

Definition 3.11. A function g is strongly m-concave if it exhibits one of the following forms:

(i) $g=\max\{\phi^{\frac{1}{m}},0\}$ for some strongly concave function ϕ if $m>0$;

(ii) $g=\exp(\phi)$ for some strongly concave function ϕ if $m=0$;

(iii) $g=\max\{\phi^{\frac{1}{m}},0\}$ for some strongly convex function ϕ if $m<0$.

Assumption 3.12. (Noise assumption C) The conditional density $p_{\epsilon|X}$ satisfies following conditions:

(i) For any $x\in\mathcal{X}$, the conditional density $p_{\epsilon|X}$ satisfies the conditions

(i-a) $\sup_{t\in\mathbb{R}}p_{\epsilon|X}(t|X=x)<+\infty$;

(i-b) $\inf_{t\in(-\infty,+\infty)}p_{\epsilon|X}(t|X=x)>0$;

(i-c) $p_{\epsilon|X}(t|X=x)\le p_{\epsilon|X}(0|X=x),\forall t\in\mathbb{R}$.

(ii) $\sup_{x\in\mathcal{X}}p''_{\epsilon|X}(\cdot|X=x)<+\infty$;

(iii) The conditional density $p_{\epsilon|X=x},\forall x\in\mathcal{X}$ satisfies strongly m-concave condition.

Assumption 3.12 is widely used in modal regression (Feng et al., 2020; Wang et al., 2017; Chen et al., 2020). Condition (i) holds for common continuous densities with a unique global mode. Condition (ii) is key for subsequent theoretical analysis. Condition (iii) is typical from a statistical viewpoint (Doss & Wellner, 2013; Feng et al., 2020) as it holds for common symmetric and skewed noise distributions, e.g., chi-square distribution, student's t distribution, skewed normal distribution and normal distribution.

Theorem 3.13. *Let Assumptions 3.2, 3.12 be true and $f_{\eta,t}$ be bounded. For any $t\in(-\infty,0)\cup(0,+\infty)$, we take $t=-\log n^{-\beta},\beta\in(0,\min\{\frac{1}{8M^2},\frac{2-s-2s\zeta}{16M^2}\})$, $\lambda=n^{-\zeta},\zeta\in(\frac{1}{2},\frac{2-s}{2s})$ and $\eta=n^{-\frac{1+2\zeta}{4}}$. For any $0<\delta<1$, with confidence at least $1-\delta$, there holds*

$$\|\mathcal{P}(f_{\mathbf{z},t})-f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}\lesssim\log^{-1}n\log(1/\delta).$$

Corollary 3.14. *Under the zero-mode noise assumption, the ground truth function f^* is a conditional mode function (Sager & Thisted, 1982; Yao & Li, 2013; Chen et al., 2016; Feng et al., 2020), i.e., $f^*(x)=\arg\max_t p_{Y|X}(t|X=x)$, where $p_{Y|X}$ is the conditional density of output Y. Theorem 3.13 shows the estimator $\mathcal{P}(f_{\mathbf{z},t})$ can approach f^* with an sample-dependent error of less than $O(\log^{-1}n)$, which verifies the robustness of our T-SpAM to the outliers in that T-SpAM would focus on the high conditional density region. Recall the approximation rate $O(n^{-1/2})$ derived in Corollary 3.10, the slower approximation rate $O(\log^{-1}n)$ we obtained in Theorem 3.13 indicates the sacrifice for dealing with more complex zero-mode noise distributions, e.g., heavy-tailed noise and skewed noise.*

With proper parameter selection strategies, Theorems 3.6-3.13 together ensure the approximation performance of T-SpAM under three specific different noise assumptions. (Li et al., 2021a) stated that TERM become robust when taking $t<0$. Indeed, we here strengthen the theoretical guarantees of TERM from the perspective of learning theory, in the sense that giving more specific parameter selection suggestions in the face of specific data analysis.

In addition, we also investigate the variable selection consistency of our method in Theorem E.2 (provided in **Appendix E**). Theorem E.2 illustrates that T-SpAM can identify the desired variables by taking properly λ and weight $\tau_j,j=1,...,p$. Indeed, the current analysis extends Theorem 4 in (Wang et al., 2017) from a linear regularized modal regression to the nonlinear T-SpAM. Moreover, it is interesting to further explore variable selection analysis by replacing the parameter conditions here with the incoherence assumptions (e.g. Assumption 4 in (Lv et al., 2018)).

4. Extensions of T-SpAM: Classification and Multi-objective Learning

Inspired by the idea in (Ravikumar et al., 2009), the T-SpAM can be extended to tilted additive logistic regression for classification problem. Assume that input $X\in\mathcal{X}$ and output $Y\in\{0,1\}$. Given observations $\mathbf{z}=\{(x_i,y_i)\}_{i=1}^n$, the T-SpAM induced by logistic loss can be formulated as

$$P(Y=1|X=x)=\frac{\exp(\sum_{j=1}^p\sum_{i=1}^n\alpha_{ji}^{\mathbf{z},t}K_j(x_{ij},x))}{1+\exp(\sum_{j=1}^p\sum_{i=1}^n\alpha_{ji}^{\mathbf{z},t}K_j(x_{ij},x))}$$

where

$$\boldsymbol{\alpha}^{\mathbf{z},t} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{np}} \{\frac{1}{t}\log(\frac{1}{n}\sum_{i=1}^{n} e^{t(y_i\mathbf{K}_i^T\boldsymbol{\alpha} - \log(1+e^{\mathbf{K}_i^T\boldsymbol{\alpha}}))}) + \lambda\sum_{j=1}^{p}\tau_j\|\boldsymbol{\alpha}_j\|_2\}. \tag{8}$$

Remark 4.1. In view of Lemma 1 in (Li et al., 2021a), the tilted additive logistic regression can be tuned to achieve robust classification (by taking $t \in (-\infty, 0)$) and imbalanced classification (by taking $t \in (0, +\infty)$).

In real-world application, a multi-objective extension of T-SpAM can be formulated to tackle complicated learning problems, e.g., mitigating label noise and addressing imbalance data simultaneously. By classifying the samples into different groups $g \in [G]$ (e.g., the groups can be associated with the classes in classification task), a multi-objective TERM can be formulated as

$$\mathcal{J}_{\mathbf{z}}(t,\gamma,f) := \frac{1}{t}\log(\frac{1}{|G|}\sum_{g\in G} e^{t\mathcal{R}_{\mathbf{z},g}(\gamma,f)})$$

where

$$\mathcal{R}_{\mathbf{z},g}(\gamma,f) = \frac{1}{\gamma}\log(\frac{1}{|g|}\sum_{z_i\in g} e^{\gamma\ell(f(x_i),y_i)}).$$

Then, the T-SpAM for two-objective learning problem can be represented as

$$f_{\mathbf{z},t,\gamma} = \sum_{j=1}^{p} f_{\mathbf{z},t,\gamma,j} = \sum_{j=1}^{p}\sum_{i=1}^{n} K_{ji}(x_i,\cdot)\alpha_{ji}^{\mathbf{z},t,\gamma}$$

with

$$\boldsymbol{\alpha}^{\mathbf{z},t,\gamma} = \arg\min_{\alpha\in\mathbb{R}^{np}} \{\frac{1}{t}\log(\frac{1}{|G|}\sum_{g\in G} e^{\frac{t}{\gamma}\log(\frac{1}{|g|}\sum_{z_i\in g} e^{\gamma\ell(\mathbf{K}_i^T\boldsymbol{\alpha},y_i)})}) + \lambda\sum_{j=1}^{p}\tau_j\|\boldsymbol{\alpha}_j\|_2\}.$$

Here, we can select $\ell(f(x),y)$ as the squared loss for regression and as the logistic loss for classification.

Remark 4.2. Lemma 7 in (Li et al., 2021a) demonstrates that the multi-objective T-SpAM is equivalent to T-SpAM when $\gamma = t$. Indeed, the scale parameter t controls the tiled level between groups $g \in [G]$, and the weight τ impacts the tilted level between samples in group $g \in [G]$. Detailed optimization procedures of T-SpAM with logistic loss and multi-objective loss are provided in **Appendix F**.

5. Experiments

This section conducts the empirical evaluations on synthetic data and real-world data to verify the effectiveness of our T-SpAM in terms of robust regression prediction, robust classification, accurate imbalanced classification and the ability to handle multi-objective learning problem.

In all synthetic experiments, we independently generate training dataset, validation dataset and test dataset, where the hyper-parameters t and λ are tuned in grids $\{\pm 0.1, \pm 0.5, \pm 1, \pm 2\}$ and $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ on validation dataset. For regression task, we use the *average squared error (ASE)* to describe the divergence between the $f^*(x)$ and the prediction $\hat{f}(x)$, i.e., ASE $= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x) - f^*(x))^2$. For simplicity, T-SpAM(K) and T-SpAM(F) refer to the T-SpAM by kernel matrix calculation (see (4)) and by random Fourier feature acceleration (see (13)), respectively. Inspired by (Feng et al., 2016), variable selection results are measured in terms of the *average selection percentage (ASP)*, which refers to the average probability of variables selected correctly.

5.1. Synthetic data experiments

For regression task, the datasets are generated by the data generating model $Y = f^*(X) + \epsilon$. Following the experimental design in (Chen et al., 2020), we consider $f^(X) = \sum_{j=1}^{8} f_j(X_j)$ with

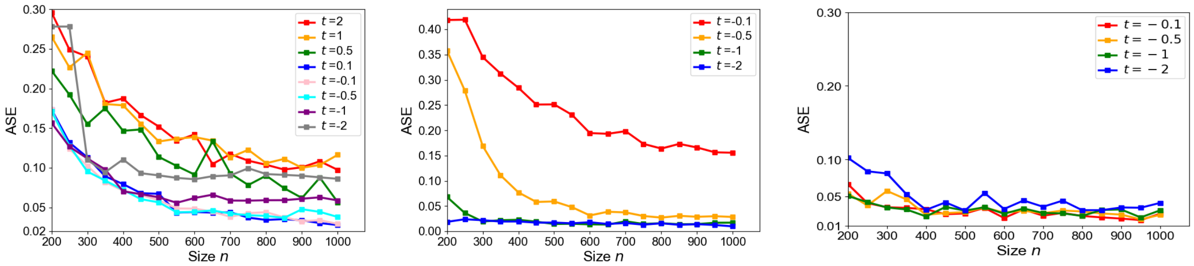
$$f_1(u) = -2\sin(2u), f_2(u) = 8u^2, f_3(u) = \frac{7\sin u}{2-\sin u},$$

$$f_4(u) = 6e^{-u}, f_5(u) = u^3 + \frac{3}{2}(u-1)^2, f_6(u) = 5u,$$

$$f_7(u) = 10\sin(e^{-u/2}), f_8(u) = -10\widetilde{\phi}(u, \frac{1}{2}, \frac{4}{5}^2),$$

where $\widetilde{\phi}$ is the normal cumulative distribution with mean $\frac{1}{2}$ and standard deviation $\frac{4}{5}$. Let the sample size $n = 200$ and dimension $p = 100$ (including 92 irrelevant variables). Each input $X_j, j = 1, ..., 100$, is generated from $U(-1, 1)$. For training data and validation data, three types of noises are considered here and named ϵ^A, ϵ^B and ϵ^C, respectively. Here, the ϵ^A is drawn from a skewed zero-mean distributions $\{0.8\mathcal{N}(-2,1) + 0.2\mathcal{N}(8,1)\}$, the ϵ^B follows the skewed zero-mode distribution $\{0.8\mathcal{N}(0,1) + 0.2\mathcal{N}(20,1)\}$ and the ϵ^C is generated from heavy-tailed Student's t distribution. To obtain ASE, the test data are generated by $Y = f^*(X)$.

We compare our proposed approch (T-SpAM(F) and T-SpAM(K)) with several related methods, e.g., Lasso (Tibshirani, 1994), SpAM (Ravikumar et al., 2009) and RMR (Wang et al., 2017). We repeat each experiment for 50 times and report the average results under three types of noise distributions. Table 2 shows that, with a small $|t|$, T-SpAM(F),



(a) The ASE under noise ϵ^A

(b) The ASE under noise ϵ^B

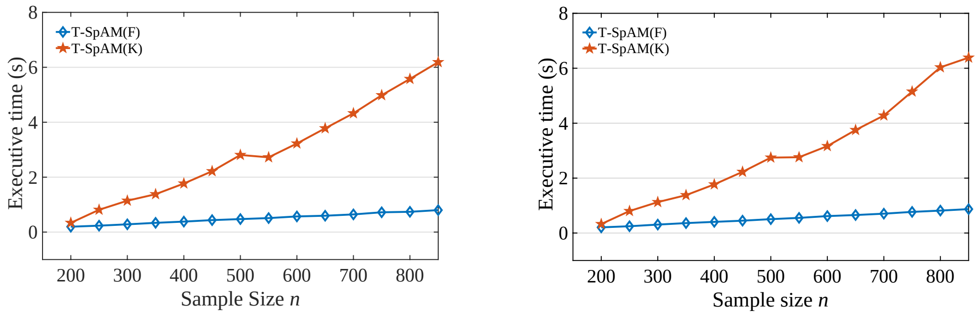
(c) The ASE under noise ϵ^C

Figure 2. The ASE of T-SpAM(F) under different t and sample size n

Table 2. The averaged performance on synthetic data under noises ϵ^A (Top), ϵ^B (Middle) and ϵ^C (Bottom).

Method		ASP	ASE	Method		ASP	ASE
$t=-0.1$	T-SpAM (F)	0.923	0.173 ± 0.042	$t=0.1$	T-SpAM (F)	0.919	0.173 ± 0.037
	T-SpAM (K)	0.925	0.184 ± 0.048		T-SpAM (K)	0.918	0.237 ± 0.050
$t=-0.5$	T-SpAM (F)	0.908	0.171 ± 0.041	$t=0.5$	T-SpAM (F)	0.906	0.222 ± 0.056
	T-SpAM (K)	0.875	0.244 ± 0.055		T-SpAM (K)	0.905	0.268 ± 0.056
$t=-1.0$	T-SpAM (F)	**0.925**	$\mathbf{0.157} \pm \mathbf{0.046}$	$t=1.0$	T-SpAM (F)	0.913	0.265 ± 0.626
	T-SpAM (K)	0.865	0.289 ± 0.065		T-SpAM (K)	0.905	0.269 ± 0.055
$t=-2.0$	T-SpAM (F)	0.810	0.278 ± 0.097	$t=2.0$	T-SpAM (F)	0.881	0.296 ± 0.068
	T-SpAM (K)	0.700	0.416 ± 0.087		T-SpAM (K)	0.905	0.277 ± 0.054
–	Lasso	0.830	1.037 ± 0.119	–	SpAM	0.918	0.204 ± 0.049
	RMR	0.380	0.597 ± 0.141				
Method		ASP	ASE	Method		ASP	ASE
$t=-0.1$	T-SpAM (F)	0.505	0.418 ± 0.077	$t=-0.5$	T-SpAM (F)	0.573	0.358 ± 0.099
	T-SpAM (K)	0.675	0.272 ± 0.067		T-SpAM (K)	0.715	0.196 ± 0.043
$t=-1.0$	T-SpAM (F)	0.938	0.068 ± 0.077	$t=-2.0$	T-SpAM (F)	**1.000**	$\mathbf{0.019} \pm \mathbf{0.005}$
	T-SpAM (K)	0.878	0.144 ± 0.042		T-SpAM (K)	0.955	0.110 ± 0.025
–	Lasso	0.655	0.713 ± 0.057	–	SpAM	0.588	0.350 ± 0.078
	RMR	0.610	0.425 ± 0.112				
Method		ASP	ASE	Method		ASP	ASE
$t=-0.1$	T-SpAM (F)	1.000	0.041 ± 0.013	$t=-0.5$	T-SpAM (F)	**1.000**	$\mathbf{0.037} \pm \mathbf{0.012}$
	T-SpAM (K)	0.993	0.152 ± 0.041		T-SpAM (K)	0.993	0.158 ± 0.042
$t=-1.0$	T-SpAM (F)	1.000	0.042 ± 0.013	$t=-2.0$	T-SpAM (F)	1.000	0.084 ± 0.037
	T-SpAM (K)	0.990	0.193 ± 0.054		T-SpAM (K)	0.885	0.316 ± 0.093
–	Lasso	0.865	1.546 ± 0.269	–	SpAM	0.988	0.094 ± 0.060
	RMR	0.843	0.247 ± 0.062				

T-SpAM(K) and SpAM have a satisfactory performance under zero-mean noise ϵ^A, while Lasso and RMR produce a large *ASE* value as they can only fit functions linearly. Moreover, when t tends to be small, T-SpAM(F) and T-SpAM(K) outperforms other competitors under zero-mode noise ϵ^B in the sense that it can select all variables correctly and obtain the smallest ASE. This result implies the robustness of T-SpAM(F) to skewed zero-mode noise and supports the findings in Theorem 3.13. Furthermore, we investigate the impact of t and n on the performance of T-SpAM(F) in



(a) Executive time under noise ϵ^A

(b) Executive time under noise ϵ^B

Figure 3. The executive time for T-SpAM(K) and T-SpAM(F) when taking $t = -0.5$.

Table 3. The accuracy on contaminated data (left) and imbalanced data (right) respectively.

Robust Classification				Imbalanced Classification			
	Method	ASP	Accuracy		Method	ASP	Accuracy
$r_1 = 0.1$	T-SpAM(F)	1.000	**0.887 ± 0.028**	$r_2 = 0.05$	T-SpAM(F)	1.000	**0.832 ± 0.042**
	ℓ_1-SVM	0.430	0.583 ± 0.043		ℓ_1-SVM	1.000	0.500 ± 0.000
	SAM	0.980	0.761 ± 0.048		SAM	1.000	0.800 ± 0.046
	SpAM	0.940	0.791 ± 0.058		SpAM	1.000	0.715 ± 0.096
$r_1 = 0.3$	T-SpAM(F)	0.530	**0.702 ± 0.066**	$r_2 = 0.10$	T-SpAM(F)	1.000	**0.880 ± 0.031**
	ℓ_1-SVM	0.530	0.572 ± 0.054		ℓ_1-SVM	1.000	0.500 ± 0.000
	SAM	0.380	0.598 ± 0.042		SAM	1.000	0.860 ± 0.037
	SpAM	0.410	0.632 ± 0.047		SpAM	1.000	0.819 ± 0.058
$r_1 = 0.5$	T-SpAM(F)	0.270	**0.554 ± 0.066**	$r_2 = 0.15$	T-SpAM(F)	1.000	**0.910 ± 0.025**
	ℓ_1-SVM	0.230	0.534 ± 0.067		ℓ_1-SVM	1.000	0.500 ± 0.000
	SAM	0.030	0.496 ± 0.042		SAM	1.000	0.885 ± 0.032
	SpAM	0.010	0.499 ± 0.054		SpAM	1.000	0.863 ± 0.052

Table 4. The accuracy on contaminated and imbalanced classification data.

Method	$r_1 = 0.3, r_2 = 0.1$	$r_1 = 0.3, r_2 = 0.5$	$r_1 = 0.1, r_2 = 0.1$	$r_1 = 0.1, r_2 = 0.15$
T-SpAM(F)	**0.655 ± 0.078**	**0.681 ± 0.071**	**0.782 ± 0.042**	**0.828 ± 0.037**
SpAM	0.616 ± 0.106	0.680 ± 0.089	0.705 ± 0.080	0.791 ± 0.064
SAM	0.603 ± 0.074	0.639 ± 0.053	0.760 ± 0.074	0.800 ± 0.048
ℓ_1-SVM	0.500 ± 0.007	0.501 ± 0.015	0.500 ± 0.002	0.500 ± 0.002

Figure 2. Clearly, with sample size n increasing, the *ASE* has a downward trend for all t, which verifies the findings in Theorems 3.3, 3.6 and 3.13. Besides, we can see that the ASE becomes smaller as $|t|$ tends to 0 for noise ϵ^A, which is consistent with the result in Theorem 3.6. Moreover, the decreasing of *ASE* as t decreases verifies Theorem 3.13. Figure 3 reports the executive times of T-SpAM(F) and T-SpAM(K) under noises ϵ^A and ϵ^B. Together with the results from Table 2, we can see that T-SpAM(F) performs as well as T-SpAM(K) but runs faster than T-SpAM(K), which verifies the effectiveness of random Fourier features technique.

Robust classification: For classification, we consider the discriminant function used in (Zhao & Liu, 2012). The

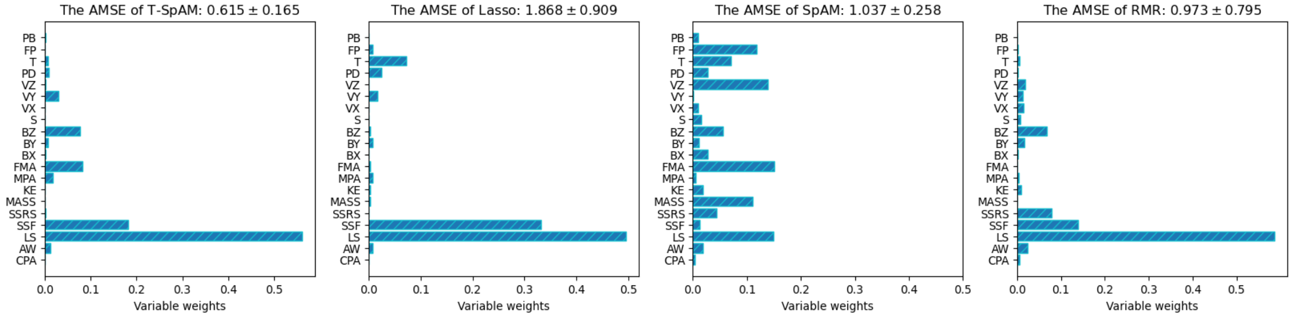


Figure 4. AMSE and the weights of variables.

discriminant function is additive and formulated as follows

$$f^*(x_i) = (x_{i1} - 0.5)^2 + (x_{i2} - 0.5)^2 - 0.08, \quad (9)$$

where $x_{ij} = (W_{ij} + U_i)/2$ and W_{ij}, U_i are independently from $U(0, 1)$ for $i = 1, \ldots, 200$, $j = 1, \ldots, 100$. The label y_i is 0 when $f(x_i) \le 0$ and 1 otherwise. We add the noise into data by randomly flipping the label with some certain ratios r_1 (see Table 3). Here, the competitors include SpAM (induced by logistic loss) (Ravikumar et al., 2009), SAM (Zhao & Liu, 2012) and linear ℓ_1-SVM (Zhu et al., 2003). The variable selection results and classification accuracy on test set are presented in Table 3, which entails the T-SpAM(F) behaves better in robust classification.

Imbalanced classification: Now we check the property of T-SpAM to handle imbalanced classification issue. The synthetic data are constructed from (9) with $N = 200, p = 10$. We compare our method with SpAM(induced by logistic loss)(Ravikumar et al., 2009), SAM(Zhao & Liu, 2012), and ℓ_1-SVM(Zhu et al., 2003) under some extreme ratios r_2 of negative class in the population. Table 3 empirically verifies that the T-SpAM(F) has an advantage in dealing with imbalanced data.

Multi-objective learning: The effectiveness of multi-objective T-SpAM is checked under imbalanced and robust classification setting. For simplicity, we only consider the combinations of $r_1 \in \{0.1, 0.3\}$ and $r_2 \in \{0.10, 0.15\}$. Based on (9), the synthetic data are generated with $n = 200, p = 10$. Table 4 implies the T-SpAM(F) is slightly superior to the other methods.

5.2. Real-world data experiment

In this section, we evaluate the performance of T-SpAM on Coronal Mass Ejections (CME) data. The CME data (https://cdaw.gsfc.nasa.gov/CME list/) consists of a output variable (arrival time of coronal mass ejections) and 20 input variables, including center projection angle (CPA), angle width (AW), linear speed (LS), SND speed final (SSF), SND speed 20RS (SSRS), MASS, kinetic energy (KE), measurement position angle (MPA), field magnitude average (FMA), BX, BY, BZ, Speed (S), VX, VY, VZ, proten density (PD), Temperature (T), flow pressure (FP), plasma beta (PB). We randomly split the CME data into three parts: 86 observations for training, 22 observations for validation, and 27 observations for test. To evaluate model performance, the *average mean squared error (AMSE)* is obtained by repeating the experiment 50 times. To the end, our T-SpAM(F) enjoys the smallest AMSE (0.615 ± 0.165 by taking $t = -0.1$) compared with SpAM (1.037 ± 0.258), RMR (0.973 ± 0.795) and Lasso (1.868 ± 0.909). Moreover, Figure 4 shows that LS, SSF, FMA and BZ are significant in arrival time prediction, which has been verified in (Liu et al., 2018).

6. Conclusion

In this paper, we propose a novel tilted sparse additive models (T-SpAM) that can be capable of a variety of learning tasks, e.g., robust estimation, robust classification, imbalanced classification, and multi-objective learning. Under some common skewed or symmetric noise assumptions, theoretical guarantees on generalization bound, function approximation, and variable selection consistency are established for T-SpAM. In practice, empirical evaluations support the advanced performance of our approach.

Acknowledgements

This work was supported by Yunnan provincial major science and technology special plan projects under Grant 202202AD080003, National Natural Science Foundation of China under Grant Nos. 12071166, 61671480, HZAU-AGIS Cooperation Fund No. SZYJY2023010, the Fundamental Research Funds for the Central Universities of China under Grant No. 2662023LXPY005, the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-008. Prof. Dacheng Tao is partially supported by Australian Research Council Project FL-170100117.

References

Agarwal, R., Frosst, N., Zhang, X., Caruana, R., and Hinton, G. E. Neural additive models: Interpretable machine learning with neural nets. *arXiv:2004.13912v1*, 2020.

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Borkar, V. S. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2):294–311, 2002.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Caponnetto, A. and Vito, E. D. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Chen, H., Wang, X., Deng, C., and Huang, H. Group sparse additive machine. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 198–208. 2017.

Chen, H., Wang, Y., Zheng, F., Deng, C., and Huang, H. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2020.3005144, 2020.

Chen, Y. C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. Nonparametric modal regression. *The Annals of Statistics*, 44(2):489–514, 2016.

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

Cucker, F. and Zhou, D. X. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Doss, C. R. and Wellner, J. A. Global rates of convergence of the mles of log-concave and s-concave densities. *The Annals of Statistics*, 44(3):954–981, 2013.

Feng, Y., Huang, X., Shi, L., Yang, Y., and Suykens, J. A. Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, 16:993–1034, 2015.

Feng, Y., Yang, Y., and Suykens, J. A. K. Robust gradient learning with applications. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):822–835, 2016.

Feng, Y., Fan, J., and Suykens, J. A statistical learning approach to modal regression. *Journal of Machine Learning Research*, 21(2):1–35, 2020.

Guo, Z. C. and Zhou, D. X. Concentration estimates for learning with unbounded sampling. *Advances in Computational Mathematics*, 38(1):207–223, 2013.

Hastie, T. J. and Tibshirani, R. J. *Generalized additive models*. London: Chapman and Hall, 1990.

Howard, R. A. and Matheson, J. E. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

Huang, J., Horowitz, J. L., and Wei, F. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313, 2010.

J. Reddi, S., Sra, S., Poczos, B., and Smola, A. Fast stochastic methods for nonsmooth nonconvex optimization. 05 2016.

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

Kowalski, M. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324, 2009. ISSN 1063-5203.

Lee, J., Park, S., and Shin, J. Learning bounds for risk-sensitive learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. Tilted empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2021a.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021b.

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

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

Lv, S., Lin, H., Lian, H., and Huang, J. Oracle inequalities for sparse additive quantile regression in reproducing kernel hilbert space. *The Annals of Statistics*, 46(2):781–813, 2018.

Meier, L., Geer, S. V. D., and Buhlmann, P. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Osogami, T. Robustness and risk-sensitivity in markov decision processes. pp. 233–241, 2012.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1177–1184, 2007.

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

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. SpAM: sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71:1009–1030, 2009.

Reddi, S. J., Sra, S., Poczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1145–1153. 2016.

Sager, T. W. and Thisted, R. A. Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics*, 10(3):690–707, 1982.

Shi, L. Learning theory estimates for coefficient-based regularized regression. *Applied and Computational Harmonic Analysis*, 34(2):252–265, 2013.

Shi, L., Feng, Y. L., and Zhou, D. X. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31(2):286–302, 2011.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Science and Business Media, 2008.

Steinwart, I., Hush, D., and Scovel, C. Optimal rates for regularized least squares regression. In *Annual Conference on Learning Theory (COLT)*, 2009.

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

Tan, Z. and Zhang, C.-H. Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics*, 47(5):2567 – 2600, 2019.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 73(3):267–288, 1994.

Wang, C. and Zhou, D.-X. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.

Wang, X., Chen, H., Cai, W., Shen, D., and Huang, H. Regularized modal regression with applications in cognitive impairment prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1448–1458, 2017.

Wang, Y., Zhong, X., He, F., Chen, H., and Tao, D. Huber additive models for non-stationary time series analysis. In *International Conference on Learning Representations (ICLR)*, 2022.

Wu, Q., Ying, Y., and Zhou, D. X. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Wu, Q., Ying, Y., and Zhou, D. X. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

Yang, L., Lv, S., and Wang, J. Model-free variable selection in reproducing kernel hilbert space. *Journal of Machine Learning Research*, 17(1):2885–2908, 2016.

Yang, Z., Zhang, A., and Sudjianto, A. Gami-net: An explainable neural network based on generalized additive models with structured interactions. *arXiv:2003.07132*, 2020.

Yao, W. and Li, L. A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3): 656–671, 2013.

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

Yuan, M. and Zhou, D. X. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

Zhao, T. and Liu, H. Sparse additive machine. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 1435–1443. PMLR, 21–23 Apr 2012.

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

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. 1-norm support vector machines. pp. 49–56. MIT Press, 2003.

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

Appendix to "Tilted Sparse Additive Models"

A. Notations

Some used notations are summarized in Table 5.

Table 5. Notations

Notations	Descriptions
$\mathcal{X}, \mathcal{Y}$	the input space and the output space, respectively
X, Y	random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively
x, y	realizations of X and Y, respectively
ϵ	the noise variable specified by the residual $Y - f^*(X)$
p	the dimension of input variable X
n	the sample size
$\mathbf{z}$	a set of n-size realizations of X, Y, i.e., $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$
ρ	the joint probability distribution of $\mathcal{X} \times \mathcal{Y}$
$\rho_{\mathcal{X}}$	the marginal distribution of X
$p_{Y\|X}$	the conditional density of Y conditioned on X
$p_{\epsilon\|X}$	the conditional density of ϵ conditioned on X
$\mathcal{L}^2_{\rho^2_{\mathcal{X}}}$	the function space of square-integrable functions with respect to $\rho_{\mathcal{X}}$
$\mathcal{H}_K$	the data independent reproducing kernel Hilbert space
$\mathcal{H}_{\mathbf{z}}$	the data dependent hypothesis space
f^*	the unknown ground truth function
$\mathcal{R}(t, f)$	the population version of tilted empirical risk minimization framework
$\mathcal{R}_{\mathbf{z}}(t, f)$	the tilted empirical risk minimization framework
$\mathcal{G}(t, f)$	the population version of stepping-stone objective function
$\mathcal{G}_{\mathbf{z}}(t, f)$	the empirical version of stepping-stone objective function
$\mathcal{E}_\rho(f)$	the objective function for conditional mean function
$\mathcal{E}_M(f)$	the objective function for conditional mode function

B. Proof of Theorem 1

Proof sketch: Usually, the generalization analysis of traditional ERM-based algorithm can be conducted by considering the expected risk and empirical risk as the expectation of random error variable and the average of random independent error observations (Cucker & Zhou, 2007; Steinwart & Christmann, 2008). However, $\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t}))$ and $\mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))$ associated with TERM do not enjoy this property due to the additional logarithmic function. Hence, the concentration inequalities (e.g., used in (Wu et al., 2007; Steinwart & Christmann, 2008)) can not be used to bound the generalization of T-SpAM directly. In this part, we overcome this difficulty by introducing the stepping-stone objectives

$$\mathcal{G}(t, f) := \frac{1}{t} \int_{\mathcal{Z}} e^{t(f(x)-y)^2} d\rho(x, y)$$

and

$$\mathcal{G}_{\mathbf{z}}(t, f) := \frac{1}{nt} \sum_{i=1}^n e^{t(f(x_i)-y_i)^2}.$$

To the end, the main building blocks in generalization consistency analysis are presented in Figure 5.

We assume $\ell(\mathcal{P}f(X), Y) \leq M_\ell$ and $\ell'(\mathcal{P}f(X), Y) \leq M_{\ell'}$ for any $f(X)$ and Y. To bound $\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z},t}(t, \mathcal{P}(f_{\mathbf{z},t}))$, we first give the upper bound of $\|f_{\mathbf{z},t}\|_K$.

Lemma B.1. *Under Assumption 1, there hold*

$$\|f_{\mathbf{z},t}\|_K \leq M_\ell^2 \tilde{\kappa} n^{\frac{1}{2}} \lambda^{-1} \tilde{\tau}^{-1}, \quad \forall t \in (-\infty, 0) \cup (0, +\infty)$$

where $\tilde{\tau} = \min_{j=1,\ldots,p} \tau_j$ *is a positive constant.*

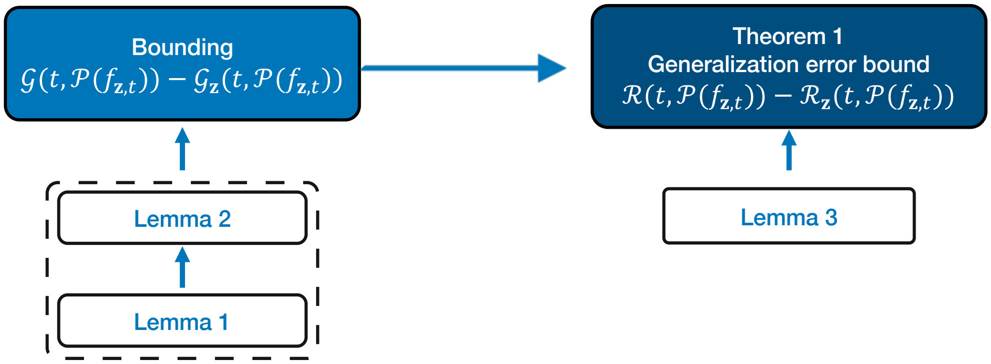

Bounding
$\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))$
Theorem 1
Generalization error bound
$\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{R}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))$
Lemma 2
Lemma 1
Lemma 3

Figure 5. An illustration of the main building blocks in generalization consistency analysis. We first establish the upper bound of $\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))$ with the help of Lemmas B.1-B.3. Lemma B.4 investigates the relation between $\mathcal{R}(t,f_{\mathbf{z},t})-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t})$ and $\mathcal{G}(t,f_{\mathbf{z},t})-\mathcal{G}_{\mathbf{z}}(t,f_{\mathbf{z},t})$. Finally, Theorem 1 can be obtained by combining Lemma B.4 and the bound of the generalization error $\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))$.

Proof. From the definition of $f_{\mathbf{z},t}$, we have

$$\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t})+\lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},t})\le\mathcal{R}_{\mathbf{z}}(t,0)+\lambda\Omega_{\mathbf{z}}(0).$$

Under Assumption 1, we come to the following two conclusions, i.e.,

$$\lambda\Omega(f_{\mathbf{z},t})=\lambda\sum_{j=1}^{p}\tau_j\|\boldsymbol{\alpha}_j^{\mathbf{z},t}\|_2\le\mathcal{R}_{\mathbf{z}}(t,0)-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t})\le\frac{1}{t}\log(\frac{\sum_{i=1}^n e^{t\ell(0,y_i)}}{\sum_{i=1}^n e^{t\ell(f_{\mathbf{z},t}(x_i),y_i)}})$$

Then, if $t\in(-\infty,0)$, we have

$$\frac{1}{t}\log(\frac{\sum_{i=1}^n e^{t\ell(0,y_i)}}{\sum_{i=1}^n e^{t\ell(f_{\mathbf{z},t}(x_i),y_i)}})\le-\frac{1}{|t|}\log e^{-|t|M_\ell^2}=M_\ell^2.$$

Also, if $t\in(0,+\infty)$, the similar result can be obtained, i.e.,

$$\frac{1}{t}\log(\frac{\sum_{i=1}^n e^{t\ell(0,y_i)}}{\sum_{i=1}^n e^{t\ell(f_{\mathbf{z},t}(x_i),y_i)}})\le\frac{1}{t}\log e^{tM_\ell^2}=M_\ell^2.$$

Consequently, we see that

$$\sum_{j=1}^{p}\|\boldsymbol{\alpha}_j^{\mathbf{z},t}\|_2\le\frac{M_\ell^2}{\lambda\min_{j=1,\dots,p}\tau_j},\ \forall z=(x,y)\in\mathcal{Z},\ t\in(\infty,0)\cup(0,+\infty)$$

In connection with

$$\|f_{\mathbf{z},t}\|_K\le\tilde{\kappa}\sum_{j=1}^{p}\sum_{i=1}^{n}|\alpha_{ji}^{\mathbf{z},t}|\le\tilde{\kappa}n^{\frac{1}{2}}\sum_{j=1}^{p}\|\boldsymbol{\alpha}_j^{\mathbf{z},t}\|_2,$$

we get the desired result of Lemma B.1. □

Lemma B.1 illustrates a ball

$$\mathcal{B}_r=\{f\in\mathcal{H}_K:\|f\|_K\le r\}$$

that covers the estimator $f_{\mathbf{z},t}$, where

$$r=\tilde{\kappa}n^{\frac{1}{2}}M_\ell^2\lambda^{-1}\tilde{\tau}^{-1}. \tag{10}$$

Next, we use the ℓ_2-empirical covering number (e.g, used in (Zhou, 2002; Anthony & Bartlett, 1999; Chen et al., 2020; Feng et al., 2020)) to measure the capacity of $\mathcal{B}_r$.

Definition B.2. Let $\mathcal{F}$ be a set of measurable functions on $\mathcal{X}$. Given input samples $\mathbf{x} = \{x_i\}_{i=1}^n$, we define the ℓ_2-empirical metric as

$$d_{2,\mathbf{x}}(f_1, f_2) = (\frac{1}{n}\sum_{i=1}^n (f_1(x_i) - f_2(x_i))^2)^{\frac{1}{2}}.$$

Then the ℓ_2-empirical covering number of function space $\mathcal{F}$ is defined as

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_n \sup_{\mathbf{x}} \inf_m \{m \in \mathbb{N} : \exists \{f_j\}_{j=1}^m \subset \mathcal{F}, s.t., \mathcal{F} \subset \cup_{j=1}^m \{f \in \mathcal{F} : d_{2,\mathbf{x}}(f, f_j) < \epsilon\}\}, \forall \epsilon > 0.$$

Indeed, the empirical covering number of $\mathcal{B}_r$ has been investigated extensively in learning theory literatures (Steinwart & Christmann, 2008). There are some detailed examples, e.g., Examples 1-2 in (Guo & Zhou, 2013), Theorem 2 in (Shi et al., 2011) and Lemma 3 in (Shi, 2013).

The following uniform concentration inequality established in (Wu et al., 2007; Steinwart & Christmann, 2008) is used for bounding $\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z},t}(t, \mathcal{P}(f_{\mathbf{z},t}))$.

Lemma B.3. *Let $\mathcal{T}$ be a measurable function set on $\mathcal{Z}$. Suppose that there are some constants B, c and $\vartheta \in [0, 1]$ such that $\|h\|_\infty \le B$, $\mathbb{E}h^2 \le c(\mathbb{E}h)^\vartheta$ for each $h \in \mathcal{T}$. If for $0 < s < 2$, $\log \mathcal{N}_2(\mathcal{T}, \epsilon) \le a\epsilon^{-s}, \forall \epsilon > 0, a > 0$, then for any $\delta \in (0, 1)$ and given sample set $\mathbf{z} = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \in \mathcal{Z}^n$, there holds*

$$\mathbb{E}h - \sum_{i=1}^n h(z_i) \le \frac{\omega^{1-\vartheta}(\mathbb{E}h)^\vartheta}{2} + c_s\omega + 2(\frac{c\log\frac{1}{\delta}}{n})^{\frac{1}{2-\vartheta}} + \frac{18B\log\frac{1}{\delta}}{n}$$

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

$$\omega = \max\{c^{\frac{2-s}{4-2\vartheta+s\vartheta}}(\frac{a}{n})^{\frac{2}{4-2\vartheta+s\vartheta}}, B^{\frac{2-s}{2+s}}(\frac{a}{n})^{\frac{2}{2+s}}\}$$

and c_s is a constant depending on s.

We then present the concentration estimation for $\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))$. For this purpose, we first define a function-based random variable set as

$$\mathcal{V} = \{h(z) := h_f(z) = \frac{1}{t}e^{t\ell(\mathcal{P}(f)(x),y)} : f \in \mathcal{B}_r, z \in \mathcal{Z}\}, t \in (-\infty, 0) \cup (0, +\infty),$$

where r is defined in (10).

Under Assumption 1, for any $f_1, f_2 \in \mathcal{B}_r$, there holds

$$|h_{f_1}(z) - h_{f_2}(z)| = |\frac{1}{t}e^{t\ell(\mathcal{P}(f_1)(x),y)} - \frac{1}{t}e^{t\ell(\mathcal{P}(f_2)(x),y)}| \le \widetilde{M}|f_1(x) - f_2(x)|, \tag{11}$$

where $\widetilde{M} = M_{\ell'}e^{4tM_\ell^2}, \forall t \in (0, +\infty)$ and $\widetilde{M} = M_{\ell'}, \forall t \in (-\infty, 0)$.

Then, we get

$$\log \mathcal{N}_2(\mathcal{V}, \epsilon) \le \log \mathcal{N}_2(\mathcal{B}_r, \epsilon\widetilde{M}^{-1}) \le \log \mathcal{N}_2(\mathcal{B}_1, \epsilon\widetilde{M}^{-1}r^{-1}) \le c_s\widetilde{M}^s r^s p^{1+s}\epsilon^{-s},$$

where the s value is given in Theorem 1 and this inequality comes from the covering number bounds for $\mathcal{H}_{K_j}$ with $K_j \in C^\nu$ (see Theorem 2 in (Shi et al., 2011) and Lemma 3 in (Shi, 2013) for more details). It is trivial to verify that

$$\|h\|_\infty \le \frac{1}{t}e^{tM_\ell^2}\mathbf{1}_{\{0<t<+\infty\}} + \frac{1}{|t|}\mathbf{1}_{\{-\infty<t<0\}} := \widetilde{V}, \tag{12}$$

and

$$\mathbb{E}h^2 \quad \le \quad \mathbb{E}\|h\|_\infty^2 = \widetilde{V}^2(\mathbb{E}h)^0.$$

Now applying Lemma B.3 to the function-based random variable set $\mathcal{V}$ with $B = \widetilde{V}$, $a = \widetilde{M}^s r^s p^{1+s}$, $\vartheta = 0$ and $c = \widetilde{V}^2$, we have, for any $f \in \mathcal{B}_r$ and $0 < \delta < 1$,

$$\begin{aligned} & \mathcal{G}(t, \mathcal{P}(f)) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f)) \\ \le \quad & (\frac{1}{2} + c_s)\max\{\widetilde{V}^{\frac{2-s}{2}}(\frac{\widetilde{M}^s r^s p^{1+s}}{n})^{\frac{1}{2}}, \widetilde{V}^{\frac{2-s}{2+s}}(\frac{\widetilde{M}^s r^s p^{1+s}}{n})^{\frac{2}{2+s}}\} + 2(\frac{\widetilde{V}^2\log\frac{1}{\delta}}{n})^{\frac{1}{2}} + \frac{18\widetilde{V}\log\frac{1}{\delta}}{n} \end{aligned}$$

with confidence at least $1-\delta$.

For any $t \in (-\infty, 0)$, we get $\widetilde{V} = |t|^{-1}$, $r = M_\ell^2 \tilde{\kappa} \tilde{\tau}^{-1} n^{\frac{1}{2}} \lambda^{-1}$ and $\widetilde{M} = M_{\ell'}$. For any $t \in (0, +\infty)$, we have $r = M_\ell^2 \tilde{\kappa} \tilde{\tau}^{-1} n^{\frac{1}{2}} \lambda^{-1}$, $\widetilde{V} = \frac{1}{t} e^{4tM_\ell^2}$ and $\widetilde{M} = M_{\ell'} e^{tM_\ell^2}$.

By taking $\lambda = n^{-\zeta}$, Lemma B.3 gives that

$$\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \lesssim \begin{cases} |t|^{\frac{s-2}{2+s}} n^{\max\{-\frac{1}{2}, \frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (-\infty, -1] \\ |t|^{-1} n^{\max\{-\frac{1}{2}, \frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (-1, 0) \\ |t|^{-1} e^{tM_\ell^2} n^{\max\{-\frac{1}{2}, \frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (0, 1) \\ |t|^{\frac{s-2}{2+s}} e^{tM_\ell^2} n^{\max\{-\frac{1}{2}, \frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in [1, +\infty) \end{cases}$$

with confidence at least $1-\delta$, where $a \lesssim b$ means there exists a positive constant c such that $a \leq cb, \forall a, b \in \mathbb{R}$.

To the end, we get back to the concerned point through following lemma.

Lemma B.4. *Let Assumption 1 be true. There holds*

$$\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \leq e^{M_\ell^2 |t|} |\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))|, \ \ \forall t \in (-\infty, 0),$$

and

$$\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \leq |\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))|, \ \ \forall t \in (0, +\infty).$$

Proof. From the mean value theorem, we have

$$\begin{aligned} \mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) &= \frac{1}{t} \log \int_{\mathcal{Z}} e^{t\ell(\mathcal{P}(f_{\mathbf{z},t})(x), y)} - \frac{1}{t} \log \frac{1}{n} \sum_{i=1}^{n} e^{t\ell(\mathcal{P}(f_{\mathbf{z},t})(x_i), y_i)} \\ &= \frac{1}{\xi} (\frac{1}{t} \int_{\mathcal{Z}} e^{t\ell(\mathcal{P}(f_{\mathbf{z},t})(x), y)} - \frac{1}{nt} \sum_{i=1}^{n} e^{t\ell(\mathcal{P}(f_{\mathbf{z},t})(x_i), y_i)}) \\ &= \frac{1}{\xi} \Big(\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \Big), \end{aligned}$$

where ξ is between $\int_{\mathcal{Z}} e^{t\ell(\mathcal{P}(f_{\mathbf{z},t})(x_i), y_i)}$ and $\frac{1}{n} \sum_{i=1}^{n} e^{t\ell(\mathcal{P}(f_{\mathbf{z},t})(x_i), y_i)}$. It is trivial to see that

$$\frac{1}{\xi} \Big(\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \Big) \leq e^{4|t|M_\ell^2} |\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))|, \forall t \in (-\infty, 0),$$

and

$$\frac{1}{\xi} \Big(\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \Big) \leq |\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t}))|, \forall t \in (0, +\infty).$$

We complete the proof of lemma B.4. □

In connection with Lemmas B.1-B.3, for any $0 < \delta < 1$ and any fixed $t \in (-\infty, 0) \cup (0, +\infty)$, there holds

$$\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) \lesssim n^{\max\{\frac{s-2+2s\zeta}{4}, -\frac{1}{2}\}} \log(1/\delta).$$

with confidence at least $1-\delta$. This completes the proof of Theorem 1.

C. Proof of Theorem 2

Proof sketch: In what follows, we shall make efforts to conduct the function approximation analysis. Clearly, the data-generating model $Y = f^*(X) + \epsilon$ and Assumption 3 together ensure that the ground truth function f^* is a conditional mean function (Wu et al., 2006; Caponnetto & Vito, 2007; Wang & Zhou, 2011; Feng et al., 2015), i.e.,

$$f^*(x) = \int_{\mathcal{Y}} y d\rho(y|x),$$

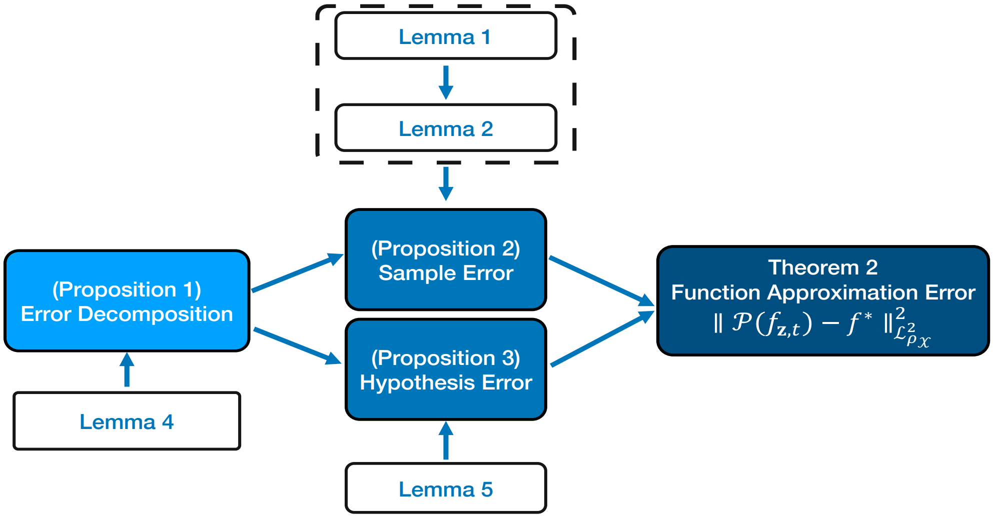

Lemma 1
Lemma 2
(Proposition 2) Sample Error
(Proposition 1) Error Decomposition
Theorem 2 Function Approximation Error $\| \mathcal{P}(f_{\mathbf{z},t}) - f^* \|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$
(Proposition 3) Hypothesis Error
Lemma 4
Lemma 5

Figure 6. An illustration of the main building blocks in function approximation analysis. The error decomposition in Proposition C.2 supports Propositions C.4 and C.6. The strategy for bounding the sample error (Proposition C.4) is similar to that for generalization error analysis (Section B). The upper bound of $\|\boldsymbol{\alpha}_j^\eta\|_2, j = 1, .., p$ in Lemma 5 is key to bound the hypothesis error in Proposition C.6. Finally, Theorem 2 is obtained by combining the results in Propositions C.2-C.6.

where

$$f^* = \arg\min_{f \in \mathcal{H}_K} \mathcal{E}_\rho(f) \;\text{ and }\; \mathcal{E}_\rho(f) := \int_\mathcal{Z} (f(x) - y)^2 d\rho(x, y).$$

In view of the above discussions, Figure 6 presents the main building blocks in function approximation analysis.

Firstly, we measure the deviations between the excess risk terms $\mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*)$ and $\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)$.

Lemma C.1. *Let Assumptions 1-3 be true. There holds*

$$\mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) \le \mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*) + M_{t,f^*},$$

where $M_{t,f^*} = 16M^4|t|e^{|t|(M+\|f^*\|_\infty)^2}$ *for any* $t \in (-\infty, 0)$ *and* $|t|(M + \|f^*\|_\infty)^4$ *for any* $t \in (0, +\infty)$.

Proof. From the definitions of $\mathcal{E}_\rho(f)$ and $\mathcal{R}(t, f)$, we have

$$\begin{aligned}
& \mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) - (\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)) \\
= & \int_\mathcal{Z} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2 - (y - f^*(x))^2 d\rho(x, y) - \frac{1}{t}\Big[\log \int_\mathcal{Z} e^{t[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2} - \log \int_\mathcal{Z} e^{t(y - f^*(x))^2} d\rho(x, y)\Big] \\
= & \int_\mathcal{Z} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2 - (y - f^*(x))^2 d\rho(x, y) - \frac{1}{t}\log\Big[1 + \frac{\int_\mathcal{Z} e^{t[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2} d\rho(x, y)}{\int_\mathcal{Z} e^{t(y - f^*(x))^2} d\rho(x, y)} - 1\Big].
\end{aligned}$$

Taking $t \in (-\infty, 0)$ gives

$$\begin{aligned}
& \mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) - (\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)) \\
\le & \int_\mathcal{Z} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2 - (y - f^*(x))^2 d\rho(x, y) \\
+ & \frac{1}{|t| \int_\mathcal{Z} e^{-|t|(y - f^*(x))^2} d\rho(x, y)} \Big[\int_\mathcal{Z} e^{-|t|[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2} d\rho(x, y) - \int_\mathcal{Z} e^{-|t|(y - f^*(x))^2} d\rho(x, y)\Big] \\
\le & e^{|t|(M+\|f^*\|_\infty)^2} \Big[\int_\mathcal{Z} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2 - (y - f^*(x))^2 d\rho(x, y) \\
& \qquad + \frac{1}{|t|}\Big(\int_\mathcal{Z} e^{-|t|[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2} d\rho(x, y) - \int_\mathcal{Z} e^{-|t|(y - f^*(x))^2} d\rho(x, y)\Big)\Big].
\end{aligned}$$

By applying Taylor's Theorem to both $\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t}))$ and $\mathcal{G}(t, f^*)$, we get

$$\begin{aligned}
& \mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) - (\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)) \\
\leq \; & |t| e^{|t|(M+\|f^*\|_\infty)^2} \Big[\int_{\mathcal{Z}} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^4 e^{\xi_1} - (y - f^*(x))^4 e^{\xi_2} d\rho(x,y) \Big], \\
\leq \; & 16 M^4 |t| e^{|t|(M+\|f^*\|_\infty)^2},
\end{aligned}$$

where ξ_1 is between 0 and $t[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2$, and ξ_2 is between 0 and $t(y - f^*(x))^2$.

Moreover, if $t \in (0, +\infty)$, we can deduce

$$\begin{aligned}
& \mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) - (\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)) \\
\leq \; & \int_{\mathcal{Z}} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2 - (y - f^*(x))^2 d\rho(x,y) \\
+ \; & \frac{1}{t \int_{\mathcal{Z}} e^{t(y - f^*(x))^2} d\rho(x,y)} \Big[\int_{\mathcal{Z}} e^{t(y-f^*(x))^2} d\rho(x,y) - \int_{\mathcal{Z}} e^{t[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2} d\rho(x,y) \Big] \\
\leq \; & \Big[\int_{\mathcal{Z}} [y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2 - (y - f^*(x))^2 d\rho(x,y) \\
& \qquad + \frac{1}{|t|} \Big(\int_{\mathcal{Z}} e^{t(y-f^*(x))^2} d\rho(x,y) - \int_{\mathcal{Z}} e^{t[y - \mathcal{P}(f_{\mathbf{z},t})(x)]^2} d\rho(x,y) \Big) \Big] \\
\leq \; & |t| (M + \|f^*\|_\infty)^4.
\end{aligned}$$

This completes the proof. □

In the following, we make a direct error decomposition.

Proposition C.2. *(Error Decomposition) Under Assumptions 1-3, there holds*

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}_{\rho^2_{\mathcal{X}}}} \leq E_1 + E_2 + \eta \|f^*\|_K^2 + M_{t,f^*},$$

where

$$E_1 = \mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) + \mathcal{R}_{\mathbf{z}}(t, f^*) - \mathcal{R}(t, f^*)$$

and

$$E_2 = \mathcal{R}_{\mathbf{z}}(t, f_{\mathbf{z},t}) + \Omega_{\mathbf{z}}(f_{\mathbf{z},t}) - \mathcal{R}_{\mathbf{z}}(t, f_{\eta,t}) - \eta \|f_{\eta,t}\|_K^2.$$

Proof. The definition of $f_{\eta,t}$ implies that

$$\mathcal{R}_{\mathbf{z}}(t, f_{\eta,t}) + \eta \|f_{\eta,t}\|_K^2 - \eta \|f^*\|_K^2 \leq \mathcal{R}_{\mathbf{z}}(t, f^*).$$

Based on Lemma C.1, we can deduce that

$$\begin{aligned}
\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}_{\rho^2_{\mathcal{X}}}} \leq \; & \mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) \\
\leq \; & \mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*) + M_{t,f^*} \\
\leq \; & \mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) + \mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t, f^*) + \mathcal{R}_{\mathbf{z}}(t, f^*) - \mathcal{R}(t, f^*) + M_{t,f^*} \\
\leq \; & E_1 + \mathcal{R}_{\mathbf{z}}(t, f_{\mathbf{z},t}) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},t}) - \mathcal{R}_{\mathbf{z}}(t, f^*) + M_{t,f^*} \\
\leq \; & E_1 + \mathcal{R}_{\mathbf{z}}(t, f_{\mathbf{z},t}) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},t}) - (\mathcal{R}_{\mathbf{z}}(t, f_{\eta,t}) + \eta \|f_{\eta,t}\|_K^2 - \eta \|f^*\|_K^2) + M_{t,f^*} \\
\leq \; & E_1 + E_2 + \eta \|f^*\|_K^2 + M_{t,f^*}.
\end{aligned}$$

□

In learning theory literatures, we call E_1 , E_2 as the sample error and the hypothesis error, respectively. The sample error E_1 describes the divergence between the generalization error terms $\mathcal{R}_{\mathbf{z}}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t}))$ and $\mathcal{R}_{\mathbf{z}}(t, f^*) - \mathcal{R}(t, f^*)$. The hypothesis error E_2 characterizes the difference between the regularized empirical formulations in $\mathcal{H}_K$ and $\mathcal{H}_{\mathbf{z}}$.

Lemma C.3. *Let ξ be a random variable on a probability space $\mathcal{Z}$ with variance σ satisfying $|\xi - \mathbb{E}\xi| \le M_\xi$ almost surely for some constant M_ξ and for all $z \in \mathcal{Z}$. Then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) - \mathbb{E}\xi \le \frac{2M_\xi \log(1/\delta)}{3n} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}}.$$

Proposition C.4. *(Bounding sample error E_1) Let Assumption 1 be true and each $K_j \in C^v$, $j = 1, ..., p$. For any $\delta \in (0,1)$ and $f \in \mathcal{H}_K$, there holds*

$$E_1 \lesssim \begin{cases} |t|^{\frac{s-2}{2+s}} e^{4|t|M^2} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (-\infty,-1] \\ |t|^{-1} e^{4|t|M^2} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (-1,0) \\ |t|^{-1} e^{8tM^2} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (0,1) \\ |t|^{\frac{s-2}{2+s}} e^{8tM^2} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in [1,+\infty) \end{cases}$$

with confidence at least $1-\delta$, where r, $\widetilde{M}$, $\widetilde{V}$ are defined in (10), (11) and (12), respectively.

Proof. The proof proceeds similarly to the generalization consistency analysis in Section B. After defining a function-based variable set as

$$\mathcal{T} = \{h_f(z) = \frac{1}{t}e^{t[\mathcal{P}(f)(x)-y]^2} - \frac{1}{t}e^{t[f^*(x)-y]^2} : f \in \mathcal{B}_r, g \in \mathcal{H}_K\},$$

Lemma B.3 holds for the function set $\mathcal{T}$ with $B = \widetilde{V}$, $a = \widetilde{M}^s r^s p^{1+s}$, $\vartheta = 0$ and $c = \widetilde{V}^2$, where $\widetilde{M}$ and $\widetilde{V}$ are defined in (11) and (12), respectively. Then we have

$$\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t})) - (\mathcal{G}(t,f^*) - \mathcal{G}_{\mathbf{z}}(t,f^*)) \lesssim \\ \begin{cases} |t|^{\frac{s-2}{2+s}} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (-\infty,-1] \\ |t|^{-1} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (-1,0) \\ |t|^{-1} e^{4tM^2} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in (0,1) \\ |t|^{\frac{s-2}{2+s}} e^{4tM^2} n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}} p^{\frac{2(1+s)}{2+s}} \log(1/\delta), & t \in [1,+\infty) \end{cases}$$

According to Lemma B.4, for any $f \in \mathcal{H}_K$, we have

$$\mathcal{R}(t,f) - \mathcal{R}_{\mathbf{z}}(t,f) = \frac{1}{\xi}(\mathcal{G}(t,f) - \mathcal{G}_{\mathbf{z}}(t,f)),$$

where ξ is between $\int_{\mathcal{Z}} e^{t(f(x)-y)^2}d\rho$ and $\frac{1}{n}\sum_{i=1}^{n} e^{t(f(x_i)-y_i)^2}$. Then we have

$$\begin{aligned} &\Big(\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))\Big) - \Big(\mathcal{R}(t,f^*) - \mathcal{R}_{\mathbf{z}}(t,f^*)\Big) \\ &\qquad - \Big\{\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t})) - (\mathcal{G}(t,f^*) - \mathcal{G}_{\mathbf{z}}(t,f^*))\Big\} \\ = \quad &(\frac{1}{\xi_1} - 1)\Big(\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))\Big) + (1 - \frac{1}{\xi_2})\Big(\mathcal{G}(t,f^*) - \mathcal{G}_{\mathbf{z}}(t,f^*)\Big) \end{aligned}$$

where ξ_1 is between $\int_{\mathcal{Z}} e^{t(\mathcal{P}(f_{\mathbf{z},t})(x)-y)^2}d\rho$ and $\frac{1}{n}\sum_{i=1}^{n} e^{t(\mathcal{P}(f_{\mathbf{z},t})(x_i)-y_i)^2}$ and ξ_2 is between $\int_{\mathcal{Z}} e^{t(f^*(x)-y)^2}d\rho$ and $\frac{1}{n}\sum_{i=1}^{n} e^{t(f^*(x_i)-y_i)^2}$. If $t \in (-\infty, 0)$, we further have

$$\begin{aligned} &\Big(\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))\Big) - \Big(\mathcal{R}(t,f^*) - \mathcal{R}_{\mathbf{z}}(t,f^*)\Big) \\ &\qquad - \Big\{\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t})) - (\mathcal{G}(t,f^*) - \mathcal{G}_{\mathbf{z}}(t,f^*))\Big\} \\ = \quad &e^{4|t|M^2}\Big\{\Big(\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))\Big) + \Big(\mathcal{G}_{\mathbf{z}}(t,f^*) - \mathcal{G}(t,f^*)\Big)\Big\} \end{aligned}$$

According to Lemma B.3, we have

$$\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))\lesssim\begin{cases}|t|^{\frac{s-2}{2+s}}n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}}p^{\frac{2(1+s)}{2+s}}\log(1/\delta), & t\in(-\infty,-1]\\ |t|^{-1}n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}}p^{\frac{2(1+s)}{2+s}}\log(1/\delta), & t\in(-1,0)\end{cases}$$

We define the following function set

$$\tilde{\mathcal{T}}=\{h_f(z)=\frac{1}{t}e^{t[f(x)-y]^2}:f\in\mathcal{H}_K\}.$$

For any $h_f\in\tilde{\mathcal{T}}$, we have $|h_f-\mathbb{E}h_f|\le 2\|h_f\|_\infty\le\widetilde{V}$ and $\sigma^2=\mathbb{E}h_f^2-(\mathbb{E}h_f)^2\le\|h_f\|_\infty^2\le\widetilde{V}^2$. Then, from Lemma C.3, we have

$$\mathcal{G}_{\mathbf{z}}(t,f^*)-\mathcal{G}(t,f^*)\le\frac{4\log(1/\delta)}{3|t|n}+\sqrt{\frac{2\log(1/\delta)}{t^2n}}\lesssim|t|^{-1}n^{-\frac{1}{2}}\sqrt{\log(1/\delta)}$$

Moreover, for $t\in(0,+\infty)$, we have

$$\begin{aligned}&\Big(\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{R}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))\Big)-\Big(\mathcal{R}(t,f^*)-\mathcal{R}_{\mathbf{z}}(t,f^*)\Big)\\&\qquad-\Big\{\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))-(\mathcal{G}(t,f^*)-\mathcal{G}_{\mathbf{z}}(t,f^*))\Big\}\\=\;&(1-\frac{1}{\xi_1})\Big(\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))\Big)+(1-\frac{1}{\xi_2})\Big(\mathcal{G}(t,f^*)-\mathcal{G}_{\mathbf{z}}(t,f^*)\Big)\end{aligned}$$

We also have

$$\mathcal{G}_{\mathbf{z}}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))\le\frac{4e^{4tM^2}\log(1/\delta)}{3nt}+\sqrt{\frac{2e^{4tM^2}\log(1/\delta)}{t^2n}}\lesssim t^{-1}e^{4tM^2}n^{-\frac{1}{2}}\sqrt{\log(1/\delta)}.$$

and

$$\mathcal{G}(t,f^*)-\mathcal{G}_{\mathbf{z}}(t,f^*)\lesssim\begin{cases}t^{-1}e^{4tM^2}n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}}p^{\frac{2(1+s)}{2+s}}\log(1/\delta), & t\in(0,1)\\ t^{\frac{s-2}{2+s}}e^{4tM^2}n^{\max\{-\frac{1}{2},\frac{s-2+2s\zeta}{4}\}}p^{\frac{2(1+s)}{2+s}}\log(1/\delta), & t\in[1,+\infty).\end{cases}$$

We get the desired result by combining the above inequalities. □

To bound the hypothesis error E_2, we first illustrate a key property for the coefficient of $f_{\eta,t}$.

Lemma C.5. *Under Assumption 2, we have*

$$\tau_j\|\boldsymbol{\alpha}_j^\eta\|_2\le M_{t,n,\eta},\forall j=1,...,p,$$

where $M_{t,n,\eta}=\frac{2(M+\|f_{\eta,t}\|_\infty)e^{t(M+\|f_{\eta,t}\|_\infty)}}{n^{\frac{1}{2}}\eta}$ *for any* $t\in(-\infty,0)\cup(0,+\infty)$.

Proof. From the definition of $f_{\eta,t}$, we can deduce that

$$\nabla\mathcal{R}_{\mathbf{z}}(t,f_{\eta,t})+\eta\sum_{j=1}^p\tau_j\nabla\|f_{\eta,t,j}\|_K^2=0.$$

Based on $f_{\eta,t}(x_i)=\mathbf{K}_i^T\boldsymbol{\alpha}^\eta$, simple computations show that

$$\frac{\sum_{i=1}^n 2e^{t(f_{\eta,t}(x_i)-y_i)^2}(f_{\eta,t}(x_i)-y_i)K_{ji}}{\sum_{i=1}^n e^{t(f_{\eta,t}(x_i)-y_i)^2}}+\eta\tau_j(\boldsymbol{\alpha}_j^\eta)^T\widetilde{\mathbf{K}}_j=0,\forall j=1,...,p,$$

where $\widetilde{\mathbf{K}}_j=(K_j(x_{ij},x_{sj}))_{i,s=1}^n\in\mathbb{R}^{n\times n},j=1,...,p$. Consequently, for any $j=1,...,p,$, we have the following equivalent form

$$\Big(\frac{2e^{t(f_{\eta,t}(x_1)-y_1)^2}(f_{\eta,t}(x_1)-y_1)}{\sum_{i=1}^n e^{t(f_{\eta,t}(x_i)-y_i)^2}},...,\frac{2e^{t(f_{\eta,t}(x_n)-y_n)^2}(f_{\eta,t}(x_n)-y_n)}{\sum_{i=1}^n e^{t(f_{\eta,t}(x_i)-y_i)^2}}\Big)\widetilde{\mathbf{K}}_j=-\eta\tau_j(\boldsymbol{\alpha}_j^\eta)^T\widetilde{\mathbf{K}}_j,$$

Since $\widetilde{\mathbf{K}}_j, j = 1, ..., p$ is positive-definite, we further get

$$\Big(\frac{2e^{t(f_{\eta,t}(x_1)-y_1)^2}(f_{\eta,t}(x_1) - y_1)}{\sum_{i=1}^n e^{t(f_{\eta,t}(x_i)-y_i)^2}}, ..., \frac{2e^{t(f_{\eta,t}(x_n)-y_n)^2}(f_{\eta,t}(x_n) - y_n)}{\sum_{i=1}^n e^{t(f_{\eta,t}(x_i)-y_i)^2}}\Big)^T = -\eta\tau_j\boldsymbol{\alpha}_j^\eta, \forall j = 1, ..., p.$$

For any $j = 1, ..., p$, it follows that

$$\tau_j\|\boldsymbol{\alpha}_j^\eta\|_2 = \frac{1}{\eta}\sqrt{\frac{\sum_{i=1}^n[2e^{t(f_{\eta,t}(x_i)-y_i)^2}(f_{\eta,t}(x_i) - y_i)]^2}{(\sum_{i=1}^n e^{t(f_{\eta,t}(x_i)-y_i)^2})^2}} \leq \frac{2\|e^{|t|(f_{\eta,t}(x)-y)^2}(f_{\eta,t}(x) - y)\|_\infty}{n^{\frac{1}{2}}\eta} \leq M_{t,n,\eta},$$

where $M_{t,n,\eta} = \frac{2(M+\|f_{\eta,t}\|_\infty)e^{|t|(M+\|f_{\eta,t}\|_\infty)^2}}{n^{\frac{1}{2}}\eta}, \forall t \in (-\infty, 0) \cup (0, +\infty)$. This completes the proof. □

Proposition C.6. *(Bounding hypothesis error E_2) Under Assumption 2, the hypothesis error E_2 satisfies*

$$E_2 \leq \lambda\sum_{j=1}^p \tau_j\|\boldsymbol{\alpha}_j^\eta\|_2 \leq 2\lambda p M_{t,n,\eta}.$$

Proof. From the definition of $f_{\mathbf{z},t}$, we know that

$$\mathcal{R}_{\mathbf{z}}(t, f_{\mathbf{z},t}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},t}) \leq \mathcal{R}_{\mathbf{z}}(t, f_{\eta,t}) + \lambda\Omega_{\mathbf{z}}(f_{\eta,t}).$$

Then,

$$\begin{aligned} E_2 &= \mathcal{R}_{\mathbf{z}}(t, f_{\mathbf{z},t}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},t}) - (\mathcal{R}_{\mathbf{z}}(f_{\eta,t}) + \lambda\Omega_{\mathbf{z}}(f_{\eta,t})) + \lambda\Omega_{\mathbf{z}}(f_{\eta,t}) - \eta\|f_{\eta,t}\|_K^2 \\ &\leq \lambda\Omega_{\mathbf{z}}(f_{\eta,t}) - \eta\|f_{\eta,t}\|_K^2 \leq \lambda\Omega_{\mathbf{z}}(f_{\eta,t}) \end{aligned}$$

The desired result follows by combining the above inequality with Lemma C.5 □

Set $\eta = n^{-\frac{1}{4}}\lambda^{\frac{1}{2}}$, $\lambda = n^{-\zeta}$, and $t = \pm n^{-\beta}$. Combining Propositions C.2-C.6, we have

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 \lesssim n^{\max\{\frac{4\beta+s-2+2s\zeta}{4}, \frac{2\beta-1}{2}, \frac{-1-2\zeta}{4}, -\beta\}}\log(1/\delta)$$

with confidence at least $1 - \delta$. This completes the proof of Theorem 2.

D. Proof of Theorem 3

Proof sketch: The strategy for Theorem 3 is similar to that for Theorem 2 except that we need to reestablish the relation between $\mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*)$ and $\mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}(t, f^*)$.

Lemma D.1. *Let Assumptions 1-2 and 4 be true. When t value is fixed and $t \in (-\frac{1}{2(2M_f+M_0)^2}, 0)$, there holds*

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 = \mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_\rho(f^*) \lesssim \mathcal{G}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{G}(t, f^*).$$

Proof. For any $f \in \mathcal{H}_K$, we can see that

$$\begin{aligned} \mathcal{R}(t, f) &= \frac{1}{t}\log\int_{\mathcal{Z}} e^{t(f(x)-y)^2}d\rho(x, y) = \frac{1}{t}\log\int_{\mathcal{X}}\int_{\mathbb{R}} e^{t(f(x)-u-f^*(x))^2}p_{\epsilon|X}(u)dud\rho_{\mathcal{X}}(x) \\ &= \frac{1}{t}\log\int_{\mathcal{X}} E(f(x) - f^*(x))d\rho_{\mathcal{X}}(x), \end{aligned}$$

where $E(m) = \int_{\mathbb{R}} e^{t(u-m)^2}p_{\epsilon|X}(u)du$.

By computing the derivative of $E(f(x)-f^*(x))$ w.r.t $f(x)-f^*(x)$, we get

$$\begin{aligned}
\mathcal{R}'(t,f) &= \frac{\int_{\mathcal{X}} E'(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)}{t\int_{\mathcal{X}} E(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)} \\
&= \frac{-2\int_{\mathcal{X}}\int_{\mathbb{R}} e^{t[u-(f(x)-f^*(x))]^2}[u-(f(x)-f^*(x))]p_{\epsilon|X=x}(u)dud\rho_{\mathcal{X}}(x)}{\int_{\mathcal{X}} E(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)} \\
&= \frac{-2\int_{\mathcal{X}}\int_{\mathbb{R}} me^{tm^2}p_{\epsilon|X=x}(f(x)-f^*(x)+m)dmd\rho_{\mathcal{X}}(x)}{\int_{\mathcal{X}} E(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)},
\end{aligned}$$

where $m=u-(f(x)-f^*(x))$. Under noise Assumption 3, there holds $\int_{\mathbb{R}} me^{-|t|m^2}p_{\epsilon|X}(m)dm=0$. Thus, we can deduce that $\mathcal{R}'(t,f)=0$ holds only when $f(x)=f^*(x)$ for any $x\in\mathcal{X}$. Furthermore,

$$\begin{aligned}
\mathcal{R}''(t,f) &= \frac{\int_{\mathcal{X}} E''(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)}{t\int_{\mathcal{X}} E(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)} \\
&= \frac{\int_{\mathcal{X}}\int_{\mathbb{R}} e^{-|t|m^2}(2-4|t|(u-m)^2)p_{\epsilon|X}(m+f(x)-f^*(x))dmd\rho_{\mathcal{X}}(x)}{\int_{\mathcal{X}} E(f(x)-f^*(x))d\rho_{\mathcal{X}}(x)},
\end{aligned}$$

where $m=u-(f(x)-f^*(x))$. Then, when $2-4|t|(u-m)^2>0$, i.e., $t\in(-\frac{1}{2(M+\|f^*\|_\infty)^2},0)\cup(0,\frac{1}{2(M+\|f^*\|_\infty)^2})$, we have $\mathcal{R}''(t,f)>0$. The above-discussed results show that $f=f^*$ is the unique minimizer of $\mathcal{R}(t,f)$. Then we can deduce that

$$\begin{aligned}
&\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{R}(t,f^*) \\
=\ & \frac{1}{|t|}\Big[\log\int_{\mathcal{X}} E(0)d\rho_{\mathcal{X}}(x)-\log\int_{\mathcal{X}} E(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))d\rho_{\mathcal{X}}(x)\Big] \\
=\ & \frac{1}{|t|\xi_1}\int_{\mathcal{X}} E(0)-E(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))d\rho_{\mathcal{X}}(x) \\
=\ & \frac{1}{|t|\xi_1}\int_{\mathcal{X}}[-E'(0)(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))-\frac{E''(\xi)}{2}(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))^2]d\rho_{\mathcal{X}}(x) \\
=\ & \frac{1}{|t|\xi_1}\int_{\mathcal{X}}\frac{-E''(\xi_2)}{2}(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))^2d\rho_{\mathcal{X}}(x),
\end{aligned}$$

where ξ_1 is between $\int_{\mathcal{X}} E(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))d\rho_{\mathcal{X}}(x)$ and $\int_{\mathcal{X}} E(0)d\rho_{\mathcal{X}}(x)$, and ξ_2 is between 0 and $\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x)$. Under the condition $t\in(-\frac{1}{2(M+\|f^*\|_\infty)^2},0)$, the inequality $m^2\le(M+\|f^*\|_\infty+\|p_{\epsilon|X}\|_\infty)^2$ yields

$$\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}(t,f^*)=\frac{1}{t}\int_{\mathcal{X}}\frac{E''(\xi)}{2}(\mathcal{P}(f_{\mathbf{z},t})(x)-f^*(x))^2d\rho_{\mathcal{X}}(x)\ge\widetilde{C}[\mathcal{E}_\rho(\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{E}_\rho(f^*)],$$

where

$$\widetilde{C}=[1-2|t|(M+\|f^*\|_\infty)^2]e^{-|t|(M+\|p_{\epsilon|X}\|_\infty+\|f^*\|_\infty)^2}\int_{\mathbb{R}}p_{\epsilon|X}(u)du$$

is a positive constant. This completes the proof. □

Lemma D.1 in connection with the idea of error decomposition in Proposition C.2 yield that

$$\begin{aligned}
\|\mathcal{P}(f_{\mathbf{z},t})-f^*\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} &\le \mathcal{E}(\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{E}(f^*)\le\widetilde{C}^{-1}(\mathcal{G}(t,\mathcal{P}(f_{\mathbf{z},t}))-\mathcal{G}(t,f^*)) \\
&\le \widetilde{C}^{-1}(E_1+E_2+\eta\|f^*\|_K^2)\lesssim E_1+E_2+\eta\|f^*\|_K^2.
\end{aligned}$$

Let t be fixed and satisfy $t\in(-\frac{1}{2(M+\|f^*\|_\infty+\|p_{\epsilon|X}\|_\infty)^2},0)$. Combining the above decomposition with Propositions C.4-C.6, we have

$$\|\mathcal{P}(f_{\mathbf{z},t})-f^*\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}\lesssim(n^{\frac{s-2}{4}}\lambda^{-\frac{s}{2}}+n^{-\frac{1}{2}}+n^{-\frac{1}{2}}\eta^{-1}\lambda+\eta)\log(1/\delta)$$

with confidence at least $1-\delta$. Moreover, the desired result in Theorem 3 follows by setting $n^{-\frac{1}{2}}\eta^{-1}\lambda=\eta$ and $\lambda=n^{-\zeta}$.

E. Proof of Theorem 4

The data-generating model $Y = f^*(X) + \epsilon$ and Assumption 5 together ensure that the ground truth function f^* is conditional mode function (Sager & Thisted, 1982; Yao & Li, 2013; Feng et al., 2020; Wang et al., 2017), i.e.,

$$f^* = \arg\max_t p_{Y|X}(t|X = x),$$

with

$$f^* = \arg\min_{f \in \mathcal{H}_K} \mathcal{E}_M(f), \ \text{ where } \ \mathcal{E}_M(f) := -\int_{\mathcal{Z}} p_{Y|X}(f(x)|X = x) d\rho_{\mathcal{X}}(x).$$

The data-generating model $Y = f^*(X) + \epsilon$ and Assumption 5 ensure that $f^* = f_M^*$ is true. The proof of Theorem 4 proceeds similarly to Theorem 2 except that we need to consider the difference between the two excess risk terms $\mathcal{E}_M(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_M(f^*)$ and $\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)$ under Assumption 5.

Lemma D.2. *Under Assumptions 5, we can deduce that*

$$\mathcal{E}_M(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_M(f^*) \le \pi^{-\frac{1}{2}} |t|^{3/2} [\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)] + |t|^{-1} M_{\pi,\epsilon}, \forall t \in (-\infty, 0),$$

and

$$\mathcal{E}_M(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_M(f^*) \le \pi^{-\frac{1}{2}} t^{3/2} [\mathcal{R}(t, \mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t, f^*)] + 2(\|p_{\epsilon|X}\|_\infty + \frac{\sqrt{t} e^{4tM^2}}{\sqrt{\pi}}), \forall t \in (0, +\infty),$$

where $M_{\pi,\epsilon} = \pi^{-\frac{1}{2}} \|p''_{\epsilon|X}(\cdot|X = x)\|_\infty \int_{\mathbb{R}} u^2 e^{-u^2} du$ is a positive constant.

Proof. From the model assumption $\epsilon = Y - f^*(X)$, we have

$$\int_{\mathcal{Z}} -p_{Y|X}(f(x)|X = x) d\rho_{\mathcal{X}}(x) = \int_{\mathcal{Z}} -p_{\epsilon|X}(f(x) - f^*(x)|X = x) d\rho_{\mathcal{X}}(x).$$

For any $f \in \mathcal{H}_K$, direct computations show

$$\begin{aligned}
& \mathcal{E}_M(f) - \frac{|t|^{3/2}}{\sqrt{\pi}} \mathcal{R}(t, f) \\
= \ & \int_{\mathcal{Z}} -p_{Y|X}(f(x)|X = x) d\rho_{\mathcal{X}}(x) - \frac{|t|^{3/2}}{\sqrt{\pi} t} \log \int_{\mathcal{Z}} e^{t(f(x)-y)^2} \\
= \ & \int_{\mathcal{Z}} -p_{\epsilon|X}(f(x) - f^*(x)|X = x) d\rho_{\mathcal{X}}(x) - \frac{|t|^{3/2}}{\sqrt{\pi} t} \log \int_{\mathcal{X}} \int_{\mathbb{R}} e^{t(v - f(x) + f^*(x))^2} p_{\epsilon|X}(v|X = x) dv \rho_{\mathcal{X}}(x) \\
= \ & \int_{\mathcal{Z}} -p_{\epsilon|X}(f(x) - f^*(x)|X = x) d\rho_{\mathcal{X}}(x) - \frac{|t|^{3/2}}{\sqrt{\pi} t} \log \int_{\mathcal{X}} \int_{\mathbb{R}} e^{tu^2} p_{\epsilon|X}(u + f(x) - f^*(x)|X = x) du \rho_{\mathcal{X}}(x).
\end{aligned}$$

By applying Taylor's Theorem to density function $p_{\epsilon|X}(\cdot)$, for any $t \in (-\infty, 0)$, we get

$$\begin{aligned}
& -\frac{|t|^{3/2}}{t\sqrt{\pi}} \log \int_{\mathcal{X}} \int_{\mathbb{R}} e^{tu^2} p_{\epsilon|X}(u + f(x) - f^*(x)|X = x) du \rho_{\mathcal{X}}(x) \\
\le \ & \frac{|t|^{1/2}}{\sqrt{\pi}} \int_{\mathcal{X}} \int_{\mathbb{R}} e^{-|t|u^2} p_{\epsilon|X}(u + f(x) - f^*(x)|X = x) du \rho_{\mathcal{X}}(x) \\
= \ & \frac{|t|^{1/2}}{\sqrt{\pi}} \int_{\mathcal{X}} \int_{\mathbb{R}} e^{-|t|u^2} [p_{\epsilon|X}(f(x) - f^*(x)|X = x) + u p'_{\epsilon|X}(f(x) - f^*(x)|X = x) \\
& \qquad + \frac{u^2}{2} p''_{\epsilon|X}(\xi|X = x)] du \rho_{\mathcal{X}}(x),
\end{aligned}$$

where ξ is between $f(x) - f^*(x)$ and $f(x) - f^*(x) + u$. We then have $\int_{\mathbb{R}} u e^{-|t|u^2} du = 0$ and $\int_{\mathbb{R}} e^{-|t|u^2} du = \frac{\sqrt{\pi}}{\sqrt{|t|}}$ for any $t \in (-\infty, 0)$. Therefore,

$$\begin{aligned}
|\mathcal{E}_M(f) - \frac{|t|^{3/2}}{\sqrt{\pi}} \mathcal{R}(t, f)| &= \int_{\mathcal{X}} -p_{Y|X}(f(x)|X = x) d\rho_{\mathcal{X}}(x) - \frac{|t|^{1/2}}{\sqrt{\pi}} \int_{\mathcal{Z}} e^{-|t|(f(x)-y)^2} \\
= \ & \frac{1}{2\sqrt{\pi}|t|} \int_{\mathcal{X}} \int_{\mathbb{R}} e^{u^2} u^2 p''_{\epsilon|X}(\xi|X = x) du d\rho_{\mathcal{X}}(x) \le \frac{\|p''_{\epsilon|X}(\cdot|X = x)\|_\infty}{2\sqrt{\pi}|t|} \int_{\mathbb{R}} u^2 e^{-u^2} du.
\end{aligned}$$

For any $t \in (0, +\infty)$, there holds

$$\begin{aligned}\mathcal{E}_M(f) - \frac{t^{3/2}}{\sqrt{\pi}}\mathcal{R}(t,f) &= -[\int_{\mathcal{X}} p_{Y|X}(f(x)|X=x)d\rho_{\mathcal{X}}(x) + \frac{\sqrt{t}}{\sqrt{\pi}}\log\int_{\mathcal{Z}} e^{t(f(x)-y)^2}] \\ &\geq -(\|p_{\epsilon|X}\|_\infty + \frac{4t^{3/2}M^2}{\sqrt{\pi}}).\end{aligned}$$

For any $t \in (-\infty, 0)$, we get

$$|\mathcal{E}_M(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_M(f^*) - \frac{|t|^{3/2}}{\sqrt{\pi}}[\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t,f^*)]| \leq \frac{\|p''_{\epsilon|X}(\cdot|X=x)\|_\infty}{\sqrt{\pi}|t|}\int_{\mathbb{R}} u^2e^{-u^2}du.$$

We complete the proof by denoting $M_{\pi,\epsilon} = \frac{\|p''_{\epsilon|X}(\cdot|X=x)\|_\infty}{\sqrt{\pi}}\int_{\mathbb{R}} u^2e^{-u^2}du$. □

From Lemma D.2, the difference between $\mathcal{E}_M(f)$ and $\frac{t^{3/2}}{\sqrt{\pi}}\mathcal{R}(t,f)$ always exists for any $t \in (0,+\infty)$. Therefore, we only focus on the function approximation performance when $t \in (-\infty, 0)$.

Lemma D.3. *For any $t \in (-\infty, 0)$, there holds*

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \frac{|t|^{3/2}}{\sqrt{\pi}}[E_1 + E_2 + \eta\|f^*\|^2_K] + M_{\pi,\epsilon}|t|^{-1},$$

where E_1 and E_2 are bounded respectively in Propositions C.4-C.6.

Proof. Similar with the error decomposition in Proposition C.2 and D.2, we get

$$\begin{aligned}\mathcal{E}_M(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_M(f^*) &\leq \frac{|t|^{3/2}}{\sqrt{\pi}}(\mathcal{R}(t,\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{R}(t,f^*_M)) + M_{\pi,\epsilon}|t|^{-1} \\ &\leq \frac{|t|^{3/2}}{\sqrt{\pi}}[E_1 + E_2 + \eta\|f^*\|^2_K] + M_{\pi,\epsilon}|t|^{-1}.\end{aligned}$$

Following the Theorem 19 in (Feng et al., 2020), under Assumption 5, one can conclude that

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \mathcal{E}_M(\mathcal{P}(f_{\mathbf{z},t})) - \mathcal{E}_M(f^*).$$

This completes the proof. □

For any $t \in (-\infty, 0)$, we get $\widetilde{V} = |t|^{-1}$, $r = \tilde{\kappa}\tilde{\tau}^{-1}n^{\frac{1}{2}}|t|^{-1}\lambda^{-1}$ and $\widetilde{M} = 4M$. Setting $|t|^{\frac{3}{2}}n^{-\frac{1}{2}}\eta^{-1}\lambda = |t|^{\frac{3}{2}}\eta$, $\lambda = n^{-\zeta}$ and $t = \log n^{-\beta}$, based on Propositions C.4-C.6 and Lemma D.3, we get

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim [\tilde{O}(n^{\max\{\frac{16\beta M^2+s-2+2s\zeta}{4},\frac{8\beta M^2-1}{2},\frac{1-2\zeta}{4},\beta\}}) + O(\log^{-1}n)]\log(1/\delta).$$

When $0 < \beta \leq \min\{\frac{1}{8M^2}, \frac{2-s-2s\zeta}{16M^2}\}$ and $\frac{2-s}{2s} > \zeta \geq \frac{1}{2}$, we have

$$\|\mathcal{P}(f_{\mathbf{z},t}) - f^*\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \log^{-1}n\log(1/\delta).$$

This completes the proof of Theorem 4.

E. Variable Selection Consistency

In this section, we aim to investigate the variable selection consistency of T-SpAM.

Denote $\mathcal{S}^* = \{1, ..., p^*\}$ with $p^* \leq p$ as the set of truly informative variables. Theoretically, we expect the ℓ_2-norm $\|\boldsymbol{\alpha}^{\mathbf{z}}_j\|_2 = 0$ for any $j \in \{p^*+1, ..., p\}$. In practice, we screen out the informative variables through $\mathcal{S}^{\mathbf{z}} = \{j : \|\boldsymbol{\alpha}^{\mathbf{z}}_j\|_2 \geq v, j = 1, ..., p\}$ with a positive threshold value v. It is meaningful to investigate the relation between $\mathcal{S}^{\mathbf{z}}$ and $\mathcal{S}^*$.

Similar with the assumptions in (Yang et al., 2016; Wang et al., 2017; Chen et al., 2020), we require some parametric conditions for our analysis.

Assumption E.1. There is a positive constant c_τ such that $\max_{j\in S^*}\tau_j \le c_\tau \le \min_{l\notin S^*}\tau_l$.

Theorem E.2. *Under Assumptions 3.2-3.5, we have* $\mathcal{S}_{\mathbf{z}} \subset \mathcal{S}^*$ *by taking* $2c_\tau^{-\frac{1}{2}}\tilde{\kappa}^{\frac{1}{2}}M^{\frac{1}{2}}C_{n,t}^{\frac{1}{2}}+2\sqrt{2}\tilde{\kappa}^{-\frac{3}{2}}|t|^{\frac{1}{2}}MC_{n,t} \le \lambda \le C_{n,t}$, *where* $C_{n,t}=\tilde{\kappa}^2 n^{\frac{1}{2}}|t|^{-1}\tilde{\tau}^{-1}M^{-1}e^{|t|M^2}$ *and* $\tilde{\tau}=\min_{j=1,\ldots,p}\tau_j$.

Theorem E.2 illustrates that T-SpAM can identify the truly informative variables by taking properly λ and weight $\tau_j, j = 1,...,p$. Indeed, the current analysis extends Theorem 4 in (Wang et al., 2017) from a linear regularized modal regression to the nonlinear T-SpAM. Moreover, it is interesting to further explore variable selection analysis by replacing the parameter conditions here with the incoherence assumptions (e.g. Assumption 4 in (Lv et al., 2018)).

Proof. From the definition of $\boldsymbol{\alpha}^{\mathbf{z},t}$, we can deduce that

$$\frac{\partial \frac{1}{t}\log(\frac{1}{n}\sum_{i=1}^n e^{t(f_{\mathbf{z},t}(x_i)-y_i)^2})}{\partial \boldsymbol{\alpha}_j} + \lambda\tau_j\frac{\boldsymbol{\alpha}_j}{\|\boldsymbol{\alpha}_j\|_2}=0$$

for any $\boldsymbol{\alpha}_j^{\mathbf{z},t}, j\in\{1,...,p\}$ satisfying $\|\boldsymbol{\alpha}_j^{\mathbf{z},t}\|_2\neq 0$.

Direct computation shows that

$$2\sum_{i=1}^n e^{t((f_{\mathbf{z},t}(x_i)-y_i)^2-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t}))}(f_{\mathbf{z},t}(x_i)-y_i)K_{ji}=\lambda\tau_j\frac{\boldsymbol{\alpha}_j^{\mathbf{z},t}}{\|\boldsymbol{\alpha}_j^{\mathbf{z},t}\|_2}.$$

Taking ℓ_2-norm on the both sides, we derive that

$$2\|\sum_{i=1}^n e^{t((f_{\mathbf{z},t}(x_i)-y_i)^2-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t}))}(f_{\mathbf{z},t}(x_i)-y_i)K_{ji}\|_2=\lambda\tau_j.$$

Suppose that $\|\boldsymbol{\alpha}_j^{\mathbf{z},t}\|_2\neq 0$ for j-th informative variable. Under Assumption 1, we get

$$\begin{aligned} & 2\|\sum_{i=1}^n e^{t((f_{\mathbf{z},t}(x_i)-y_i)^2-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t}))}(f_{\mathbf{z},t}(x_i)-y_i)K_{ji}\|_2 \\ \le\ & 2\sqrt{n}\tilde{\kappa}\sum_{i=1}^n |e^{t((f_{\mathbf{z},t}(x_i)-y_i)^2-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t}))}(f_{\mathbf{z},t}(x_i)-y_i)| \end{aligned}$$

The reproducing property assures that $\|f_{\mathbf{z},t}\|_\infty \le \tilde{\kappa}\|f_{\mathbf{z},t}\|_K$. Moreover, Lemma B.1 shows that

$$\|f_{\mathbf{z},t}\|_K \le \tilde{\kappa}n^{\frac{1}{2}}|t|^{-1}\lambda^{-1}\tilde{\tau}^{-1}e^{|t|M^2},\ \ \forall t\in(-\infty,0)\cup(0,+\infty),$$

where $\tilde{\tau}=\min_{j=1,\ldots,p}\tau_j$ is positive constant. Under Assumption 1, we can further deduce that

$$\begin{aligned} & 2\sqrt{n}\tilde{\kappa}\sum_{i=1}^n |e^{t((f_{\mathbf{z},t}(x_i)-y_i)^2-\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t}))}(f_{\mathbf{z},t}(x_i)-y_i)| \\ \le\ & 2\sqrt{n}\tilde{\kappa}\sum_{i=1}^n e^{|t|((\|f_{\mathbf{z},t}\|_\infty+M)^2+\|\mathcal{R}_{\mathbf{z}}(t,f_{\mathbf{z},t})\|_\infty)}(\|f_{\mathbf{z},t}\|_\infty+M) \\ \le\ & 2\tilde{\kappa}e^{2|t|(\|f_{\mathbf{z},t}\|_\infty+M)^2}(\|f_{\mathbf{z},t}\|_\infty+M)n^{\frac{3}{2}} \end{aligned}$$

By setting $\lambda \le M^{-1}\tilde{\kappa}^2 n^{\frac{1}{2}}|t|^{-1}\tilde{\tau}^{-1}e^{|t|M^2}$, we have

$$2\tilde{\kappa}e^{2|t|(\|f_{\mathbf{z},t}\|_\infty+M)^2}(\|f_{\mathbf{z},t}\|_\infty+M)n^{\frac{3}{2}} \le 4\tilde{\kappa}^3 n^2|t|^{-1}\tilde{\tau}^{-1}\lambda^{-1}e^{8\tilde{\kappa}^4 n|t|^{-1}\tilde{\tau}^{-2}\lambda^{-2}}.$$

Then

$$\lambda^2 \le 4\tilde{\kappa}^3 n^2|t|^{-1}\tilde{\tau}^{-1}\tau_j^{-1}e^{|t|M^2}e^{8\tilde{\kappa}^4 n|t|^{-1}\tilde{\tau}^{-2}\lambda^{-2}e^{2|t|M^2}}.$$

By taking logarithmic function on both side, we get

$$\lambda^2 \log \frac{\lambda^2}{C} \leq 8\tilde{\kappa}^4 n|t|^{-1}\tilde{\tau}^{-2}e^{2|t|M^2},$$

where $C = 4\tilde{\kappa}^3 n^2 |t|^{-1}\tilde{\tau}^{-1}\tau_j^{-1}e^{|t|M^2}$. By setting $\lambda \geq 2\tilde{\kappa}^{3/2}n|t|^{-\frac{1}{2}}\tilde{\tau}^{-\frac{1}{2}}\tau_j^{-\frac{1}{2}}e^{\frac{|t|M^2}{2}}$, we derive

$$\lambda \leq e^{\frac{2\sqrt{2}\kappa^2 n^{\frac{1}{2}}|t|^{-\frac{1}{2}}\tilde{\tau}^{-1}e^{|t|M^2}}{2}} + 2\tilde{\kappa}^{3/2}n|t|^{-\frac{1}{2}}\tilde{\tau}^{-\frac{1}{2}}\tau_j^{-\frac{1}{2}}e^{\frac{|t|M^2}{2}}.$$

Under Assumption 5, for any $j \notin \mathcal{S}^*$, the above inequality guarantees that

$$\lambda \leq e^{\frac{2\sqrt{2}\kappa^2 n^{\frac{1}{2}}|t|^{-\frac{1}{2}}\tilde{\tau}^{-1}e^{|t|M^2}}{2}} + 2\tilde{\kappa}^{3/2}n|t|^{-\frac{1}{2}}\tilde{\tau}^{-\frac{1}{2}}c_\tau^{-\frac{1}{2}}e^{\frac{|t|M^2}{2}}.$$

This inequality contradicts with the parameter condition in Theorem 5. Therefore we have $\|\boldsymbol{\alpha}_j^{\mathbf{z}}\|_2 = 0$ for any $j \notin \mathcal{S}^*$. □

F. Optimization

In spite of the rich representation power of kernel-based algorithm, it suffers from the high computational cost with large-scale data. Random Fourier features have shown potential for accelerating the training associated with kernel methods and may achieve even better results (Rahimi & Recht, 2007; Li et al., 2021b). The main idea of random Fourier acceleration is to approximate kernel evaluation $K_j(\cdot,\cdot), j = 1, ..., p$ by

$$K_j(x_{ij}, x_{tj}) \approx \psi_j(x_{ij})^T \psi(x_{tj}),\ \forall i, t \in \{1, ..., n\},$$

where $\psi_j : \mathbb{R} \to \mathbb{R}^d$ is a random Fourier feature map constructed by the Algorithm 1 in (Rahimi & Recht, 2007).

Denote $\boldsymbol{\psi}(x_i) = (\psi_1(x_{i1})^T, \psi_2(x_{i2})^T, \ldots, \psi_p(x_{ip})^T)^T \in \mathbb{R}^{pd}$. Recalling the optimization objectives in (6) and (8)

$$\hat{W} = \mathop{\arg\min}_{W=(w_1^T, w_2^T, \ldots, w_p^T)^T} \left\{\mathcal{R}_{\boldsymbol{z},\psi}(t, W) + \lambda\Omega_{\mathbf{z}}(W)\right\}, \tag{13}$$

where

$$\mathcal{R}_{\boldsymbol{z},\psi}(t, W) := \frac{1}{t}\log(\frac{1}{n}\sum_{i=1}^n e^{t\ell(W^T\boldsymbol{\psi}(x_i), y_i)}) \quad \text{and} \quad \Omega_{\mathbf{z}}(W) = \sum_{j=1}^p \tau_j \|w_j\|_2.$$

The loss function $\ell(\cdot,\cdot)$ can be selected as least squared loss for regression task and logistic loss for classification task. Simple computation shows that

$$\nabla_W \mathcal{R}_{\boldsymbol{z},\psi}(t, W) = \frac{1}{n}\sum_{i=1}^n \frac{e^{t\ell\left(W^T\boldsymbol{\psi}(x_i), y_i\right)}}{\sum_{j=1}^n e^{t\ell(W^T\boldsymbol{\psi}(x_j), y_j)}} \nabla_W \ell(W^T\boldsymbol{\psi}(x_i), y_i),$$

where $\nabla_W \ell(W^T\boldsymbol{\psi}(x_i), y_i) = (W^T\boldsymbol{\psi}(x_i) - y_i)\boldsymbol{\psi}(x_i)$ for least squared loss and $\nabla_W \ell(W^T\boldsymbol{\psi}(x_i), y_i) = \frac{e^{W^T\boldsymbol{\psi}(x_i)}}{1+e^{W^T\boldsymbol{\psi}(x_i)}}\boldsymbol{\psi}(x_i) - y_i\boldsymbol{\psi}(x_i)$ for logistic loss.

Let M be the number of inner loop, S be the number of outer loop, and η_0 be the step size. We suppose that I_k with $|I_k| = b$ is a set randomly picked from $\{1, \ldots, n\}$. Denote $\nabla\mathcal{R}_{\boldsymbol{z},\psi}^{I_k}$ as the gradient related to I_k samples. To overcome the non-smoothness property of the regularizer $\Omega_{\mathbf{z}}$, we introduce the following the proximal operator as (Kowalski, 2009; Boyd & Vandenberghe, 2004)

$$\text{prox}_{\eta_0, \lambda\Omega_{\boldsymbol{z}}}(W) := (S_{\lambda\eta_0}(w_1)^T, S_{\lambda\eta_0}(w_2)^T, \ldots, S_{\lambda\eta_0}(w_p)^T)^T,$$

where

$$S_{\lambda\eta_0}(w_j) = \begin{cases} 0, & \text{if } \|w_j\|_2 \leq \lambda\eta_0 \\ \frac{\|w_j\|_2 - \lambda\eta_0}{\|w_j\|_2} w_j, & \text{otherwise.} \end{cases} \quad \text{for } j = 1, \ldots, p.$$

Following the non-convex optimization algorithm ProxSVRG in(J. Reddi et al., 2016), the detailed steps for solving (13) are summarized in Algorithm 1.

Algorithm 1 ProxSVRG for (13)

```
Input: Observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$, $\widetilde{W}^0 = W_M^0 = W^0 \in \mathbb{R}^{pd}$, $t \in (-\infty, 0) \cup (0, +\infty)$
for $s = 0; s < S; s = s + 1$ do
    $W_0^{s+1} = W_M^s$
    $g^{s+1} = \nabla_W \mathcal{R}_{\boldsymbol{z},\psi}(t, \widetilde{W}^s)$
    for $k = 0; k < M; k = k + 1$ do
        Uniformly randomly pick $I_k \subset \{1, \ldots, n\}$ such that $|I_k| = b$
        $v_k^{s+1} = \nabla_W \mathcal{R}_{\boldsymbol{z},\psi}^{I_k}(t, W_k^{s+1}) - \nabla_W \mathcal{R}_{\boldsymbol{z},\psi}^{I_k}(\widetilde{W}^s) + g^{s+1}$
        $W_{k+1}^{s+1} = \text{prox}_{\eta_0, \lambda\Omega_{\mathbf{z}}}(W_k^{s+1} - \eta_0 v_k^{s+1})$
    end
    $\widetilde{W}^{s+1} = W_M^{s+1}$
end
Output: $\hat{W} = W_M^S$
```

Moreover, the optimization for multi-objective learning can also be achieved through replacing the gradient $\nabla_W \mathcal{R}_{\boldsymbol{z},\psi}(t, W)$ in Algorithm 1 with

$$\nabla_W \mathcal{R}_{\boldsymbol{z},\psi}(t, \gamma, W) = \sum_{g \in G} \sum_{x \in g} \frac{1}{|g|} \frac{\left(\frac{1}{|g|} \sum_{z \in g} e^{\gamma \ell(W^T \boldsymbol{\psi}(x), y)}\right)^{\left(\frac{t}{\gamma} - 1\right)}}{\sum_{g' \in G} \left(\frac{1}{|g'|} \sum_{z \in g'} e^{\gamma \ell(W^T \boldsymbol{\psi}(x), y)}\right)^{\frac{t}{\gamma}}} e^{\gamma \ell(W^T \boldsymbol{\psi}(x), y)} \nabla_W \ell(W^T \boldsymbol{\psi}(x), y).$$