Quantile Additive Trend Filtering

Zhi Zhang **Kyle Ritscher** **Oscar Hernan Madrid Padilla**
Department of Statistics and Data Science
University of California, Los Angeles

Abstract

This paper introduces and analyzes quantile additive trend filtering, a novel approach to model the conditional quantiles of the response variable given multivariate covariates. Under the assumption that the true model is additive, and that the components are functions whose rth order weak derivatives have bounded total variation, our estimator is a constrained version of quantile trend filtering within additive models. The primary theoretical contributions are the error rate of our estimator in both fixed and growing input dimensions. In the fixed dimension case, we show that our estimator attains a rate that mirrors the non-quantile minimax rate for additive trend filtering, featuring the main term $n^{-2r/(2r+1)}$. For growing input dimension (d), our rate has an additional polynomial factor $d^{(2r+2)/(2r+1)}$. We propose a practical algorithm for implementing quantile additive trend filtering using dimension-wise backfitting. Experiments in both real data and simulations confirm our theoretical findings. We provide a public implementation of the algorithm at https://github.com/zzh237/QATF.

1 INTRODUCTION

Quantile regression has been widely used for analyzing various types of data, especially in the presence of heteroscedasticity or non-normal error distributions. It has been considered robust, with broad applications ranging from economics and finance to environmental science and healthcare (Coad et al., 2006; Sanderson et al., 2006; Perlich et al., 2007; Benoit and Van den Poel, 2009; Wasko and Sharma, 2014; Pata and Schindler, 2015; Belloni et al., 2023).

Additive models and trend filtering have emerged as powerful tools for handling high-dimensional and nonparametric estimation problems. Additive models, introduced by Friedman and Stuetzle (1981), reduce dimensionality while providing interpretable predictions, and have been utilized in many fields, including finance (Hastie and Tibshirani, 1987), marketing (Breiman and Friedman, 1985), healthcare (Hastie, 2017), environmental pollution assessment (Hastie, 2017), policy analysis (Hastie, 2017), energy demand forecastingFasiolo et al. (2021), social science (Fasiolo et al., 2021), and politics (Petersen et al., 2016).

Trend filtering, proposed by Mammen and Van De Geer (1997); Kim et al. (2009), is a nonparametric method which fits a piecewise polynomial function. It has garnered significant interest in nonparametric research (Rudin et al., 1992; Tibshirani, 2014; Kim et al., 2009; Sadhanala and Tibshirani, 2019), and is particularly useful in scenarios with in-homogeneous smoothness or sharp changes in the underlying data, such as image processing (Rudin et al., 1992), final market crash (Gu and Mulvey, 2021), shifts in the dynamics of macroeconomics (Mulvey and Liu, 2016), and understanding of earthquakes (Yano and Kano, 2022).

Motivated by data sets with heavy-tails and outliers, we build our model based on quantile regression. To accommodate the prevalent application of additive models, we assume the underlying function is additive, and we use trend filtering as a regularization technique. This leads us to a more robust, flexible and interpretable framework, known as quantile additive trend filtering, which combines the benefits of quantile regression, additive models, and trend filtering.

1.1 Summary of Contributions

The proposed model in this paper is quantile additive trend filtering (QATF), which offers a dual benefit of dimension reduction and robust regression, and has not yet been studied. We conduct an examination of risk bounds for QATF and develop an algorithm for

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

learning the QATF model. The techniques we developed could also be used for analyzing the other additive models with different classes of functions. The most related work to our work is the paper Madrid Padilla and Chatterjee (2022), which focuses on the univarite quantile trend filtering, but does not consider additive models. We also provide non-asymptotic analysis, in contrast to Narayan et al. (2023), which only addresses asymptotic behavior of the estimator. Moreover, the papers Gasthaus et al. (2019) and Tagasovska and Lopez-Paz (2019) primarily focus on methodology and experimentation.

QATF is built around the univariate trend filtering estimator, defined by constraining according to the sum of ℓ_1 norms of discrete derivatives of the component functions, with a quantile loss applied. When the underlying quantile function is additive, with components whose $(r-1)$th derivatives have total variation bounded by V, we derive error rates for $(r-1)$th order quantile additive trend filtering. This rate is $n^{-2r/(2r+1)}V^{2/(2r+1)}\max\{1, V^{(2r-1)/(2r+1)}\}$ for a fixed input dimension d (under weak assumptions). The main term $n^{-2r/(2r+1)}V^{2/(2r+1)}$, is the same as the non-quantile minimax rate in Sadhanala and Tibshirani (2019). It also includes an extra term, which is of order $\mathcal{O}(1)$ if $V = \mathcal{O}(1)$. Such a value of V is referred to as canonical scaling, as described in Madrid Padilla and Chatterjee (2022); Sadhanala et al. (2016).

Additionally, we explore error rates for QATF where the input dimension (d) increases with the sample size (n). We show the proposed estimator exhibits an error rate that includes a polynomial factor of $d^{(2r+2)/(2r+1)}$ multiplied by the fixed dimension error rate. When r is large, the exponent $d^{(2r+2)/(2r+1)}$ approaches 1, recovering the linear dependence on d in the standard additive trend filtering model.

Our rates are minimax up to logarithmic factors in fixed dimension cases, and off by a small factor in the growing dimension setting. This aligns closely with the minimax rates for mean estimation in additive trend filtering, as discussed in Sadhanala and Tibshirani (2019). However, Sadhanala and Tibshirani (2019) primarily considers sub-Gaussian errors, where we only introducing minor assumptions, proving that the constrained quantile trend filtering estimator is robust to heavy-tailed (with no moment conditions required) distributions of the errors. In addition, the proof techniques we employ differ substantially from those in the cited work, see our discussion in Section B of the Appendix.

Finally, we develop an algorithm for performing quantile trend filtering in additive models, which involves a backfitting approach for each dimension. We validate the algorithm in extensive simulation settings consisting of additive models in the presence of noise, and then demonstrate the utility of our method by constructing prediction intervals using the publicly available 2024 World Happiness dataset.

The review of additive models, trend filtering, and quantile regression is given in Section M of Appendixs.

2 QUANTILE TREND FILTERING FOR ADDITIVE MODELS

2.1 Notations

Before we define the estimator, let us introduce the necessary definitions and notation. Given a distribution Q, and a set of i.i.d. points $X = \{X^1, \ldots, X^n\} \subseteq [0,1]^d$, $i = 1, \ldots, n$, from Q, we define the L_2, euclidean and empirical inner products $\langle \cdot, \cdot \rangle_{L_2}$ $\langle \cdot, \cdot \rangle_2$ $\langle \cdot, \cdot \rangle_n$ as below.

Let $\mathcal{X}$ be $[0,1]^d$, for two functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$, we define

$$\langle f, g \rangle_{L_2} = \int_{\mathcal{X}} f(x)g(x)dQ(x), \langle f, g \rangle_2 = \sum_{i=1}^{n} f(X^i)g(X^i),$$

And $\langle f, g \rangle_n = \frac{1}{n}\langle f, g \rangle_2$. Hence, we define the L_2 norm $\|\cdot\|_{L_2}$ as $\|f\|_{L_2}^2 = \langle f, f \rangle_{L_2} = \int_{\mathcal{X}} f(x)^2 dQ(x)$. We have empirical norm $\|\cdot\|_n$ as $\|f\|_n^2 = \langle f, f \rangle_n = \frac{1}{n}\sum_{i=1}^n f(X^i)^2$. The $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are defined as $\|f\|_2^2 = n\|\cdot\|_n^2$, and $\|f\|_\infty = \max_{i=1,\ldots n} |f(X^i)|$. We also use the $\|\cdot\|$ as the common euclidean norm, and we use $\|\cdot\|_\infty$ as the vector ℓ_∞ norm. For sequences a_n and b_n, we will also use the notation $a_n \lesssim b_n$ to denote that $a_n \leq Cb_n$ for an absolute constant C. We write $a_n = O(b_n)$ if there exists constants $C > 0$ and $n_0 > 0$ such that $n \geq n_0$ implies that $a_n \leq Cb_n$. Furthermore, if $a_n = O(b_n)$ and $b_n = O(a_n)$ then we write $a_n = \Theta(b_n)$ or $a_n \asymp b_n$. For a sequence of random variables $x^{(n)}$ and a positive sequence a_n we write $x^{(n)} = O_{\text{pr}}(a_n)$ if for every $\epsilon > 0$ there exists $M > 0$ such that $\mathbb{P}\left(|x^{(n)}| \geq Ma_n\right) < \epsilon$ for all n. Let $\mathbb{I}\{\mathcal{A}\}$ be the indicator function over the set $\mathcal{A}$ such that $\mathbb{I}\{\mathcal{A}\} = \begin{cases} 1, & \text{if } a \in \mathcal{A}, \\ 0, & \text{if } a \notin \mathcal{A}. \end{cases}$ For a functional ν, we let $B_\nu(\epsilon)$ for the ν-ball of radius $V > 0$, i.e., $B_\nu(V) = \{f : \nu(f) \leq V\}$. So $B_n(V)$ for the $\|\cdot\|_n$-ball of radius V, and $B_{L_2}(V)$ for the $\|\cdot\|_{L_2}$-ball of radius V, and $B_\infty(V)$ for the $\|\cdot\|_\infty$-ball of radius V.

For an additive function $f = \sum_{j=1}^d f_j$, its input is a vector $X^i = (X_1^i, \ldots, X_d^i) \in \mathbb{R}^d$, and we have $f(X^i) = \sum_{j=1}^d f_j(X_j^i)$. In this paper, we reserve index i for data point, and j for the component, and for any function f, we use f^i as a shorthand for $f(X^i)$, and f_j^i for $f_j(X_j^i)$.

Next, we introduce our specific model and data framework.

2.2 Model and Data Description

Suppose we are given data $\{(X^i, Y^i)\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$, where $\mathcal{X} = [0,1]^d$, generated as

$$Y^i = f_0(X^i) + \epsilon^i = \sum_{j=1}^d f_{0j}(X_j^i) + \epsilon^i, \tag{1}$$

where $\mathbb{P}(\epsilon^i \leq 0 \mid X^i) = \tau$, with

$$f_0 : [0,1]^d \to \mathbb{R}, \quad f_{0j} : [0,1] \to \mathbb{R},$$

and where $X^1, X^2, \ldots, X^n$ are i.i.d. draws from a distribution $\mathscr{F}$ in $[0,1]^d$.

Thus, $f_0(X^i) = \sum_{j=1}^d f_{0j}(X_j^i)$ is the τ-conditional quantile of Y^i given $X^i = (X_1^i, \ldots, X_d^i)$ of d dimensions, i.e. $f_0(X^i) = F^{-1}_{Y^i|X^i}(\tau)$, for $i = 1, \ldots, n$, where $F_{Y^i|X^i}(\cdot)$ denotes the conditional cumulative distribution function of Y^i given X^i.

An alternative way to write our model is to let

$$\theta_0^i \coloneqq f_0(X^i) \quad \text{and} \quad \theta_{0j}^i = f_{0j}(X_j^i),$$

so that Y^i can be written as

$$Y^i = \theta_0^i + \epsilon^i = \sum_{j=1}^d \theta_{0j}^i + \epsilon^i. \tag{2}$$

We find (2) to be helpful when both describing the methods and theory for our proposed model.

Furthermore, we collect the jth component of the inputs into a vector of n dimensions, i.e., $X_j = (X_j^1, X_j^2, \ldots, X_j^n)$ for $j = 1, \ldots, d$.

Our goal is to estimate f_0 or θ_0. In this paper we are interested in scenarios where f_0 is (or is close to) a piecewise polynomial function.

2.3 Discrete Difference Operator

The discrete difference operator is a fundamental component in defining the trend filtering estimator for a given integer r.

Let $r \in \mathbb{N}_{\geq 1}$ and denote $X = (X^1, \ldots, X^n) \in \mathbb{R}^n$ as a vector of univariate input points. Let $(X^{(1)}, \ldots, X^{(n)})$ be the ordered inputs where $X^{(1)} < \cdots < X^{(n)}$. The operator $D_n^{(X,r)} \in \mathbb{R}^{(n-r)\times n}$ is defined on the ordered inputs recursively.

First-Order Difference Operator ($r = 1$). The first-order difference operator $D_n^{(X,1)}$ is defined as:

$$D_n^{(X,1)} = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n},$$

$$D_n^{(X,1)}(\theta) = (\theta^2 - \theta^1, \ldots, \theta^n - \theta^{n-1}). \tag{3}$$

Higher-Order Difference Operators ($r \geq 2$). For higher-order difference operators, we recursively define $D_n^{(X,r)} \in \mathbb{R}^{(n-r)\times n}$ using the previous order's operator,

$$D_n^{(X,r)} = D_{n-r+1}^{(X,1)} \cdot \operatorname{diag}\left(\frac{r-1}{X^{(r)} - X^{(1)}}, \ldots, \frac{r-1}{X^{(n)} - X^{(n-r+1)}}\right) D_n^{(X,r-1)}, \tag{4}$$

where $D_{n-r+1}^{(X,1)} \in \mathbb{R}^{(n-r)\times(n-r+1)}$ is first-order operator as in Equation (3). Examples of the discrete difference operator are given in Section N.2.1 of the Appendix.

In summary, the matrix $D_n^{(X,r)}$ generalizes the discrete difference operator to account for non-uniformly spaced inputs, making it a powerful tool in trend filtering applications.

2.4 Univariate Quantile Trend Filtering

We first introduce the univariate quantile trend filtering, which focuses on signals (quantile sequences) that have bounded rth order total variation.

We first introduce a function used in quantile regression to define the loss. The check function, $\rho_\tau(u)$, for a given quantile level $\tau \in (0,1)$, is defined as follows:

$$\rho_\tau(u) = u\left(\tau - \mathbb{I}\{u < 0\}\right) = \max\left\{\tau u, (\tau - 1)u\right\}.$$

In the analysis presented in Madrid Padilla and Chatterjee (2022), the constrained univariate quantile trend filtering estimator $\widehat{\theta}^{(r)}$ is given by

$$\begin{aligned} &\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^n \rho_\tau(Y^i - \theta^i) \\ &\text{subject to} \quad \|D^{(r)}(\theta)\|_1 \leq V. \end{aligned} \tag{5}$$

Remark 2.1. a. The $D^{(r)}$ is the rth order difference operator, for a vector $\theta \in \mathbb{R}^n$, $D^{(0)}(\theta) = \theta$, $D^{(1)}(\theta) = (\theta^2 - \theta^1, \ldots, \theta_n - \theta_{n-1})$, and $D^{(r)}(\theta)$, for $r \geq 2$, as recursively defined as $D^{(r)}(\theta) = D^{(1)}(D^{(r-1)}(\theta))$. Note that $D^{(r)}(\theta) \in \mathbb{R}^{n-r}$.
b. $D^{(r)}$ in the univariate quantile trend filtering estimator is equal to $D_n^{(X,r)}$ in (4) when x equals the grid $(1/n, 2/n, \ldots, n/n)$.

Due to ℓ_1 penalization constrained univariate quantile trend filtering estimator enforce $D^{(r)}(\widehat{\theta}^{(r)})$ to be sparse. It is known that for $\theta \in \mathbb{R}^n$, $D^{(r)}(\theta)$ has k nonzero entries if and only if $\theta = (f(1/n), f(2/n), \ldots, f(n/n))^\top$ for a *discrete spline function* f, consisting of $(k+1)$ polynomials of degree $r-1$ (see Proposition D.3 in Guntuboyina et al. (2020)). For this reason, just like the usual trend filtering estimators, the univariate quantile trend filtering estimator fits discrete splines. For the precise definition of a discrete spline see Section 2 in Mangasarian and Schumaker (1971).

2.5 Quantile Additive Trend Filtering via Discrete Difference Operator

Given the definition of the discrete difference operator, the quantile additive trend filtering (QATF) estimator regularizes each component function based on the total variation of its rth order discrete derivative. To define the estimator, we introduce the constrained set $K^{(X,r)}(V)$, and the estimator $\widehat{\theta}^{(r)}$ is derived from the following optimization problem:

$$\min_{\theta \in K^{(X,r)}(V)} \sum_{i=1}^{n} \rho_\tau \left(Y^i - \sum_{j=1}^{d} \theta_j^i \right), \tag{6}$$

where the feasible set is defined as:

$$K^{(X,r)}(V) := \left\{ \theta : \theta = \sum_{j=1}^{d} \theta_j, \sum_{j=1}^{d} \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1 \leq V, \right. \\ \left. \mathbb{1}^\top \theta_j = 0,\, j = 1, \ldots, d \right\}. \tag{7}$$

The $S_j \in \mathbb{R}^{n \times n}$ is a permutation matrix such that

$$S_j \theta_j = \left(\theta_j^{(1)}, \theta_j^{(2)}, \ldots, \theta_j^{(n)} \right), \quad j = 1, \ldots, d,$$

where $\theta_j^{(1)} < \cdots < \theta_j^{(n)}$ among $\theta_j = (\theta_j^1, \theta_j^2, \ldots, \theta_j^n)$. The $\mathbb{1}$ is a vector with all entries equal to 1, ensuring that $\mathbb{1}^\top \theta_j = 0$ for identifiability. The positive constant V controls the additive ℓ_1 penalization. As shown by Tibshirani (2014) and Sadhanala and Tibshirani (2019), this regularization results in piecewise polynomial components of degree $(r-1)$.

Compared to the constrained additive trend filtering formulation in Sadhanala and Tibshirani (2019), where Sub-Gaussian errors are assumed, the formulation in Equation (6) is designed to handle models with heavy-tailed distributions, such as Cauchy and t-distributed errors, without assuming the existence of moments. This makes the problem formulation and analysis more challenging.

2.6 Falling Factorial Representation

The falling factorial representation offers an alternative method to express QATF estimates. As seen in Tibshirani (2014); Wang et al. (2015), this representation is useful for deriving rapid error rates in trend filtering.

Definition 2.2 (Definition 2.1 in Sadhanala and Tibshirani (2019))**.** Given the knot points $(t^1 < \cdots < t^n)$ where $t^1 < \cdots < t^n \in [0,1]$, we define the falling factorial basis as follows:

$$h_i^{(t)}(t) = \prod_{l=1}^{i-1} (t - t^l), \quad i = 1, \ldots, r+1, \\ h_{i+r+1}^{(t)}(t) = \prod_{l=1}^{r} (t - t^{i+l}) \cdot \mathbb{I}\{t > t^{i+r}\}, \\ i = 1, \ldots, n-r-1. \tag{8}$$

Then we define a linear subspace of function as

$$\mathcal{H}_j = \left\{ \sum_{i=1}^{n} \alpha_j^i h_i^{(X_j)} : \alpha_j^1, \ldots, \alpha_j^n \in \mathbb{R} \right\},$$

where $\alpha_j^1, \ldots, \alpha_j^n$ are coefficients for the jth dimension.

The space $\mathcal{H}_j$ is defined as the span of the $(r-1)$th order falling factorial basis $\{h_1^{(X_j)}, \ldots, h_n^{(X_j)}\}$ over the jth dimension. These functions represent piecewise polynomial functions of order $r-1$.

If $f_j \in \mathcal{H}_j$, then $f_j = \alpha_j^1 h_1^{(X_j)} + \alpha^2 h_2^{(X_j)} + \cdots + \alpha_j^n h_n^{(X_j)}$. This results in a continuous piecewise polynomial function, with a global polynomial structure determined by $\alpha_j^1, \ldots, \alpha_j^{r+1}$. Notably, when $\alpha_j^{i+r+1} \neq 0$, f_j exhibits changes in its rth derivative at the knot t^{i+r}, as well as in all lower-order derivatives.

Based on the Definition 2.2, we will give an equivalent formulation of problem in Equation (6). Before we do that, we introduce an important definition.

Definition 2.3 (Total Variation)**.** For a function $f : [0,1] \to \mathbb{R}$, its total variation is defined as:

$$TV(f) = \sup \left\{ \sum_{i=1}^{p} |f(z_{i+1}) - f(z_i)| : \right. \\ \left. z_1 < \cdots < z_p, \text{ is a partition of } [0,1] \right\}. \tag{9}$$

For simplicity, we use $TV^{(r)}$ to denote the total variation of the $(r-1)$th weak derivative of f, i.e., $TV^{(r)}(f) = TV(f^{(r-1)})$.

Total Variation is discussed further in Appendix N.1.

Based on the relationship $\theta_j^i = f_j(X_j^i)$ and the arguments above, the following problem is equivalent to the problem in Equation (6). Define the function class $\mathcal{F}^{(r)}(V)$ and the estimator $\widehat{f}_{\mathcal{H}}^{(r)}$ as follows:

$$\mathcal{F}^{(r)}(V) := \left\{ f = \sum_{j=1}^{d} f_j : f_j \in \mathcal{H}_j, \sum_{j=1}^{d} TV(f_j^{(r-1)}) \leq V, \right. \\ \left. \sum_{i=1}^{n} f_j(X_j^i) = 0 \right\}, \tag{10}$$

$$\min_{f\in\mathcal{F}^{(r)}(V)} \sum_{i=1}^{n} \rho_\tau \left(Y^i - \sum_{j=1}^{d} f_j(X_j^i) \right). \tag{11}$$

Here, $\mathcal{H}_j$ is the span of the falling factorial basis over the jth dimension, defined in Equation (8), and $\sum_{i=1}^{n} f_j(X_j^i) = 0$ is required for identifiability.

Penalized Version of the Estimator. Using general Lagrange duality theory, the problem in Equation (6) can be reformulated as a penalized version by introducing a positive tuning parameter λ:

$$\begin{aligned} &\min_{\theta_1,\ldots,\theta_d\in\mathbb{R}^n} \sum_{i=1}^{n} \rho_\tau(Y^i - \sum_{j=1}^{d} \theta_j^i) + \lambda \sum_{j=1}^{d} \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1 \\ &\text{subject to} \quad \mathbb{1}^\top \theta_j = 0, \quad j = 1,\ldots,d. \end{aligned} \tag{12}$$

For appropriate values of V and λ, Equations (6) and (12) are equivalent. For practical implementation, (12) will be employed in our experiments.

3 MAIN RESULTS

In this section, we present the main results of the paper concerning the theoretical guarantees of the QATF estimator. In our analysis, we adopt key assumptions that guide the theoretical foundation of the work. These are only describe them here at a high level, but are provided and explained in full mathematical detail in Section D of the Appendix.

First, we assume the input data lies within $[0,1]^d$, with the maximum gap between consecutive values controlled by $O(\frac{\log n}{n})$, ensuring a well-behaved distribution for analysis. Remark D.2 explains that this assumption is not restrictive.

Next, we assume the additive functions are centered, meaning their sum across input dimensions is zero, ensuring model identifiability. This follows standard practice in additive models.

Finally, we introduce an assumption for the conditional distribution function of the outcome variable, which guarantees reliable quantile estimates and is commonly applied in related statistical literature.

3.1 Results for QATF Estimator

To begin, we introduce notation for a loss function, Δ_n^2. This loss function is not used for training; rather, it is used to quantify the quality of the estimators. The mathematical formulation of this loss function is as follows: let $X^1,\ldots,X^n$ be given data inputs, and consider a function $\delta : [0,1]^d \to \mathbb{R}$. Define Δ_n^2 as a mapping from functions to real numbers

$$\Delta_n^2(\delta) = \frac{1}{n} \sum_{i=1}^{n} \min\left\{ \left|\delta(X^i)\right|, \delta(X^i)^2 \right\}, \tag{13}$$

which, up to constants, is the Huber loss (Huber, 1992). We then provide the risk bounds for estimators in Equation (6) and (11). However, the primary challenges encountered in our study arise from the use of quantile loss, as opposed to least squares loss, and the incorporation of an additive model with potentially growing dimension. Both parts present considerable difficulties in proving risk bounds. Thus, we give the bounds based on Huber-type loss in (13), as this loss naturally appears as a lower bound to the quantile population loss. See (35) in the Appendix for a more detailed explanation. For these reasons, the Huber loss is a natural loss function in quantile regression (Belloni et al., 2023; Madrid Padilla and Chatterjee, 2022; Ye and Padilla, 2021). From a statistical perspective, the Huber loss provides insights into the convergence rate of the estimator toward the true value, as it captures some notion of the "average" error, albeit in a slightly weaker sense than the usual least squares loss.

Theorem 3.1. Let $\{Y^i\}_{i=1}^n$ be any sequence of independent random variables which satisfies Assumption D.4 and let f_0 be as defined in (1), where $f_0(X^i)$ represents the τ-conditional quantile of Y^i given X^i. Suppose Assumption D.1 holds on the data inputs, and Assumption D.3 holds on f_0. Assume that the dimension d of the input space is fixed, and that the underlying regression function is additive, $f_0 = \sum_{j=1}^{d} f_{0j}$. Here, $f_0 \in \mathcal{F}^{(r)}(V)$ and the components f_{0j}, $j = 1,\ldots,d$, are r times weakly differentiable, with $\sum_{j=1}^{d} TV(f_{0j}^{(r-1)}) = V^*$. If V is chosen such that $V \geq V^* = \sum_{j=1}^{d} TV(f_{0j}^{(r-1)})$, then

$$\begin{aligned} &\Delta_n^2(\widehat{f}_{\mathcal{H}}^{(r)} - f_0) = \\ &\quad O_{pr}\Big(n^{-\frac{2r}{2r+1}} V^{\frac{2}{2r+1}} \max\Big\{1, V^{\frac{2r-1}{2r+1}}\Big\}\Big). \end{aligned} \tag{14}$$

Remark 3.2. Theorem 3.1 extends Theorem 1 in Sadhanala and Tibshirani (2019) to quantile regression, differing from their Theorem 1 by the inclusion of an additional factor: $\max\{1, V^{(2r-1)/(2r+1)}\}$, which can go to infinity if V grows to infinity. However, under the canonical scaling $V^* = \mathcal{O}(1)$ as mentioned in Madrid Padilla and Chatterjee (2022); Sadhanala et al. (2016), one can choose $V = \mathcal{O}(1)$ as well, thus the above term is also $\mathcal{O}(1)$. Theorem 3 in Sadhanala and Tibshirani (2019) presents a lower bound, detailed in Equation (39). This lower bound aligns with our upper bound rate stated in Equation (15) of our main manuscript, which holds for fixed dimension d, differing only by logarithmic factors. Hence, in the fixed dimension d, our bound is minimax optimal. Furthermore, the lower bound of their paper is for the Sub-Gaussian case, our upper bound is for arbitrary error terms that do not require moment conditions.

Theorem 3.1 assumes that τ is fixed. For the cases where $\tau \to 0$ or $\tau \to 1$, see the discussion in Section G of the Appendix.

3.2 Error Bounds for A Growing Dimension d

In this subsection, we allow the input dimension d to grow with the sample size n. Furthermore, to achieve an error rate that scales linearly with the dimension d, we assume the input points are i.i.d. from a continuous distribution on $[0,1]^d$ that decomposes into independent marginals. For more details, refer to Section H of the Appendixs.

We now state our main result in the growing d case, whose proof can be found in Appendix I, J.

Theorem 3.3. Let $\{Y^i\}_{i=1}^n$ be any sequence of independent random variables which satisfies Assumption D.4 and let f_0 be as defined in (1), where $f_0(X^i)$ represents the τ-conditional quantile of Y^i given X^i. Let $\{X^i\}_{i=1}^n$ be the input points which satisfy Assumption D.1 and H.1. Also, suppose that f_0 satisfies Assumption D.3. Assume the underlying regression function is additive, $f_0 = \sum_{j=1}^d f_{0j}$, where $f_0 \in \mathcal{F}^{(r)}(V)$. The components f_{0j}, $j = 1,\dots,d$, are k times weakly differentiable, and $\sum_{j=1}^d TV(f_{0j}^{(r-1)}) = V^*$. If V is chosen such that $V \geq V^* = \sum_{j=1}^d TV(f_{0j}^{(r-1)})$, then

$$\Delta_n^2(\widehat{f}_{\mathcal{H}}^{(r)} - f_0) = O_{pr}\Big(d^{\frac{2r+2}{2r+1}} n^{-\frac{2r}{2r+1}} V^{\frac{2}{2r+1}} \cdot \max\Big\{1, V^{\frac{2r-1}{2r+1}}\Big\}\Big). \tag{15}$$

Remark 3.4. Theorem 3.3 generalizes Theorem 2 in Sadhanala and Tibshirani (2019) to the quantile regression setting, with a difference that our upper bound contains an extra term. This is the factor $\max\{1, V^{(2r-1)/(2r+1)}\}$, which can go to infinity if V grows to infinity. However, under the canonical scaling $V^* = \mathcal{O}(1)$ one can choose $V = \mathcal{O}(1)$ as well and thus the above term is also $\mathcal{O}(1)$. The factor $d^{(2r+2)/(2r+1)}$ compared to Theorem 2 in Sadhanala and Tibshirani (2019) is d, and for large r, these two terms tend to be the same, making our rate essentially minimax up to log factors.

4 BACKFITTING ALGORITHM

In this section, we describe the algorithm for fitting our method. The algorithm is based on the idea of backfitting, which was introduced in Härdle and Hall (1993) for additive models. Two algorithms closely related to backfitting for additive models are the alternating least squares and alternating conditional expectations methods, as mentioned in Van Der Burg and de Leeuw (1983) and Breiman and Friedman (1985). Other works that have explored backfitting idea include Buja et al. (1989); Nielsen and Sperlich (2005); Ravikumar et al. (2009); Petersen et al. (2016); Tibshirani (2017).

To solve the original problem (12), we use the backfitting approach described in Algorithm 1. Specifically, the algorithm cycles over $j = 1,\dots,d$, and at each step updates the estimate for component j by applying univariate quantile trend filtering to the jth partial residual (i.e., the current residual excluding component j). The univariate quantile trend filtering utilizes the functions described in Brantley et al. (2020).

Algorithm 1 Backfitting for quantile additive trend filtering

1: **Inputs:** Responses $Y^i \in \mathbb{R}$ and input points $X^i \in \mathbb{R}^d$, $i = 1,\dots,n$, and set a value for $\lambda > 0$.
2: Set $t = 0$ and initialize $\theta_j^{(0)} = 0$, for $j = 1,\dots,d$.
3: **for** $t = 1, 2, 3, \dots$ (until convergence) **do**
4: **for** $j = 1, 2, \cdots, d$ **do**
5: (i) Set response u_j^i, $u_j^i = Y^i - \sum_{l<j} \theta_l^{i(t)} - \sum_{l>j} \theta_l^{i(t-1)}$, for $i = 1,\dots,n$.
6: (ii) $\theta_j^{(t)} = \arg\min_{\theta_j} \sum_{i=1}^n \rho_\tau(u_j^i - \theta_j^i) + \lambda \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1$, subject to $\mathbb{1}^\top \theta_j = 0$.
7: **end for**
8: **end for**
9: Return $\widehat{\theta}_j$, $j = 1,\dots,d$ (parameters $\theta_j^{(t)}$ at convergence).

4.1 Backfitting with ADMM

The main computational burden of Algorithm 1 is Line 6. We tackle this by using the alternating directions method of multipliers (ADMM) algorihtm, see Boyd et al. (2011), as follows. We let $u_j^i = Y^i - \sum_{l<j} \theta_l^i - \sum_{l>j} \theta_l^i$, then, we solve

$$\min_{\theta_j \in \mathbb{R}^n} \sum_{i=1}^n \rho_\tau(u_j^i - \theta_j^i) + \lambda \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1, \quad \text{subject to} \quad \mathbb{1}^\top \theta_j = 0. \tag{16}$$

How to solve (16) is given in Section K of Appendix.

Algorithm 1 is equivalent to block coordinate descent (BCD), also called exact blockwise minimization, applied to Problem (12) over the coordinate blocks θ_j, $j = 1,\dots,d$. To summarize the broader theoretical framework, a general treatment of the Block Coordinate Descent (BCD) method is given by Tseng (2001). Since Equation (12) decomposes into smooth plus separable terms, which satisfies a convex criterion of BCD,

see Appendix Theorem L.1 for details. Then it immediately holds that the iterates of our algorithm converges to the minimal point of the objection function. Hence, our algorithm for quantile additive trend filtering is guaranteed to find the minimizer.

4.2 Computational Complexity

In Algorithm 1, line 3, the outer loop usually takes 20 iterations to converge. In each iteration, the backfitting process is conducted, involving d distinct fits of quantile trend filtering. For each fit of quantile trend filtering, the ADMM procedure in Section K of the Appendix requires three updates (primal, dual, and slack). In the worst case, each update can be computed in $O(n)$ time (treating r as a constant), leading to an overall ADMM complexity of $O(nm)$, where m denotes the maximum number of ADMM iterations. There are d components, therefore one full iteration of standard backfitting ADMM updates can be done in linear time $O(dnm)$. In practice, we find m tends to be small, typically around 10. Thus, the overall complexity of Algorithm 1 amounts to $O(dn)$. The details about the calculation of $O(dn)$ can be found in Section K of Appendix. The software and hardware setting for Algorithm 1 can be found in Appendix P.

4.3 Examples

We conclude this section with two examples, visualized in Figure 1 and Figure 2, illustrating the performance of our quantile additive trend filtering algorithm. Full details are given in Section O of the Appendix.

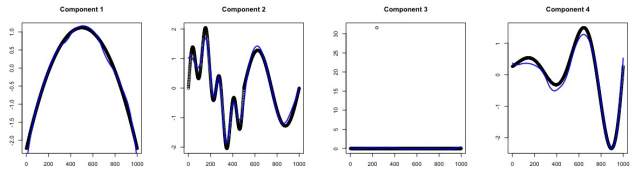
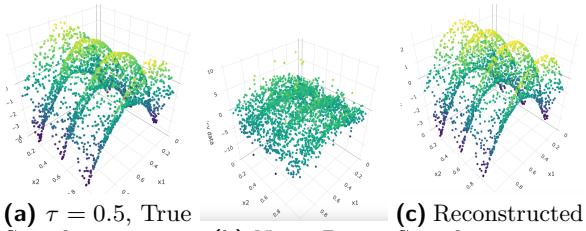


Figure 1: Component estimates of Quantile Additive Trend Filtering with $\tau = 0.5$ are plotted in blue, and the true component functions $f_{0j}(x) = a_j g_j(x) - b_j$ in black, with a_j and b_j chosen such that f_{0j} has an empirical mean of zero and empirical norm $\|f_{0j}\|_n = 1$. For this scenario, $g_1(x) = \frac{1}{2}x^2$, $g_2(x) = \frac{3}{2}\sin(4\pi x) + \mathbb{1}_{x \le \frac{1}{2}} \cdot \sin(16\pi x)$, g_3 is a dummy dimension (where only 1 randomly assigned point takes non-zero value), and $g_4(x) = e^{3x}\sin(4\pi x)$.

5 SIMULATED EXPERIMENTS

In this section, we conduct a comparison study to evaluate the performance of quantile additive trend filtering (QATF) of orders $r = 1$ and $r = 0$, denoted as QATF1 and QATF0 respectively, against existing benchmark methods.



(a) $\tau = 0.5$, True Signal **(b)** Noisy Data **(c)** Reconstructed Signal

Figure 2: Figure a plots true function $f_{0j}(x) = a_j g_j(x) - b_j$ on 0.5 quantile, where $g_1(x) = \frac{1}{2}\cos(6\pi x) + 0.1)$ and $g_2(x) = -(x - \frac{1}{2})^2$. Figure b plots the real data with heavy tailed noise added. Figure c plots QATF estimators $\widehat{f}$ on this Scenario, for a reconstruction of the true signal.

5.1 Benchmark Methods

We compared against the usual mean regression penalized additive trend filtering models of orders $r = 1$ and $r = 0$ (ATF1 and ATF0), as well as quantile smoothing splines (QS) (Koenker et al., 1994). Both ATF and QS can be implemented with minor modifications to step (ii) in Algorithm 1. Specifically, ATF is implemented by replacing the check function with the least squares loss, following the approach of Sadhanala and Tibshirani (2019), and QS simply substitutes the trend filtering penalty with the smoothing spline penalty. We implement the univariate fits using the R packages `trendfilter` (Tibshirani and Taylor, 2011) for ATF, and `fields` (Nychka et al., 2017) for QS.

5.2 Experimental Setup

We sample $n \in \{500, 1000, 2500\}$ points in $d = 10$ dimensions, assigning inputs $X_j = (X_j^1 \dots X_j^n)$, where X_j^i is sampled from a d-dimensional uniform distribution within the interval $[0, 1]^d$.

The component functions are defined as $f_{0j} = a_j g_j - b_j$ for $j = 1, \dots, d$. For $g_j(x)$, we use sinusoids with Doppler-like spatially varying frequencies:

$$g_j(x) = \sin\left(\frac{2\pi}{(x + 0.1)^{j/10}}\right), \quad j = 1, \dots, 10. \quad (17)$$

The true data is then generated as:

$$Y^i = \sum_{j=1}^{d} f_{0j}(X_j^i) + \epsilon^i, \quad i = 1, \dots, n.$$

This structure allows for significant heterogeneity both within and between the component functions. The errors $\{\epsilon^i\}_{i=1}^n$ are independent with $\epsilon^i \sim F^i$, varying across different scenarios.

For feature selection, we consider values of λ such that $\log_{10}(\lambda)$ is in a grid of 50 evenly spaced points in $[-2 + r, 4 + r]$ for QATF0 and QATF1, $[-7, 3]$ for

ATF0 and ATF1, and $[-16, 0]$ for QS. We then choose the λ to be the value that minimizes the estimator $\widehat{f}$'s mean squared error (MSE), $\frac{1}{n}\sum_{i=1}^{n}(\widehat{f} - f_0)^2$, with f_0 representing the true vector of quantiles. The grid size was chosen to experimentally encompass all variability in the MSE resulting from changes in λ, and we chose to use MSE for feature selection in order to ensure fair competition with other methods in the comparison experiments. Feature selection via Cross-Validation is used for the experiments on real data in Section 6.

Table 1: Average mean squared error, $\frac{1}{n}\sum_{i=1}^{n}(f_0^i - \widehat{f}^i)^2$, averaging over 50 Monte Carlo simulations for the different methods considered. The best MSE is listed with bold text. S = Scenario.

n	S	τ	QATF1	QATF0	QS	ATF1	ATF0
500	1	0.5	0.9831	1.0136	1.0738	**0.5800**	0.6194
1000	1	0.5	0.6132	0.7426	0.7304	**0.4011**	0.4583
2500	1	0.5	0.3111	0.4096	0.4325	**0.2055**	0.2751
500	2	0.5	**2.0341**	2.3507	2.2444	165.2800	2794.1800
1000	2	0.5	**1.3132**	1.5889	1.5223	1412.6500	78147.2000
2500	2	0.5	**0.6243**	0.7877	0.8078	33.5683	2623.1000
500	3	0.5	**1.1926**	1.5098	1.3299	4.1947	4.0913
1000	3	0.5	**0.7359**	0.9847	0.8901	3.9147	3.8723
2500	3	0.5	**0.3468**	0.4963	0.4965	3.6066	3.4550
500	4	0.5	**0.9751**	1.1981	1.0976	3.5949	5.4504
1000	4	0.5	**0.5600**	0.7544	0.6707	1.0105	1.1889
2500	4	0.5	**0.2391**	0.3647	0.3604	0.6629	0.9073
500	4	0.8	**1.7977**	2.4720	1.9304	NA	NA
1000	4	0.8	**1.0583**	1.4232	1.2408	NA	NA
2500	4	0.8	**0.5405**	0.7116	0.6917	NA	NA
500	4	0.2	**1.7558**	2.3276	1.7912	NA	NA
1000	4	0.2	**1.0611**	1.4834	1.2454	NA	NA
2500	4	0.2	**0.5364**	0.7019	0.6888	NA	NA
500	5	0.5	1.6479	1.6549	**1.6190**	1.8772	1.6511
1000	5	0.5	1.0925	**0.8612**	1.1024	1.2632	0.9551
2500	5	0.5	0.6382	**0.3466**	0.6963	0.7755	0.4624
500	6	0.5	0.6457	1.0902	0.9807	**0.6412**	0.6713
1000	6	0.5	**0.3373**	0.7372	0.6635	0.3586	0.4407
2500	6	0.5	0.2175	0.4265	0.4308	**0.2078**	0.2885
500	7	0.5	**1.9526**	2.0708	2.0007	3.5986	3.5436
1000	7	0.5	**1.3192**	1.5304	1.4419	2.0597	2.3514
2500	7	0.5	**0.7164**	0.8174	0.8542	2.9768	4.3687

5.3 Scenarios and Results

We examine the performance of QATF and benchmark methods across different scenarios:

Scenario 1: Normal Errors

In this scenario, the distributions F^i are taken as $N(0, 1)$. Given the normal errors and varying smoothness, this scenario is ideally suited for ATF. As shown in Table 1, ATF1 performs the best, with ATF0 closely following. QATF1 and QATF0 both outperform QS, likely due to splines' limitations in adapting to heterogeneous smoothness.

Scenario 2: Cauchy Errors

Here, the distributions F^i are taken as Cauchy$(0, 1)$, where the errors have no mean. As expected, ATF1 and ATF0 perform poorly in this scenario. Conversely, the quantile methods show robustness, with, consistent with the results from Scenario 1, QATF1 and QATF0 slightly outperforming QS.

Scenario 3: Log-Normal Errors

For this scenario, the distributions F^i are taken as log-normal$(0, 1)$, which is both right-skewed and heavy-tailed. In real data applications, this is another typical example of where estimating the median is more useful than estimating the mean. We see that the quantile methods perform well at this task, with QATF1 performing the best.

Scenario 4: Heteroscedastic t Errors

In this scenario, the distributions F^i are taken as $t(2)$, but with $\epsilon^i = i^{1/2}/n^{1/2}v^i$, where v^is are independent draws from F^i. The performance of the methods is similar to Scenario 2, but less extreme. We also utilize this scenario to demonstrate fits on other quantiles. This is only possible with a quantile fitting method, so ATF1 and ATF0 are represented by NAs in Table 2.

Scenario 5: Piecewise Constant Components, t Errors

For this scenario, we redefine the function $g_j(x)$ as:

$$g_j(x) = \begin{cases} 1 & \text{if } x \in [b_1, b_2) \\ -1 & \text{if } x \in [b_2, b_3) \\ \vdots & \vdots \\ (-1)^j & \text{if } x \in [b_{j+1}, b_{j+2}) \end{cases},$$

where $b_1 = 0$, $b_{j+2} = 1$, and $b_2 \dots b_{j+1}$ are squares of j evenly spaced breakpoints in $(0, 1)$. This maintains heterogeneity within and between components as in previous scenarios, but introduces a piecewise-constant structure. The distributions F^i are taken as $t(3)$. This is a scenario where we expect the k = 0 models to perform the best, which is indeed the case; QATF0 is the premier method, with ATF0 outperforming QATF1 and QS for second place.

Scenario 6: Time Series, Normal Errors

For this scenario, we take the function as in (17), but now we allow series correlation. We have an autoregressive model the error ϵ^i depends on the previous ϵ^{i-1}. Specifically, for $\tau = 0.5$, we consider.

$$\epsilon^i = \frac{0.5\epsilon^{i-1} + v^i}{0.5^2 + 1^2}, v^i \sim F^i, \text{ where } F^i \text{ as } N(0, 1).$$

With the normal errors, ATF1 performs the best, though by only a slim margin over QATF1.

Scenario 7: Time Series, t Errors

For this, we consider the same structure as Scenario 6, but with F^i as $t(2)$. As expected, in the transition to heavy tails, the quantile methods are more effective, with QATF1 demonstrating the strongest performance.

Overall, the QATF estimator performs well across all scenarios, and particularly outperforms ATF under heavy-tailed errors, thereby supporting our theoretical findings. We also provide simulation results for $\tau = 0.2$ and 0.8 for Scenarios 1, 2, 3, 5, and 6 in the Table 3 of Appendix. We only consider $\tau = 0.5$ for Scenario 7 due to the difficulties in constructing the ground truth data when $\tau \neq 0.5$

6 WORLD HAPPINESS DATA

In this section, we demonstrate the effectiveness of QATF in real data applications by performing statisti-

cal inference on the 2024 World Happiness Data. We analyze the conditional relationships between a country-level happiness index, derived from average responses to the Cantril life ladder question in the Gallup World Polls Cantril (1965); Helliwell et al. (2024), and nine predictors, such as gross national income, sourced from the 2023 World Happiness Report and World Bank Development Indicators World Bank Group (2012).

Our dataset choice is inspired by Petersen et al. (2016), who compared the fused lasso additive model (FLAM) to generalized additive models (GAM) (Hastie, 2013) using 2013 Gallup World Polls data with twelve predictors. A key benefit of our quantile method is the ability to construct prediction intervals and assess coverage without domain-specific knowledge.

To assess QATF, we conduct 10-fold cross-validation on 113 countries with complete data, split into training and testing sets. We use $r = 0$, making this an even more direct extension of Petersen et al. (2016) to the quantile case, as FLAM is equivalent to Additive Trend Filtering when $r = 0$. We evaluate coverage for 60%, 80%, and 90% prediction intervals, deeming a prediction covered if it lies between the estimated quantiles, akin to swapping the lower and upper quantile estimates in the cases where they intersect Chernozhukov et al. (2010); Vardi and Zhang (2000). For example, with a 90% prediction interval, coverage (α) is:

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\hat{q}^i_{0.05} \leq Y^i_{\text{test}} \leq \hat{q}^i_{0.95}\}, \tag{18}$$

where $\hat{q}^i_{0.05}$ and $\hat{q}^i_{0.95}$ are predicted quantile for data Y^i_{test} based on X^i_{test}. Tuning parameters for the 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, and 0.95 quantiles are selected via 10-fold cross-validation. Results in Table 2 demonstrate QATF's superior coverage compared to quantile smoothing splines.

Table 2: Coverage α of various prediction intervals, averaged over 10 train-test splits.

Interval Width	QATF	QS
90%	0.894	0.798
80%	0.790	0.716
60%	0.634	0.602

Finally, we examine the component-wise results, demonstrating the interpretability advantages of additive models. The 0.1, 0.5, and 0.9 quantiles are trained on the entire dataset, and the component-wise results are plotted in Appendix R, Figure 4. With these individual component results, we can see, for example, that conditional on the other predictors, our model finds that Log-GDP per Capita is associated with increased happiness index scores (which is consistent with the findings in Petersen et al. (2016)). Figure 3 displays this association. When considering the quantiles specifically, we see that construction of prediction intervals relies almost exclusively on the Log-GDP per Capita and Perceptions of Corruption predictor variables, since the other variables demonstrate little to no difference in their component fits between the 0.1 and 0.9 quantiles.

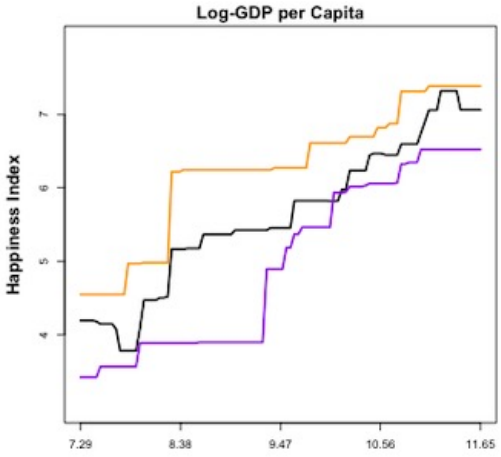


Figure 3: Log-GDP per Capita component fit in Quantile Additive Trend Filtering with $\tau = 0.1, 0.5,$ and 0.9, plotted in purple, black, and orange, respectively.

7 CONCLUSION

In total, we analyzed risk bounds for quantile additive trend filtering in both fixed and growing input dimensions. We also proposed and validated a practical algorithm using dimension-wise backfitting, effectively addressing challenges in additive models with heavy-tailed distributions.

Future work could extend this method to high-dimensional settings, incorporating sparsity penalties for enhanced performance when input dimensions are comparable to or larger than the sample size.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537.

Belloni, A., Chen, M., Madrid Padilla, O. H., and Wang, Z. (2023). High-dimensional latent panel quantile regression with an application to asset pricing. *The Annals of Statistics*, 51(1):96–121.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Benoit, D. F. and Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications*, 36(7):10475–10484.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Brantley, H. L., Guinness, J., and Chi, E. C. (2020). Baseline drift estimation for air quality data using quantile trend filtering.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.

Cantril, H. (1965). *Pattern of Human Concerns*. Rutgers University Press.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Coad, A., Rao, R., et al. (2006). Innovation and market value: a quantile regression analysis. *Economics Bulletin*, 15(13):1–10.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Ding, Q., Kang, Y., Liu, Y.-W., Lee, T. C. M., Hsieh, C.-J., and Sharpnack, J. (2022). Syndicated bandits: A framework for auto tuning hyper-parameters in contextual bandit algorithms. *Advances in Neural Information Processing Systems*, 35:1170–1181.

Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. (2019). Probabilistic forecasting with spline quantile function rnns. In *The 22nd international conference on artificial intelligence and statistics*, pages 1901–1910. PMLR.

Gu, J. and Mulvey, J. M. (2021). Factor momentum and regime-switching overlay strategy. *The Journal of Financial Data Science*, 3(4):101–129.

Guntuboyina, A., Lieu, D., Chatterjee, S., and Sen, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics*, 48(1):205–229.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer.

Härdle, W. and Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica neerlandica*, 47(1):43–57.

Hastie, T. (2013). gam: Generalized additive models. R package version 1.09.

Hastie, T. and Tibshirani, R. (1987). Non-parametric logistic and proportional odds regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):260–276.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., and Wang, S., editors (2024). *World Happiness Report 2024*. University of Oxford: Wellbeing Research Centre.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Jo, W. and Kim, D. (2023). Neural additive time-series models: Explainable deep learning for multivariate

time-series prediction. *Expert systems with applications*, 228:120307.

Johnson, N. A. (2013). A dynamic programming algorithm for the fused lasso and l 0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260.

Kang, Y., Hsieh, C.-J., and Lee, T. C. M. (2022). Efficient frameworks for generalized low-rank matrix bandit problems. *Advances in Neural Information Processing Systems*, 35:19971–19983.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM review*, 51(2):339–360.

Koenker, R. (2005). Quantile regression. *Cambridge Univ Pr*.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.

Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR.

Madrid Padilla, O. H. and Chatterjee, S. (2022). Risk bounds for quantile trend filtering. *Biometrika*, 109(3):751–768.

Mammen, E. and Van De Geer, S. (1997). Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413.

Mangasarian, O. L. and Schumaker, L. L. (1971). Discrete splines via mathematical programming. *SIAM Journal on Control*, 9(2):174–183.

Mulvey, J. M. and Liu, H. (2016). Identifying economic regimes: Reducing downside risks for university endowments and foundations. *Journal of Portfolio Management*, 43(1):100.

Narayan, T., Wang, S. L., Canini, K. R., and Gupta, M. (2023). Expected pinball loss for quantile regression and inverse cdf estimation. *Transactions on Machine Learning Research*.

Nielsen, J. P. and Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):43–61.

Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). *Fields: Tools for spatial data*. National Center for Atmospheric Research, Boulder, CO. R package version 10.3.

Pata, P. and Schindler, J. (2015). Astronomical context coder for image compression. *Experimental Astronomy*, 39:495–512.

Perlich, C., Rosset, S., Lawrence, R. D., and Zadrozny, B. (2007). High-quantile modeling for customer wallet estimation and other applications. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 977–985.

Petersen, A., Witten, D., and Simon, N. (2016). Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025. Epub 2016 Nov 10.

Ramdas, A. and Tibshirani, R. J. (2016). Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268.

Sadhanala, V. and Tibshirani, R. J. (2019). Additive models with trend filtering. *The Annals of Statistics*, 47(6):3032–3068.

Sadhanala, V., Wang, Y.-X., and Tibshirani, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. *Advances in Neural Information Processing Systems*, 29.

Sanderson, A. J., Ponman, T. J., and O'Sullivan, E. (2006). A statistically selected chandra sample of 20 galaxy clusters–i. temperature and cooling time profiles. *Monthly Notices of the Royal Astronomical Society*, 372(4):1496–1508.

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. (2021). Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*.

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. (2022). Estimation of non-crossing quantile regression process with deep requ neural networks. *arXiv preprint arXiv:2207.10442*.

Tagasovska, N. and Lopez-Paz, D. (2019). Single-model uncertainties for deep learning. *Advances in neural information processing systems*, 32.

Thielmann, A. F., Kruse, R.-M., Kneib, T., and Säfken, B. (2024). Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean. In *International Conference on Artificial Intelligence and Statistics*, pages 1783–1791. PMLR.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering.

Tibshirani, R. J. (2017). Dykstra's algorithm, admm, and coordinate descent: Connections, insights, and extensions. *Advances in Neural Information Processing Systems*, 30.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109:475–494.

Van Der Burg, E. and de Leeuw, J. (1983). Non-linear canonical correlation. *British journal of mathematical and statistical psychology*, 36(1):54–80.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate l 1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, Y.-X., Sharpnack, J., Smola, A., and Tibshirani, R. (2015). Trend filtering on graphs. In *Artificial Intelligence and Statistics*, pages 1042–1050. PMLR.

Wang, Y.-X., Smola, A., and Tibshirani, R. (2014). The falling factorial basis and its statistical applications. In *International Conference on Machine Learning*, pages 730–738. PMLR.

Wasko, C. and Sharma, A. (2014). Quantile regression for investigating scaling of extreme precipitation with temperature. *Water Resources Research*, 50(4):3608–3614.

World Bank Group (2012). *World Development Indicators 2012*. World Bank Publications.

Yano, K. and Kano, M. (2022). l1 trend filtering-based detection of short-term slow slip events: Application to a gnss array in southwest japan. *Journal of Geophysical Research: Solid Earth*, 127(5):e2021JB023258.

Ye, S. S. and Padilla, O. H. M. (2021). Non-parametric quantile regression via the k-nn fused lasso. *Journal of Machine Learning Research*, 22(111):1–38.

Checklist

1. For all models and algorithms presented, check if you include:
   - **a)** A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   - **b)** An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   - **c)** (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
2. For any theoretical claim, check if you include:
   - **a)** Statements of the full set of assumptions of all theoretical results. [Yes]
   - **b)** Complete proofs of all theoretical results. [Yes]
   - **c)** Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   - **a)** The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   - **b)** All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   - **c)** A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   - **d)** A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - **a)** Citations of the creator If your work uses existing assets. [Yes]
   - **b)** The license information of the assets, if applicable. [Yes]
   - **c)** New assets either in the supplemental material or as a URL, if applicable. [Yes]
   - **d)** Information about consent from data providers/curators. [Not Applicable]
   - **e)** Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   - **a)** The full text of instructions given to participants and screenshots. [Not Applicable]
   - **b)** Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   - **c)** The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A BROADER IMPACT AND ETHICAL STATEMENT

This paper contributes to the ongoing development of Machine Learning, adhering to standard ethical guidelines in research. Our work, aligning with common advancements in the field, does not present any unique ethical dilemmas or societal consequences that require special emphasis. We recognize the importance of responsible use and development of Machine Learning technologies and their potential impact on society. However, our research does not delve into areas that might raise significant ethical or societal concerns. We commit to ethical research practices and consider the broader implications of our work to be aligned with the typical advancements in Machine Learning.

B CONTRIBUTION OF PROOF TECHNIQUES

Our proof process establishes the Huber type loss (defined in (13)) is upper bounded by the localized Rademacher width, as delineated in Lemmas E.5 to Theorem E.12. We then establish the upper bound of local Rademacher width in terms of Huber type metric by the local Rademacher width in terms of ℓ_2 type metric, which needs us to convert the ℓ_2 norm in local Rademacher width to the Huber loss function in its local Rademacher width. This conversion is detailed in E.13 to E.25, with specific emphasis on Lemmas E.17, E.21, E.22, E.24, and is culminated in Equation (142) of Lemma E.25. In addition, our proof executes this conversion without amplifying the dimensionality. In particular, this is achieved by orthogonally projecting the estimation error into the null space of the trend filtering operator and its orthogonal complement, while controlling for dimensionality, as demonstrated in Lemmas I.2, I.4, I.3, and I.5. As a result, we maintain a minimax rate for arbitrary errors, irrespective of moment conditions. This part of our proof technique represents one of the significant technical contributions of this work.

C PROOF SKETCH

Below we summarize our proofs for Theorems 3.1 and 3.3, with the detailed proofs provided in the Appendix.

Step 1. The first step in our analysis is to show that the convergence rate of our estimator depends on a local Rademacher complexity, which involves a constraint enforced by the total variation and one constraint consisting of a ball of the form $\{\delta : \Delta_n^2(\delta) \leq r\}$.

Step 2. The geometry of the set $\{\delta : \Delta_n^2(\delta) \leq r\}$ is extremely complicated because the function $\Delta_n(\cdot)$ does not satisfy the triangle inequality. To deal with this, we use the formula $\|\delta\|_2^2 \leq \max\{\|\delta\|_\infty, 1\}\Delta_n^2(\delta)$. The idea behind this is to replace the set $\{\delta : \Delta_n^2(\delta) \leq r\}$ with a set of the form $\{\delta : \|\delta\|_2^2 \leq \tilde{r}\}$ for an appropriate $\tilde{r} > 0$. However, to do this we must control $\|\delta\|_\infty$ for δ in the set $\{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta_n^2(\delta) \leq t^2\}$. This in itself is challenging, and it is done by orthogonally projecting the vectors $\delta(X)$ into the null space of the trend filtering operator resulting in p_j and projecting onto the orthogonal complement leading to q_j, where $j = 1, \ldots, d$. This has to be done for each dimension j.

Step 3. Controlling the ℓ_∞ of the projections onto both spaces is nontrivial, but we achieve this in Lemmas E.17, E.20 and E.22. The next step is to obtain bounds on $\|p\|_\infty$ and $\|q\|_\infty$, where $p = p_1 + \cdots + p_d$ and $q = q_1 + \cdots + q_d$, then obtain the $\Delta_n^2(p)$ and $\Delta_n^2(q)$ afterwards. Lemmas E.23, E.24, and E.25 derive these steps. The higher-order term $\Delta_n^2(q)$ would need conversion to $\ell_2(q)$ to obtain the optimal rate. Lemma E.25 deals with that.

D ASSUMPTIONS FOR THEOREM 3.1

Before providing those main results, we give some detail description of assumptions as needed.

Assumption D.1. Assume the inputs are contained in $[0, 1]$ in each dimension such that

$$0 \leq X_j^i \leq 1, \quad i = 1, \ldots, n, \quad j = 1, \ldots, d. \tag{19}$$

Then we assume the maximum width w for any consecutive X_j^i satisfies, with high probability, that $w = O(\frac{\log n}{n})$, where w is given by (20),

$$w = \max_{\substack{i=1,\ldots,n-1 \\ j=1,\ldots,d}} \left| X_j^{(i)} - X_j^{(i-1)} \right|, \tag{20}$$

where $X_j^{(1)} < X_j^{(2)} < \cdots < X_j^{(n)}$.

Remark D.2. The $[0,1]$ assumption is minor, as we can always scale the inputs. The same constraint is used in the analysis in Sadhanala and Tibshirani (2019). The assumption for w would be satisfied with probability approaching 1 by Lemma 5 in Wang et al. (2015), for $X_j^1, \ldots, X_j^n$ that are i.i.d. from a uniform distribution $[0,1]$. In the literature of nonparametric statistics, it is common to assume covariates $X^i \in \mathbb{R}^d$ that are uniformly distributed on $[0,1]^d$, for instance, see Definition 3.4 of Györfi et al. (2002). The latter model is actually a particular case of the class of models that we consider.

Assumption D.3. For the f_0 in (1), we assume that $\sum_{i=1}^n f_{0j}(X_j^i) = 0, \quad \text{for } j = 1, \ldots, d.$

Our Assumption D.3 is consistent with Assumption B1 by Sadhanala and Tibshirani (2019), and it is required for identifiability.

Assumption D.4. There exists a constants $L > 0$ and $\underline{f} > 0$ such that for any function $\delta : [0,1]^d \to \mathbb{R}$ with $\|\delta\|_\infty \leq L$ we have that

$$\left|F_{Y^i|X^i}(f_0^i + \delta^i) - F_{Y^i|X^i}(f_0^i)\right| \geq \underline{f}\left|\delta^i\right|, \tag{21}$$

where $F_{Y^i|X^i}$ is the conditional CDF of Y^i, and $\delta^i = \delta(X^i)$. Note for a given quantile level τ, for the input X^i, we define the evaluation of f_0 at X^i as $f_0^i = F^{-1}_{Y^i|X^i}(\tau)$, where $F^{-1}_{Y^i|X^i}(\tau)$ represents the τ conditional quantile of Y^i given X^i.

Similar conditions are assumed by Assumption A of Madrid Padilla and Chatterjee (2022) and Assumption 4 of Shen et al. (2022). Assumption D.4 ensures that the conditional quantile of Y^i given X^i is uniquely defined and there is a uniform linear growth of the CDF around a neighborhood of the quantile.

E SUPPLEMENTARY LEMMAS FOR THEOREM 3.1

To complete our proofs for the theorems, we provide additional definitions.

Definition E.1. The rth order TV-ball of radius V is defined as:

$$B_{TV^{(r)}}(V) = \left\{f : TV^{(r)}(f) \leq V\right\}. \tag{22}$$

We also denote

$$B^d_{TV^{(r)}}(V) = \left\{\sum_{j=1}^d f_j : \sum_{j=1}^d TV^{(r)}(f_j) \leq V, j = 1, \ldots, d\right\}. \tag{23}$$

The rth order trend filtering operator norm ball of radius V as

$$B_{D_n^{(X,r)}}(V) = \left\{\theta \in \mathbb{R}^n : \left\|D_n^{(X,r)}(\theta)\right\|_1 \leq V\right\}. \tag{24}$$

We abbreviate $B_n(V)$ for the $\|\cdot\|_n$-ball of radius V, $B_{L_2}(V)$ for the $\|\cdot\|_2$-ball of radius V, and $B_\infty(V)$ for the $\|\cdot\|_\infty$-ball of radius V.

Definition E.2. Let $Y \in \mathbb{R}^n$ be a vector of independent random variables and for a given quantile level $\tau \in (0,1)$. For any finite function class $\mathcal{F}$, we define the Empirical Rademacher width as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi^i f^i(X^i) \mid \left\{X^i\right\}_{i=1}^n\right],$$

and its Rademacher complexity

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \xi^i f(X^i)\right].$$

Definition E.3. For $f \in \mathcal{F}^{(r)}(V)$, we define the empirical loss function as

$$\widehat{M}(f) = \sum_{i=1}^{n} \widehat{M}^i(f), \tag{25}$$

where each $\widehat{M}^i(f)$ is defined as

$$\widehat{M}^i(f) = \rho_\tau(Y^i - f(X^i)) - \rho_\tau(Y^i - f_0(X^i)). \tag{26}$$

We also define the population loss function as

$$M(f) \coloneqq \sum_{i=1}^{n} M^i(f), \tag{27}$$

where

$$M^i(f) \coloneqq \mathbb{E}\left[\rho_\tau(Y^i - f(X^i)) - \rho_\tau(Y^i - f_0(X^i))|X^i\right]. \tag{28}$$

It is well known that for our model in (1), conditioned on X^i, the true τ-quantile of Y^i, $f_0(X^i)$ satisfies

$$f_0 = \arg\min_{f\in\mathcal{F}^{(r)}(V)} M(f), \tag{29}$$

where the expectation is taken for Y^i conditioned on X^i.

Note, from (29), we know $f_0 \in \arg\min M(f)$, and $M(f_0) = 0$.

Definition E.4. The constrained quantile additive estimator defined in (11) will be

$$\widehat{f} = \operatorname*{arg\,min}_{f\in\mathcal{F}^{(r)}(V)} \widehat{M}(f), \tag{30}$$

By Definitions in (29) and (25), we have

$$\widehat{M}(\widehat{f}) - \widehat{M}(f^*) \leq 0 \tag{31}$$

as the basic inequality.

Lemma E.5. For any $f, f_0 \in \mathcal{F}^{(r)}(V)$, denote $\delta(X^i) = f(X^i) - f_0(X^i)$, then we have

$$M^i(f) - M^i(f_0) = \int_0^{\delta(X^i)} \left(F_{Y^i|X^i}(f_0(X^i) + z) - F_{Y^i|X^i}(f_0(X^i))\right) dz. \tag{32}$$

Proof. By equation (B.3) in Belloni and Chernozhukov (2011),we have

$$\rho_\tau(w - v) - \rho_\tau(w) = -v(\tau - \mathbb{I}\{w \leq 0\}) + \int_0^v \{\mathbb{I}\{w \leq z\} - \mathbb{I}\{w \leq 0\}\} dz. \tag{33}$$

Given any f and X^i, let $w = Y^i - f_0(X^i)$, $v = f(X^i) - f_0(X^i)$. Then taking conditional expectations on above

equation with respect to Y^i conditioned on X^i on both sides, we have

$$\begin{aligned} M^i(f) - M^i(f_0) &= \mathbb{E}\left[\rho_\tau(Y^i - f(X^i)) - \rho_\tau(Y^i - f_0(X^i))|X^i\right] \\ &= \mathbb{E}\left[-\delta(X^i)(\tau - \mathbb{I}\{Y^i - f(X^i) \le 0\})|X^i\right] \\ &\quad + \mathbb{E}\left[\int_0^{\delta(X^i)} \left(\mathbb{I}\{Y^i - f_0(X^i) \le z\} - \mathbb{I}\{Y^i - f_0(X^i) \le 0\}dz|X^i\right)\right] \\ &= 0 + \int_0^{\delta(X^i)} \left(F_{Y^i|X^i}(f_0(X^i) + z) - F_{Y^i|X^i}(f_0(X^i))\right) dz \\ &= \int_0^{\delta(X^i)} \left(F_{Y^i|X^i}(f_0(X^i) + z) - F_{Y^i|X^i}(f_0(X^i))\right) dz, \end{aligned} \tag{34}$$

for all $i = 1, \ldots, n$. □

Lemma E.6. Let $\Delta^2(\delta) \coloneqq n\Delta_n^2(\delta)$, assume Assumption D.4 holds, then there exists c_0 such that for all $\delta = f - f_0$, we have

$$M(f_0 + \delta) \ge c\Delta^2(\delta), \tag{35}$$

where c is a positive constant that depends on L and $\underline{f}$ in Assumption D.4.

Proof. Supposing that $\left|\delta(X^i)\right| \le L$, by Assumption D.4 and Lemma E.5, we have

$$\begin{aligned} M^i(f) - M^i(f_0) &\ge \int_0^{\delta(X^i)} \underline{f}\,|z|\,dz \\ &= \frac{\underline{f}\delta(X^i)^2}{2} \end{aligned} \tag{36}$$

Supposing that $\delta(X^i) > L$, by Lemma E.5, we have

$$\begin{aligned} M^i(f) - M^i(f_0) &\ge \int_{L/2}^{\delta(X^i)} \left(F_{Y^i|X^i}(f_0(X^i) + z) - F_{Y^i|X^i}(f_0(X^i))\right) dz \\ &\ge \int_{L/2}^{\delta(X^i)} \left(F_{Y^i|X^i}(f_0(X^i) + L/2) - F_{Y^i|X^i}(f_0(X^i))\right) dz \\ &= (\delta(X^i) - L/2)\left(F_{Y^i|X^i}(f_0(X^i) + L/2) - F_{Y^i|X^i}(f_0(X^i))\right) \\ &\ge \frac{\delta(X^i)}{2}\frac{L\underline{f}}{2} \coloneqq c\left|\delta(X^i)\right|, \end{aligned} \tag{37}$$

where the first two inequalities follow because $F_{Y^i|X^i}$ is monotone, where c is a positive constant that depends on L and $\underline{f}$ in Assumption D.4. Then for the case $\delta(X^i) < -L$, it can be handled similarly. The conclusion follows combining the three different cases. □

Lemma E.7 (Symmetrization)**.** For $t > 0$, it holds that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}^{(r)}(V), M(f) \le t^2} \left(M(f) - \widehat{M}(f)\right) \mid \left\{X^i\right\}_{i=1}^n\right] \le 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}^{(r)}(V), M(f) \le t^2} \sum_{i=1}^n \xi^i \widehat{M}^i(f) \mid \left\{X^i\right\}_{i=1}^n\right], \tag{38}$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher variables independent of $\left\{Y^i\right\}_{i=1}^n$.

Proof. Let $\tilde{Y}^1, \ldots \tilde{Y}^n$ be an independent and identically distributed copy of $Y^1, \ldots, Y^n$. Let $\tilde{M}^i$ the version of $\widehat{M}^i$ corresponding to $\tilde{Y}^1, \ldots \tilde{Y}^n$. If we define

$$X_f = \sum_{i=1}^n \left(\tilde{M}^i(f) - \widehat{M}^i(f)\right) \tag{39}$$

then we have

$$\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\mathbb{E}\left[X_f|\left\{Y^i\right\}_{i=1}^n,\left\{X^i\right\}_{i=1}^n\right]\le\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}X_f|\left\{Y^i\right\}_{i=1}^n,\left\{X^i\right\}_{i=1}^n\right]. \tag{40}$$

The expectation is taken with respect to $\tilde{Y}^1,\dots\tilde{Y}^n$ conditioned on $Y^1,\dots,Y^n,X^1,\dots,X^n$.

Then, the left-hand side of the above, can be further written as

$$\begin{aligned}
&\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\mathbb{E}\left[X_f|\left\{Y^i\right\}_{i=1}^n,\left\{X^i\right\}_{i=1}^n\right]\\
&=\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\left(\mathbb{E}\left[\sum_{i=1}^n\tilde{M}^i(f)\mid\left\{Y^i\right\}_{i=1}^n,\left\{X^i\right\}_{i=1}^n\right]-\mathbb{E}\left[\sum_{i=1}^n\widehat{M}^i(f)\mid\left\{Y^i\right\}_{i=1}^n,\left\{X^i\right\}_{i=1}^n\right]\right) \qquad (41)\\
&=\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\left(M^i(f)-\widehat{M}^i(f)\right).
\end{aligned}$$

Combing (39) and (41), we get

$$\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\left(M^i(f)-\widehat{M}^i(f)\right)\le\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\left(\tilde{M}^i(f)-\widehat{M}^i(f)\right)|\left\{Y^i\right\}_{i=1}^n,\left\{X^i\right\}_{i=1}^n\right]. \tag{42}$$

Further take the expectation with respect of $Y^1,\dots,Y^n$ on both sides, and similar to the proof of the Lemma 10 in Madrid Padilla and Chatterjee (2022), we can get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\left(M^i(f)-\widehat{M}^i(f)\right)\mid\left\{X^i\right\}_{i=1}^n\right]\\
&\le\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\left(\tilde{M}^i(f)-\widehat{M}^i(f)\right)\mid\left\{X^i\right\}_{i=1}^n\right]\\
&=\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\xi^i\left(\tilde{M}^i(f)-\widehat{M}^i(f)\right)\mid\left\{X^i\right\}_{i=1}^n\right] \qquad (43)\\
&\le\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\xi^i\tilde{M}^i(f)\mid\left\{X^i\right\}_{i=1}^n\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n-\xi^i\widehat{M}^i(f)\mid\left\{X^i\right\}_{i=1}^n\right]\\
&=2\,\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\xi^i\widehat{M}^i(f)\mid\left\{X^i\right\}_{i=1}^n\right],
\end{aligned}$$

where the first equality follows because $\xi_1(\tilde{M}^i(f)-\widehat{M}^i(f),\dots,\xi_n(\tilde{M}_n(f)-\widehat{M}_n(f))$ and $(\tilde{M}^i(f)-\widehat{M}^i(f),\dots,(\tilde{M}_n(f)-\widehat{M}_n(f))$ have the same distribution. The second equality follows because $-\xi_1,\dots,-\xi_n$ are also independent Rademacher variables. □

Lemma E.8 (Contraction principle)**.** With the notation from before we have that, for $t>0$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n\xi^i\widehat{M}^i(f)\mid\left\{X^i\right\}_{i=1}^n\right]\le n\mathcal{R}_n(\left\{f-f_0:f\in\mathcal{F}^{(r)}(V)\right\}\cap\left\{f-f_0:M(f)\le t^2\right\}). \tag{44}$$

Proof. Based on the definition of $M^i(f)$,

$$\widehat{M}^i(f)=\rho_\tau(Y^i-f(X^i))-\rho_\tau(Y^i-f_0(X^i)). \tag{45}$$

From Lemma E.9, $\widehat{M}^i(f)$ are 1-Lipschitz continuous functions, thus,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n \xi^i\widehat{M}^i(f) \mid \{X^i\}_{i=1}^n\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n \xi^i\widehat{M}^i(f) \mid \{X^i\}_{i=1}^n,\{Y^i\}_{i=1}^n\right] \mid \{X^i\}_{i=1}^n\right]\\
&\le\mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n \xi^i f^i \mid \{X^i\}_{i=1}^n,\{Y^i\}_{i=1}^n\right] \mid \{X^i\}_{i=1}^n\right]\\
&=\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n \xi^i f^i \mid \{X^i\}_{i=1}^n\right]\\
&\le\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n \xi^i (f^i-f_0^i) \mid \{X^i\}_{i=1}^n\right]+\mathbb{E}\left[\sum_{i=1}^n \xi^i f_0^i \mid \{X^i\}_{i=1}^n\right]\\
&=\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V),M(f)\le t^2}\sum_{i=1}^n \xi^i (f^i-f_0^i)(X^i) \mid \{X^i\}_{i=1}^n\right].
\end{aligned} \tag{46}$$

The third line holds since we have used the 1-Lipschitz condition, so the expectation inside does not have y anymore. □

Lemma E.9. The function $\widehat{M}^i(f)$ is 1-lipschitz.

Proof. For any f, g, and fixed τ, we have

$$\begin{aligned}
&\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right|\\
&=\left|\rho_\tau(Y^i-f(X^i))-\rho_\tau(Y^i-g(X^i))\right|\\
&=\left|\max\left\{\tau(Y^i-f(X^i)),(1-\tau)(Y^i-f(X^i))\right\}-\max\left\{\tau(Y^i-g(X^i)),(1-\tau)(Y^i-g(X^i))\right\}\right|.
\end{aligned} \tag{47}$$

It is easy to see the conclusion satisfied if $Y^i-f(X^i)$ and $Y^i-g(X^i)$ have the same signs.

Assume $Y^i-f(X^i)>0$ and $Y^i-g(X^i)\le 0$, we have Equation (47) equals

$$\begin{aligned}
&\left|\max\left\{\tau(Y^i-f(X^i)),(1-\tau)(Y^i-f(X^i))\right\}-\max\left\{\tau(Y^i-g(X^i)),(1-\tau)(Y^i-g(X^i))\right\}\right|\\
&=\left|\tau(Y^i-f(X^i))-(1-\tau)(Y^i-g(X^i))\right|\\
&=\left|\tau(g(X^i)-f(X^i))+(2\tau-1)(Y^i-g(X^i))\right|.
\end{aligned} \tag{48}$$

If $(2\tau-1)\le 0$, since $Y^i-f(X^i)>0$ and $g(X^i)>f(X^i)$, then we know $0\le(2\tau-1)(Y^i-g(X^i))\le(2\tau-1)(f(X^i)-g(X^i))$, we get

$$\begin{aligned}
\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right|&\le\left|\tau(g(X^i)-f(X^i))+(2\tau-1)(f(X^i)-g(X^i))\right|\\
&=\left|(\tau-1)(f(X^i)-g(X^i))\right|\le\left|f(X^i)-g(X^i)\right|,
\end{aligned} \tag{49}$$

On the other hand, if $(2\tau-1)\ge 0$, if further $\left|(2\tau-1)(Y^i-g(X^i))\right|\ge\tau(g(X^i)-f(X^i))$, and $\tau(g(X^i)-f(X^i))>\tau(g(X^i)-Y^i)\ge 0$, science $Y^i<f(X^i)\le g(X^i)$, then we have

$$\begin{aligned}
\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right|&\le\left|\tau(g(X^i)-Y^i)+(2\tau-1)(Y^i-g(X^i))\right|\\
&=\left|(\tau-1)(Y^i-g(X^i))\right|\le\left|(\tau-1)(f(X^i)-g(X^i))\right|\le\left|f(X^i)-g(X^i)\right|,
\end{aligned} \tag{50}$$

if $\left|(2\tau-1)(Y^i-g(X^i))\right|\le\tau(g(X^i)-f(X^i))$, then we have

$$\begin{aligned}\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right| &\le \left|\tau(g(X^i)-f(X^i))+(2\tau-1)(f(X^i)-g(X^i))\right| \\ &= \left|(\tau)(f(X^i)-g(X^i))\right| \le \left|f(X^i)-g(X^i)\right|.\end{aligned} \tag{51}$$

Thus, by (49), (50) and (51), we have

$$\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right| \le \left|f(X^i)-g(X^i)\right|.$$

For the other case, we assume $Y^i - f(X^i) \le 0$ and $Y^i - g(X^i) > 0$,

Since

$$\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right| = \left|\widehat{M}^i(g)-\widehat{M}^i(f)\right|,$$

we thus can derive the same conclusion by applying the above analysis, thus, we have

$$\left|\widehat{M}^i(f)-\widehat{M}^i(g)\right| \le \left|f(X^i)-g(X^i)\right|, \tag{52}$$

which indicates that $\widehat{M}^i(f)$ is a 1-lipschitz function. □

Proposition E.10. Let $\mathcal{F}^{(r)}(V)$ be a function class, then the following inequality is true for any $t > 0$,

$$\mathbb{P}\left(M(\widehat{f}) > t^2 \mid \{X^i\}_{i=1}^n\right) \le \frac{2n}{t^2}\mathcal{R}_n(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \{f - f_0 : M(f) \le t^2\}). \tag{53}$$

Proof. Suppose that

$$M(\widehat{f}) > t^2. \tag{54}$$

First, define $g : [0,1] \to \mathbb{R}$ as $g(u) = M\left((1-u)f_0 + u\widehat{f}\right)$. Clearly, g is a continuous function with $g(0) = 0$, and $g(1) = M(\widehat{f})$. Therefore, there exists $\mu_{\widehat{f}} \in [0,1]$ such that $g(\mu_{\widehat{f}}) = t^2$. Hence, letting $\tilde{f} = (1-\mu_{\widehat{f}})f_0 + \mu_{\widehat{f}}\widehat{f}$, we observe that by the convexity of $\widehat{M}$ and the basic inequality, we have

$$\widehat{M}(\tilde{f}) = \widehat{M}\left((1-\mu_{\widehat{f}})f_0 + \mu_{\widehat{f}}\widehat{f}\right) \le (1-\mu_{\widehat{f}})\widehat{M}(f_0) + \mu_{\widehat{f}}\widehat{M}(\widehat{f}) \le 0. \tag{55}$$

Furthermore, $\mathcal{F}^{(r)}(V)$ from (11) is a convex set, and f_0 and $\widehat{f}$ are belong to $\mathcal{F}^{(r)}(V)$, thus we know $\tilde{f}$ belongs to $\mathcal{F}^{(r)}(V)$, furthermore, we have $M(\tilde{f}) = t^2$ by construction. This implies that

$$\sup_{f\in\mathcal{F}^{(r)}(V), M(f)\le t^2} M(f) - \widehat{M}(f) \ge M(\tilde{f}) - \widehat{M}(\tilde{f}) \ge M(\tilde{f}). \tag{56}$$

The first inequality holds by the supreme, and the second inequality is held by $\widehat{M}(\tilde{f}) \le 0$. In the results shown above, it is true that

$$M(\widehat{f}) > t^2 \implies \sup_{f\in\mathcal{F}^{(r)}(V), M(f)\le t^2} M(f) - \widehat{M}(f) > t^2. \tag{57}$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(M(\widehat{f}) > t^2 \mid \{X^i\}_{i=1}^n\right) &\le \mathbb{P}\left(\sup_{f\in\mathcal{F}^{(r)}(V), M(f)\le t^2} M(f) - \widehat{M}(f) > t^2 \mid \{X^i\}_{i=1}^n\right) \\ &\le \frac{1}{t^2}\,\mathbb{E}\left[\sup_{f\in\mathcal{F}^{(r)}(V), M(f)\le t^2} M(f) - \widehat{M}(f) \mid \{X^i\}_{i=1}^n\right] \\ &\le \frac{2n}{t^2}\mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \{f - f_0 : M(f) \le t^2\}\right).\end{aligned} \tag{58}$$

The third inequality is by Lemma E.7 and E.8. □

Theorem E.11. Suppose the distribution of $Y^1, \ldots, Y^n$ obey Assumption D.4, then the following inequality is true for any $t > 0$,

$$\mathbb{P}\left(\Delta^2(\widehat{f} - f_0) > t^2 \mid \{X^i\}_{i=1}^n\right) \leq \frac{cn}{t^2}\mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \left\{f - f_0 : \Delta^2(f - f_0) \leq t^2\right\}\right), \tag{59}$$

where c is the same constant in Lemma E.6 that only depends on the distribution of $X^1, \ldots, X^n$, and the distribution of $Y^1, \ldots, Y^n$, and $\mathcal{F}^{(r)}(V)$ is a convex set with $\widehat{f}, f_0 \in \mathcal{F}^{(r)}(V)$, and $\mathcal{R}_n$ is defined in (E.2).

Proof. It is true that

$$\begin{aligned}
\mathbb{P}\left(\Delta^2(\widehat{f} - f_0) > t^2 \mid \{X^i\}_{i=1}^n\right) &\leq \mathbb{P}\left(M(\widehat{f}) > ct^2 \mid \{X^i\}_{i=1}^n\right) \\
&\leq \frac{cn}{t^2}\mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \left\{f - f_0 : M(f) \leq t^2\right\}\right) \\
&\leq \frac{cn}{t^2}\mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \left\{f - f_0 : \Delta^2(f - f_0) \leq t^2\right\}\right),
\end{aligned} \tag{60}$$

where the first inequality follows from Lemma E.6, the second inequality follows from Proposition E.10, and the third inequality follows from the fact that $\left\{f - f_0 : M(f) \leq t^2\right\} \subset \left\{f - f_0 : \Delta^2(f - f_0) \leq t^2\right\}$. □

We would need to upper bound the quantity

$$\mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \left\{f - f_0 : \Delta^2(f - f_0) \leq t^2\right\}\right), \tag{61}$$

where

$$\mathcal{F}^{(r)}(V) - f_0 = \left\{f - f_0 = \sum_{j=1}^d (f_j - f_{0j}) : f_j \in \mathcal{H}_j, \sum_{j=1}^d TV^{(r)}(f_j) \leq V, \sum_{i=1}^n f_j(X_j^i) = 0\right\}, \tag{62}$$

and

$$\Delta^2(f) = \sum_{i=1}^n \min\left\{\left|\sum_{j=1}^d f_j(X_j^i)\right|, \left(\sum_{j=1}^d f_j(X_j^i)\right)^2\right\}. \tag{63}$$

Theorem E.12. Suppose $f_j \in \mathcal{H}_j$, the space spanned by the falling factorial basis, and for the fixed points of $X^1, \ldots, X^n, Y^1, \ldots, Y^n$ obey Assumption D.4,

(a) Let

$$f_j = (f_j(X_j^1), \ldots, f_j(X_j^n)), \tag{64}$$

$f = \sum_{j=1}^d f_j$, then the following inequality is true for any $t > 0$,

$$\mathbb{P}\left(\Delta^2(\widehat{f} - f_0) > t^2 \mid \{X^i\}_{i=1}^n\right) \leq \frac{cn}{t^2}\mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \left\{f - f_0 : \Delta^2(f - f_0) \leq t^2\right\}\right), \tag{65}$$

where $\mathcal{F}^{(r)}(V)$ is defined in (11), where c is a constant that only depends on the distribution of $X^1, \ldots, X^n, Y^1, \ldots, Y^n$.

(b) Let

$$\theta_j = (\theta_j^1, \ldots, \theta_j^n), \tag{66}$$

$\theta_j^i = f_j(X_j^i)$, and let θ_0 is given by

$$\theta_0 = \min_{\theta \in K^{(X,r)}(V)} \sum_{i=1}^n \mathbb{E}\left[\rho_\tau(Y^i - \theta^i) - \rho_\tau(Y^i - \theta_0^i)|X^i\right], \tag{67}$$

then the following inequality is true for any $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\Delta^2(\widehat{f} - f_0) > t^2 \mid \{X^i\}_{i=1}^n\right) &= \mathbb{P}\left(\Delta^2(\widehat{\theta} - \theta_0) > t^2 \mid \{X^i\}_{i=1}^n\right) \\ &\leq \frac{cn}{t^2}\mathcal{R}_n\left(\left\{\theta - \theta_0 : \theta \in K^{(X,r)}(V)\right\} \cap \left\{\theta - \theta_0 : \Delta^2(\theta - \theta_0) \leq t^2\right\}\right),\end{aligned} \tag{68}$$

where $K^{(X,r)}(V)$ is defined in (7), $\widehat{\theta}$ is an estimator in $K^{(X,r)}(V)$, and c is a constant that only depends on the distribution of $X^1, \ldots, X^n$, and the distribution of $Y^1, \ldots, Y^n$.

Proof. For (a), the conclusion (the inequality) is derived based on Theorem E.11. For (b), the first equality holds because the equivalent formulation of additive trend filtering from Equation (11) and Equation (6), then a similar conclusion is derived for the second inequality. □

By the definition of $TV^{(r)}$, $TV^{(r)}$ is a seminorm $TV^{(r)}$, since its domain is contained in the space of $(r-1)$ times weakly differentiable functions, and its null space contains all $(r-1)$th order polynomials.

Definition E.13. Let $f, f_0 \in \mathcal{F}^{(r)}(V)$. Let $\delta = f - f_0 \in \mathcal{F}^{(r)}(V)$, let $\delta = \sum_{j=1}^d \delta_j$, with each $\delta_j = f_j - f_{0j}$, for $j = 1, \ldots, d$.

1. Define a subset that is contained in the space of $(r-1)$times weakly differentiable functions.

$$\mathcal{W} = \left\{g : [0,1] \to \mathbb{R}, TV^{(r)}(g) \leq V\right\}.$$

   Clearly, we have $f_j \in \mathcal{W}$, $f_{0j} \in \mathcal{W}$, and $\delta_j \in \mathcal{W}$.

2. Define a subspace of $\mathbb{R}^n$ that spanned by functions in $\mathcal{W}$ evaluated at points $X_j = (X_j^1, \ldots, X_j^n)$ as

$$W_j = \text{Span}\left\{(g(X_j^1), \ldots, g(X_j^n)), g \in \mathcal{W}\right\}. \tag{69}$$

   For any δ_j, let the vector $\delta_j(X_j)$ denote δ_j evaluated at points $X_j = (X_j^1, \ldots, X_j^n)$:

$$\delta_j(X_j) \coloneqq (\delta_j(X_j^1), \ldots, \delta_j(X_j^n)).$$

   Let the n-dimensional vector obtained by adding $\delta_j(X_j)$ be denoted as

$$\delta(X) \coloneqq \left(\sum_{j=1}^d \delta_j(X_j^1), \ldots, \sum_{j=1}^d \delta_j(X_j^n)\right) = (\delta(X^1), \ldots, \delta(X^n)).$$

   This implies that $\delta(X) = \sum_{j=1}^d \delta_j(X_j), \delta_j \in \mathcal{W}$. We will see $\delta(X^i)$ equals to $\sum_{j=1}^d \delta_j(X_j^i)$.

3. Let $\phi_\ell(\nu)$ be the ℓth degree polynomials $\ell \leq (r-1)$, Define P_j be the space spanned by all $(r-1)$th order polynomials evaluated at points $X_j = (X_j^1, \ldots, X_j^n)$ as

$$P_j \coloneqq \text{Span}\left\{(\phi_\ell(X_j^1), \ldots, \phi_\ell(X_j^n)), l = 0, \ldots, r-1\right\}. \tag{70}$$

4. Define the orthogonal complement of P_j in terms of $\|\cdot\|_n$ as

$$Q_j \subseteq W_j : Q_j = P_j^\perp, \quad Q_j \oplus P_j = W_j. \tag{71}$$

   In other words, for any $f_j(X_j) \in P_j$ and $g_j(X_j) \in Q_j$, it holds $\langle f_j, g_j \rangle_n = 0$.

5. Define the orthogonal projection operators applied to W_j for P_j and Q_j to be

$$\Pi_{P_j} : \Pi_{P_j} t \in P_j, \quad \Pi_{Q_j} : \Pi_{Q_j} t \in Q_j, \quad \text{with} \quad t \in W_j. \tag{72}$$

6. For any $\delta \in \mathcal{F}^{(r)}(V)$, with the vector $\delta(X) = (\delta(X^1), \ldots, \delta(X^n)) \in \mathbb{R}^n$, with Π_{P_j} and Π_{Q_j} defined in (72), we define Π_P and Π_Q as

$$\Pi_P : \Pi_P \delta(X) = \Pi_{P_1} \delta_1(X_j^{(1)}) + \cdots + \Pi_{P_d} \delta_d(X_d), \tag{73}$$

and

$$\Pi_Q : \Pi_Q \delta(X) = \Pi_{Q_1} \delta_1(X_j^{(1)}) + \cdots + \Pi_{Q_d} \delta_d(X_d). \tag{74}$$

Then we denote

$$p_j \coloneqq \Pi_{P_j} \delta_j(X_j) \in \mathbb{R}^n, \quad q_j \coloneqq \Pi_{Q_j} \delta_j(X_j) \in \mathbb{R}^n.$$

And let $p^i = (p_1^i, \ldots, p_d^i)$, $\quad q^i = (q_1^i, \ldots, q_d^i)$, where $p_1^i, \ldots, p_d^i$ are the elements of p^i, and $q_1^i, \ldots, q_d^i$ are the elements of q^i. And denote

$$p \coloneqq p_1 + \cdots + p_d \in \mathbb{R}^n, \quad q \coloneqq q_1 + \cdots + q_d \in \mathbb{R}^n.$$

So we have $\delta_j(X_j) = p_j + q_j$, $\Pi_P \delta(X) = p$, and $\Pi_Q \delta(X) = q$.

Lemma E.14. Let $v_j \in \mathbb{R}^{n \times r}$ be a matrix whose ith row is given by

$$(1, X_j^{(i)}, (X_j^{(i)})^2, \ldots, (X_j^{(i)})^{r-1}) \in \mathbb{R}^r, \tag{75}$$

where $X_j^{(i)}, i = 1, \ldots, n$ are the sorted data points in $[0, 1]$. By the definition of P_j in Definition E.13, it holds that

$$P_j = \text{Span}\left\{\text{columns of } v_j\right\}. \tag{76}$$

Furthermore, for $D_n^{(X_j, r)}$, it holds $\text{Span}\left\{\text{columns of } v_j\right\} = \text{null}(D_n^{(X_j, r)})$.

Proof. Clearly, P_j equals to the column space of v_j.

We know $D_n^{(X_j, r)} \in \mathbb{R}^{(n-r) \times n}$ as outlined in 4 is the discrete difference operator. It is evident that for any $\nu \in \text{Span}\left\{\text{columns of } v_j\right\}$, we have $D_n^{(X_j, r)} \nu = 0$. This follows from Lemma 1 and Assumption C1 in Sadhanala and Tibshirani (2019). Lemma 1 in Sadhanala and Tibshirani (2019) expresses the $TV^{(r)}$ of a function f in terms of the falling factorial basis. Since the $TV^{(r)}$ of an $(r-1)$th order polynomial is zero, the result follows. Furthermore, the dimension of $\text{Span}\left\{\text{columns of } v_j\right\}$ is equal to r. By the definition of $D_n^{(X_j, r)}$, the dimension of $\text{row}(D_n^{(X_j, r)})$ is equal to $n - r$. Then the dimension of the $\text{null}(D_n^{(X_j, r)})$ is equal to $n - (n - r) = r$. Thus, we have $\text{Span}\left\{\text{columns of } v_j\right\} = \text{null}(D_n^{(X_j, r)})$.

□

Lemma E.15. For $\mathcal{F}^{(r)}(V)$ defined in Definition (10), and for the $f_0 \in \mathcal{F}^{(r)}(V)$,

(a) For any $t > 0$, let a space of functions be defined as

$$\mathcal{D}(t^2) \coloneqq \left\{\delta = f - f_0 : \Delta(\delta) \leq t^2, f \in \mathcal{F}^{(r)}(V)\right\}, \tag{77}$$

We have

$$\mathcal{R}_n(\left\{\mathcal{F}^{(r)}(V) - f_0\right\} \cap \mathcal{D}(t^2)) \leq T_1 + T_2, \tag{78}$$

with T_1 and T_2 are given below

$$T_1 \coloneqq \mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) | \left\{X^i\right\}_{i=1}^n\right], T_2 \coloneqq \mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X) | \left\{X^i\right\}_{i=1}^n\right].$$

(b) For $K^{(X,r)}(V)$ in Definition (7) and for $\mathcal{D}(t^2) \coloneqq \left\{\delta \in \mathbb{R}^n, \Delta(\delta) \leq t^2\right\}$, we have for the same T_1 and T_2

$$\mathcal{R}_n(\left\{K^{(X,r)}(V) - \theta_0\right\} \cap \mathcal{D}(t^2)) \leq T_1 + T_2. \tag{79}$$

Proof. With the definitions and notations in Definition E.13.

For (a), we have

$$\begin{aligned}\sup_{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) \leq & \sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) \\ & + \sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X),\end{aligned} \tag{80}$$

taking expectation conditioned on $\left\{X^i\right\}_{i=1}^n$ on both sides, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) | \left\{X^i\right\}_{i=1}^n\right] \\ &\leq \underbrace{\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) | \left\{X^i\right\}_{i=1}^n\right]}_{T_1} + \underbrace{\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X) | \left\{X^i\right\}_{i=1}^n\right]}_{T_2}.\end{aligned} \tag{81}$$

For (b), We have

$$\begin{aligned}\sup_{\delta \in K^{(X,r)}(V) - \theta_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) \leq & \sup_{\delta \in B_{D_n^{(X,r)}}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) \\ & + \sup_{\delta \in B_{D_n^{(X,r)}}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X),\end{aligned} \tag{82}$$

taking expectation conditioned on $\left\{X^i\right\}_{i=1}^n$ on both sides, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) | \left\{X^i\right\}_{i=1}^n\right] \\ &\leq \underbrace{\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) | \left\{X^i\right\}_{i=1}^n\right]}_{T_1} + \underbrace{\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X) | \left\{X^i\right\}_{i=1}^n\right]}_{T_2}.\end{aligned} \tag{83}$$

Thus, the rest proof will proceed to bound each terms individually. □

Lemma E.16. Let $\delta \in \mathcal{F}^{(r)}(V)$. Then

$$\|\delta\|_n^2 \leq \max\left\{\|\delta\|_\infty, 1\right\} \Delta^2(\delta) n^{-1}. \tag{84}$$

Proof. We notice that

$$\begin{aligned}n\|\delta\|_n^2 &= \sum_{i:|\delta(X^i)| \leq 1} \left|\delta(X^i)\right|^2 + \sum_{i:|\delta(X^i)|>1} \left|\delta(X^i)\right|^2 \\ &\leq \sum_{i:|\delta(X^i)| \leq 1} \left|\delta(X^i)\right|^2 + \|\delta\|_\infty \sum_{i:|\delta(X^i)|>1} \left|\delta(X^i)\right| \\ &\leq \max\left\{\|\delta\|_\infty, 1\right\} \left(\sum_{i:|\delta(X^i)| \leq 1} \left|\delta(X^i)\right|^2 + \sum_{i:|\delta(X^i)|>1} \left|\delta(X^i)\right|\right) \\ &= \max\left\{\|\delta\|_\infty, 1\right\} \Delta^2(\delta).\end{aligned} \tag{85}$$

Thus we conclude that

$$\|\delta\|_n \leq \max\left\{\|\delta\|_\infty^{1/2}, 1\right\} \Delta(\delta) n^{-1/2}. \tag{86}$$

□

Lemma E.17. If $p_j \in P_j$ with $\|p_j\| = 1$, where p_j, P_j follow the definitions and notations in Definition E.13, then we have

$$\|p_j\|_\infty \leq \frac{c_3 \log n}{n^{1/2}}, \tag{87}$$

where c_3 depends on the degree r of P_j.

Proof. Let $\phi_\ell(\nu)$ be the ℓth degree orthogonal polynomials $\ell \leq (r-1)$ which have a domain in $[0,1]$ and satisfy

$$\int_0^1 \phi_\ell(\nu)\phi_{\ell'}(\nu)dt = \mathbb{1}\{l = l'\}, \tag{88}$$

then by Lemma E.14, p_j can be expressed as

$$p_{ij} = \sum_{l=0}^{r-1} a_l \phi_\ell(X_j^{(i)}), \tag{89}$$

for $i = 1, \ldots, n$ and $a_0, \ldots, a_{r-1} \in \mathbb{R}$.

Let $g_j : \mathbb{R} \to \mathbb{R}$ be defined as

$$g_j(\nu) = \sum_{l=0}^{r-1} a_l \phi_\ell(\nu), \tag{90}$$

which means $g_j(X_j^{(i)}) = p_{ij}$ for all $j = 1, \ldots d$.

If we split the interval $[0,1]$ into n pieces, and let $A_1 = [0, X_j^{(1)})$, $A_i = [X_j^{(i-1)}, X_j^{(i)})$ for all $i > 1$, and define

$$w = \max_{\substack{i=1,\ldots,n-1 \\ j=1,\ldots,d}} \left|X_j^{(i)} - X_j^{(i-1)}\right|. \tag{91}$$

then we have,

$$\begin{aligned}
&\left|\sum_{l=0}^{r-1} a_l^2 - \sum_{i=1}^{n} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)})\right| \\
&= \left|\int_0^1 g_j(u)^2 du - \sum_{i=1}^{n} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)})\right| \\
&= \left|\sum_{i=1}^{n} \int_{A_i} g_j(u)^2 du - \sum_{i=1}^{n} \int_{A_i} g_j(X_j^{(i)})^2 du\right| \\
&\leq \sum_{i=1}^{n} \int_{A_i} \left|g_j(u)^2 - g_j(X_j^{(i)})^2\right| du \\
&\leq \sum_{i=1}^{n} \int_{A_i} \left\|(g_j^2)'\right\|_\infty \left|u - X_j^{(i)}\right| du \\
&\leq \sum_{i=1}^{n} \left\|(g_j^2)'\right\|_\infty \frac{(u - X_j^{(i)})^2}{2} \Big|_{u=X_j^{(i)}}^{X_j^{(i-1)}} \\
&\lesssim \left\|(g_j^2)'\right\|_\infty \frac{\log^2 n}{n}.
\end{aligned} \tag{92}$$

The first equality follows from the definition of g_j in (90).

The last inequality follows since

$$w \lesssim \frac{c \log n}{n}, \tag{93}$$

by the Assumption D.1.

By applying this, we have

$$\left| X_j^{(i-1)} - X_j^{(i)} \right| \leq \frac{\log n}{n}. \tag{94}$$

At the same time, we have

$$\left(g(\nu)^2\right)' = \left(\sum_{l=0}^{r-1} a_l^2 \phi_\ell(\nu)^2 + \sum_{l \neq l'} a_l a_{l'} \phi_\ell(\nu) \phi_{\ell'}(\nu) \right)' \tag{95}$$

Thus, the above inequality is continued as

$$\begin{aligned} \left| \sum_{l=0}^{r-1} a_l^2 - \sum_{i=1}^{n} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)}) \right| &\lesssim \frac{\log^2 n}{n} \left\| (g_j^2)' \right\|_\infty \\ &\leq \frac{c_2 \log^2 n \, \|a\|_\infty^2}{n} \leq \frac{c_2 \log^2 n}{n} \sum_{l=0}^{r-1} a_l^2, \end{aligned} \tag{96}$$

for some constant $c_2 > 0$ that only depends on r. Thus for large enough n, we can have $\sum_{l=0}^{r-1} a_l^2 \leq \sum_{i=1}^{n} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)})$, or

$$\sum_{l=0}^{r-1} a_l^2 \lesssim \sum_{i=1}^{n} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)}) + \frac{c_2 \log^2 n}{n} \sum_{l=0}^{r-1} a_l^2,$$

which further implies

$$\sum_{l=0}^{r-1} a_l^2 \lesssim \frac{1}{1 - \frac{c_2 \log^2 n}{n}} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)}).$$

Thus, overall, we have

$$\begin{aligned} \sum_{l=0}^{r-1} a_l^2 &\lesssim \frac{1}{1 - \frac{c_2 \log^2 n}{n}} \sum_{i=1}^{n} g_j(X_j^{(i)})^2 (X_j^{(i)} - X_j^{(i-1)}) \\ &\leq \frac{1}{1 - \frac{c_2 \log^2 n}{n}} \sum_{i=1}^{n} g_j(X_j^{(i)})^2 \max_j \left| X_j^{(i)} - X_j^{(i-1)} \right| \\ &\lesssim \frac{1}{1 - \frac{c_2 \log^2 n}{n}} \frac{\log n}{n} \sum_{i=1}^{n} g_j(X_j^{(i)})^2 \\ &\leq \frac{1}{1 - \frac{c_2 \log^2 n}{n}} \frac{\log n}{n}. \end{aligned} \tag{97}$$

The last inequality holds by $\|p_j\|_2 = 1$. Furthermore, for large n, we have

$$\frac{1}{1 - \frac{c_2 \log^2 n}{n}} \leq c_2 \log n.$$

Thus, (97) can be further bounded as

$$\sum_{l=0}^{r-1} a_l^2 \lesssim \frac{c_2 \log^2 n}{n},$$

where we have absorbed the constants that depend on r and other constants into a single c_2. Therefore,

$$\begin{aligned}
\|p_j\|_\infty &= \max_{i=1,\dots,n} |p_{ij}| = \max_{i=1,\dots,n} \left| \sum_{l=0}^{r-1} a_l \phi_\ell(X_j^{(i)}) \right| \\
&\leq \max_{i=1,\dots,n} \sum_{l=0}^{r-1} \left| g_{j\ell}(X_j^{(i)}) \right| |a_l| \\
&\leq \|a\|_\infty \max_{x\in[0,1]} \sum_{l=0}^{r-1} |g_{j\ell}(x)| \\
&\leq \frac{c_2^{1/2}(\log n)}{n^{1/2}} \max_{x\in[0,1]} \sum_{l=0}^{r-1} |g_{j\ell}(x)| \\
&\leq \frac{c_3^{1/2}(\log n)}{n^{1/2}},
\end{aligned} \tag{98}$$

where the third inequality holds by using $\|a\|_\infty \leq \sqrt{\sum_{l=0}^{r-1} a_l^2}$ and the last inequality holds because of $\max_{x\in[0,1]} \sum_{l=0}^{r-1} |g_{j\ell}(x)| = O(1)$ then the claim follows. □

The following Lemma E.18, Lemma E.19 and Lemma E.20 will show that the $\|q_j\|_\infty \leq V$.

We reserve the letter S save for X_j. We will first define, for $r \geq 2$,

$$\Xi^{(r)} = \operatorname{diag}\left(S^{(r)} - S^{(1)}, \dots, S^{(n)} - S^{(n-r+1)} \right).$$

We then define the falling factorial basis matrix, $H \in \mathbb{R}^{n\times n}$, by

$$H_{ik} = h_k(S^{(i)}), \quad k, i = 1, \dots n, \tag{99}$$

here h_k denotes the falling factorial basis over S, as defined in Definition 2.2.

Lemma E.18 (Lemma 2 in Wang et al. (2014))**.** If $H^{(r-1)}$ is the $(r-1)$th order falling factorial basis matrix defined over the ordered inputs $S^{(1)} < \cdots < S^{(n)}$, and $D_n^{(S,r)}$ is the (r)th order discrete difference operator defined over the same inputs, then

$$(H^{(r-1)})^{-1} = \begin{bmatrix} C' \\ \frac{1}{(r-1)!} \cdot D_n^{(S,r)} \end{bmatrix}, \tag{100}$$

for an explicit matrix $C' \in \mathbb{R}^{r\times n}$. If we let A_i denote the ith row of a matrix A, and e_i be element of the canonical basis of subspace of $\mathbb{R}^n$, then C' has first row $C_1' = e_1^\top$, and subsequent rows

$$C_i' = \left[\frac{1}{(i-2)!} \cdot (\Xi^{(i)})^{-1} \cdot D_n^{(S,i-1)} \right], \quad i = 2, \dots r.$$

Then the last $n - r$ rows of $(H^{(r-1)})^{-1}$ are given by $D_n^{(S,r)}/(r-1)!$.

Lemma E.19 (Lemma 13 in Wang et al. (2015))**.** Let Π be the projection onto $\operatorname{row}(D_n^{(S,r)})$, the (r)th order discrete difference operator has pseudoinverse

$$(D_n^{(S,r)})^\ddagger = \Pi H_2^{(r-1)}/(r-1)!,$$

where $H_2^{(r-1)} \in \mathbb{R}^{n\times(n-r)}$ is the last $n-r$ columns of the $r-1$th order falling factorial basis matrix $H^{(r-1)}$.

Proof. The proof follows a same approach to Lemma 13 in Wang et al. (2015); for the reader's convenience, we provide it here again. We abbreviate $D = D_n^{(S,r)}$, and consider the linear system

$$DD^\top x = Db \tag{101}$$

in x, where $b \in \mathbb{R}^n$ is arbitrary. We seek an expression for $x = (DD^\top)^{-1}D^\top = (D^\dagger)^\top b$, and this will tell us the form of $D^\dagger$.

Define

$$\tilde{D} = \begin{bmatrix} C \\ D \end{bmatrix} \in \mathbb{R}^{n \times n},$$

where $C \in \mathbb{R}^{r \times n}$ is equal to $(r-1)!C'$ where C' is the matrix that collects the first row of each lower order difference operator, defined in Equation (100). From Lemma E.18, we know that

$$\tilde{D}^{-1} = H/(r-1)!,$$

where $H = H^{(r-1)}$ is falling factorial basis matrix of order $r-1$, evaluated over inputs $S^{(1)} < \cdots < S^{(n)}$. With this in mind, consider the expanded linear system

$$\begin{bmatrix} CC^\top & CD^\top \\ DC^\top & DD^\top \end{bmatrix} \begin{bmatrix} w \\ x \end{bmatrix} = \begin{bmatrix} a \\ Db \end{bmatrix}. \tag{102}$$

The second equation reads

$$DC^\top w + DD^\top x = Db,$$

and so if we can choose a in (102) so that at the solution we have $w = 0$, then x is the solution in (101). The first equation in (102) reads

$$CC^\top w + CD^\top x = a,$$

i.e.,

$$w = (CC^\top)^{-1}(a - CD^\top x).$$

That is, we want to choose

$$a = CD^\top x = CD^\top (DD^\top)^{-1} Db = C\Pi b,$$

where Π is the projection onto row space of D. Thus we can reexpress (102) as

$$\tilde{D}\tilde{D}^\top \begin{bmatrix} w \\ x \end{bmatrix} = \begin{bmatrix} C\Pi b \\ Db \end{bmatrix} = \tilde{D}\Pi b$$

and, using $\tilde{D}^{-1} = H/(r-1)!$,

$$\begin{bmatrix} w \\ x \end{bmatrix} = H^\top \Pi b/(r-1)!.$$

Finally, writing H_2 for the last $n-r$ columns of H, we have $x = H_2^\top \Pi b/(r-1)!$, as desired. □

Lemma E.20. Let $TV^{(r)}$ be defined in E.1. For $\delta \in \mathcal{F}^{(r)}(V)$, let $\delta = \sum_{j=1}^d \delta_j$, with $\delta_j \in B_{TV^{(r)}}(V)$ that is defined in Definition (22), with the definitions and notations in Definition E.13. Let the vector $\delta_j(X_j) \in \mathbb{R}^n$ be the evaluation of δ_j on the ordered inputs $X_j^{(1)}, \ldots, X_j^{(n)}$. Let $D_n^{(X_j,r)}$ be the discrete trend filtering operator defined in Definition (4), let $q_j = \Pi_{Q_j}\delta_j(X_j)$, where Q_j is row space of $D_n^{(X_j,r)}$, then we have

$$\|q_j\|_\infty \leq V \log n. \tag{103}$$

Proof. We know the Π_{Q_j} is the projection onto the row$(D_n^{(X_j,r)})$. Now let $M_j = (D_n^{(X_j,r)})^\ddagger \in \mathbb{R}^{n \times (n-r)}$, the pseudoinverse of the rth order discrete difference operator. From Lemma E.19, we know that

$$(D_n^{(X_j,r)})^\ddagger = \Pi_{Q_j} H_{2,j}^{(r-1)}/(r-1)!, \tag{104}$$

where $H_{2,j}^{(r-1)} \in \mathbb{R}^{n \times (n-r)}$ contains the last $n-r$ columns of the falling factorial basis matrix of order $(r-1)$, evaluated over $X_j^{(1)}, \ldots, X_j^{(n)}$, such that for $i \in \{1, \ldots, n\}$, and $s \in \{1, \ldots, n-r\}$, and $j \in \{1, \ldots, d\}$,

$$(H_{2,j}^{(r-1)})_{i,s} = h_{s,j}(X_j^{(i)}), \tag{105}$$

where

$$h_{s,j}(x) = \prod_{l=1}^{r-1}(x - x_j^{s+l})\mathbb{I}\{x \geq x_j^{s+l}\}, \tag{106}$$

where we reserve letter x_j for X_j. Then for e_i an element of the canonical basis of subspace of $\mathbb{R}^n$, and $P_j = \text{null}(D_n^{(X_j,r)})$, we have

$$\begin{aligned}\left\|e_i^\top M_j\right\|_\infty &\leq \left\|\Pi_{Q_j} e_i\right\|_1 \left\|H_{2,j}\right\|_\infty /(r-1)! \\ &\leq (\|e_i\|_1 + \left\|\Pi_{P_j} e_i\right\|_1) \left\|H_{2,j}\right\|_\infty /(r-1)! \\ &\leq (1 + \left\|\Pi_{P_j} e_i\right\|_1)/(r-1)!.\end{aligned} \tag{107}$$

The first inequality follows from Holder's inequality, the second from the triangle inequality and the last by the definition of $H_{2,j}^{(r-1)}$, with each entry is less equal to 1. Now, we let $\nu_1, \ldots, \nu_r$ be an orthonormal basis of P_j. Then we have

$$\left\|\Pi_{P_j} e_i\right\|_1 = \left\|\sum_{j=1}^{r}(e_i^\top \nu_j)\nu_j\right\|_1 \leq \sum_{j=1}^{r}\|\nu_j\|_\infty \|\nu_j\|_1 \leq \sum_{j=1}^{r}\|\nu_j\|_\infty n^{1/2}. \tag{108}$$

Based on Lemma E.17, we have obtain that

$$\|M_j\|_\infty = \max_{s=1\ldots n-r}\max_{i=1,\ldots,n}(M_j)_{i,s} = \max_{i=1,\ldots,n}\left\|e_i^\top M_j\right\|_\infty = O(\log n). \tag{109}$$

Then by the equivalence of Problems (6) and (11), for $\delta_j \in B_{TV^{(r)}}(V)$, we have

$$D^{(X_j,r)_n}\delta_j(X_j) \in B_{D_n^{(X,r)}}(V), \tag{110}$$

where $B_{D_n^{(X,r)}}(V)$ is defined in Definition (24). Then we get for any $\delta_j \in B_{TV^{(r)}}(V)$, we have

$$\left\|\Pi_{Q_j}\delta_j(X_j)\right\|_\infty = \left\|(D_n^{(X_j,r)})^\ddagger D_n^{(X_j,r)}\delta_j(X_j)\right\|_\infty \leq \|M_j\|_\infty \left\|D_n^{(X_j,r)}\delta_j(X_j)\right\|_1 \leq V\log n. \tag{111}$$

The last inequality follows from the (110) and (109). □

Lemma E.21. For $\delta \in \mathcal{F}^{(r)}(V)$, let $\delta = \sum_{j=1}^d \delta_j$, with $\delta_j \in B_{TV^{(r)}}(V)$ that is defined in Definition (E.1). Let $g_{j1}, \ldots, g_{jr}$ be the orthonormal basis for P_j such that $\|g_{j\ell}\| = 1$, for $l \in \{1, \ldots, r\}$, with the definitions and notations in Definition E.13. Denote $\alpha_{jl} = \left\langle \delta_j(X_j), g_{j\ell}\right\rangle_2$, and put all α_{jl} into a vector

$$\alpha = (\alpha_{11}, \ldots, \alpha_{1r}, \ldots, \alpha_{d1}, \ldots, \alpha_{dr}). \tag{112}$$

Then we have

$$\|\alpha\|_2 \leq \frac{\|p\|_2}{\lambda_{\min}(\Gamma^\top\Gamma)^{1/2}} = \frac{\|p\|_n}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}} \leq \frac{\max\left\{\|\delta\|_\infty^{1/2}, 1\right\}\Delta(\delta)n^{-1/2} + V\log n}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}, \tag{113}$$

where $\Gamma \in \mathbb{R}^{n\times rd}$ is a matrix constructed from the basis $g_{j1}, \ldots, g_{jr}$ for each $j \in \{1, \ldots, d\}$, such that the columns of Γ consist of $g_{j1}, \ldots, g_{jr}$ for all $j \in \{1, \ldots, d\}$.

Proof. Let $\text{col}(v_j)$ be the column space of v_j, where v_j is defined in Lemma E.17. Let $g_{j1}, \ldots, g_{jr}$ be the orthonormal basis for $\text{col}(v_j)$ by taking Gram-Schmidt Procedure, since $P_j = \text{col}(v_j)$, so $g_{j1}, \ldots, g_{jr}$ is a set of basis for P_j that

$$\|g_{j\ell}\| = 1, \quad l \in \{1, \ldots, r\}. \tag{114}$$

Recall

$$\alpha_{jl} = \left\langle \delta_j(X_j), g_{j\ell}\right\rangle_2, \tag{115}$$

and put all α_{jl} into a vector

$$\alpha = (\alpha_{11}, \ldots, \alpha_{1r}, \ldots, \alpha_{d1}, \ldots, \alpha_{dr}) \in \mathbb{R}^{rd}. \tag{116}$$

Then we have

$$p = \Pi_{P_1}\delta_1(X_j^{(1)}) + \cdots + \Pi_{P_d}\delta_d(X_d) = \sum_{j=1}^{d}\sum_{l=1}^{r}\left\langle \delta_j(X_j), g_{j\ell}\right\rangle_2 g_{j\ell} = \Gamma\alpha, \tag{117}$$

where $\Gamma \in \mathbb{R}^{n\times rd}$ the basis matrix constructed by basis $g_{j1}, \ldots, g_{jr}$, with $j = 1, \ldots, d$. We have

$$\|p\| = \|\Gamma\alpha\|, \tag{118}$$

and

$$\|p\| \geq \lambda_{\min}(\Gamma^\top\Gamma)^{1/2}\|\alpha\|, \tag{119}$$

thus, we have

$$\|\alpha\| \leq \frac{\|p\|}{\lambda_{\min}(\Gamma^\top\Gamma)^{1/2}} = \frac{\|p\|_n}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}. \tag{120}$$

By Lemma E.16, Lemma E.20, the triangle inequality, we have

$$\|p\|_n \leq \|q\|_n + \|\delta\|_n \leq \max\left\{\|\delta\|_\infty^{1/2}, 1\right\}\Delta(\delta)n^{-1/2} + V\log n \tag{121}$$

Thus, we have

$$\|\alpha\| \leq \frac{\max\left\{\|\delta\|_\infty^{1/2}, 1\right\}\Delta(\delta)n^{-1/2} + V\log n}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}. \tag{122}$$

□

Lemma E.22. Let $t > 0$ and for $\delta \in \mathcal{F}^{(r)}(V)$ with $\Delta^2(\delta) \leq t^2$ and $\delta_j \in B_{TV^{(r)}}(V)$. Then, with the definitions and notations in Definition E.13, it holds that

$$\|p_j\|_\infty \leq \gamma(t, d, n) \coloneqq \frac{c_3\sqrt{d}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}\left(\frac{t\log n}{n} + \frac{t^2\log n}{n} + \frac{V\log n}{n^{1/2}}\right), \tag{123}$$

where $c_3 > 0$ is a constant depends on r.

Proof. For $j = 1, \ldots, d$, let $g_{j1}, \ldots, g_{jr}$ be the orthonormal basis for P_j in Definition E.13 such that

$$\|g_{j\ell}\|_2 = 1, \quad l \in \{1, \ldots, r\}. \tag{124}$$

Then we have

$$p = \Pi_{P_1}\delta_1(X_j^{(1)}) + \ldots \Pi_{P_d}\delta_d(X_d) = \sum_{j=1}^{d}\sum_{l=1}^{r}\left\langle \delta_j(X_j), g_{j\ell}\right\rangle_2 g_{j\ell}. \tag{125}$$

Thus, for any j, we have for all i, if we express g_{jl} in component form as

$$g_{jl} = (g_{jl}^1, g_{jl}^2, \ldots, g_{jl}^n).$$

$$\begin{aligned}
|p_{ij}| &= \left|\left\langle \delta_j(X_j), g_{j1}\right\rangle_2 \cdot g_{j1}^i + \cdots + \left\langle \delta_j(X_j), g_{jr}\right\rangle_2 \cdot g_{jl}^i\right| \\
&\leq \sum_{l=1}^{r}|\alpha_{jl}|\left|g_{jl}^i\right| \leq \sum_{l=1}^{r}|\alpha_{jl}|\left(\max_l\|g_{jl}\|_\infty\right) \\
&\leq |\alpha|_1\frac{c_3\log n}{n^{1/2}} \leq \frac{c_3\sqrt{d}\log n}{n^{1/2}}\|\alpha\|_2 \\
&\leq \frac{c_3\sqrt{d}\log n\left(\max\left\{\|\delta\|_\infty^{1/2}, 1\right\}\Delta(\delta)n^{-1} + n^{-1/2}V\log n\right)}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}.
\end{aligned} \tag{126}$$

The first inequality is by Triangle Inequality, and α_{jl} is defined in Lemma E.21. The second inequality holds by the fact $X_j^i \in [0,1]$, and $g_{j1}, \ldots, g_{jr}$ are the orthonormal basis for P_j, then by Lemma E.17, $\|g_{j\ell}\|_\infty \leq \frac{c_3 \log n}{n^{1/2}}$. The last inequality follows from the Lemma E.21. Also, $c_3 > 0$ is a constant.

Furthermore, we have

$$\begin{aligned}
\|\delta\|_\infty &= \max\left\{ \max_{\{i:|\delta(X^i)|\geq 1\}} \left|\delta(X^i)\right|, \max_{\{i:|\delta(X^i)|<1\}} \left|\delta(X^i)\right| \right\} \\
&\leq \max\left\{ \sum_{i\in\{i:|\delta(X^i)|\geq 1\}} \left|\delta(X^i)\right|, 1 \right\} \\
&\leq \max\left\{ \sum_i \left|\delta(X^i)\right| \mathbb{I}\{\left|\delta(X^i)\right| \geq 1\} + \sum_i \left|\delta(X^i)\right|^2 \mathbb{I}\{\left|\delta(X^i)\right| \leq 1\}, 1 \right\} \\
&= \max\left\{\Delta^2(\delta), 1\right\} \leq \max\left\{t^2, 1\right\}.
\end{aligned} \tag{127}$$

Thus, we have $\|\delta\|_\infty^{1/2} \leq \max\{\Delta(\delta), 1\} \leq \max\{t, 1\}$.

Finally, we have

$$\begin{aligned}
\sum_{l=1}^r \left\langle \delta_j(X_j), g_{j\ell} \right\rangle_2 &\leq \sum_{l=1}^r |\alpha_{jl}| \\
&\leq \frac{c_3\sqrt{d}\left(\max\left\{\|\delta\|_\infty^{1/2}, 1\right\}\Delta(\delta) n^{-1/2} + V\log n\right)}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}} \\
&\leq \frac{c_3\sqrt{d}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}\left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V\log n\right),
\end{aligned} \tag{128}$$

where the second inequality follows from (122) and α has length rd; the last inequality follows from the condition $\Delta(\delta) \leq t$, and $\|\delta\|_\infty^{1/2} \leq \max\{t, 1\}$, and $\max\left\{t^2, t\right\} \leq t^2 + t$ for $t \geq 0$. Also from (126) we conclude that

$$\begin{aligned}
\|p_j\|_\infty &\leq \frac{c_3\sqrt{d}\log n\left(\max\left\{\|\delta\|_\infty^{1/2}, 1\right\}\Delta(\delta) n^{-1} + n^{-1/2} V\log n\right)}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}} \\
&\leq \frac{c_3\sqrt{d}\log n\left(\max\left\{\Delta^2(\delta), \Delta(\delta)\right\} n^{-1} + n^{-1/2} V\log n\right)}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}} \\
&\leq \frac{c_3\sqrt{d}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}\left(\frac{t\log n}{n} + \frac{t^2\log n}{n} + \frac{V\log n}{n^{1/2}}\right).
\end{aligned} \tag{129}$$

□

Lemma E.23 (Bounding The T_1)**.** Let $t > 0$ and for $\delta \in \mathcal{F}^{(r)}(V)$ with $\Delta^2(\delta) \leq t^2$ and $\delta_j \in B_{TV^{(r)}}(V)$. Consider the definitions and notations in Definition E.13. Let $g_{j1}, \ldots, g_{jr}$ be an orthonormal basis for P_j. It holds that

$$\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) | \left\{X^i\right\}_{i=1}^n\right] \leq \frac{c_3 d^{3/2}\sqrt{\log d}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top\Gamma)^{1/2}}\left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V\log n\right), \tag{130}$$

which goes to zero for large n, where c_3 is a constant depends on r and R, where Γ is defined in Lemma E.21.

Proof. By the Definition E.13 of Π_P,

$$\begin{aligned}
\xi^\top \Pi_P \delta(X) &= \xi^\top (p_1 + \cdots + p_d) \\
&\le \left| \sum_{l=1}^{r} \sum_{j=1}^{d} \xi^\top g_{j\ell} \delta_j(X_j)^\top g_{j\ell} \right| \le \sum_{l=1}^{r} \sum_{j=1}^{d} \left| \delta_j(X_j)^\top g_{j\ell} \right| \left| \xi^\top g_{j\ell} \right| \\
&\le d \left[\left(\max_{l=1,\ldots,r, j=1,\ldots,d} \left| \xi^\top g_{jl} \right| \right) \left(\max_{j=1,\ldots,d} \sum_{l=1}^{r} \left| \delta_j(X_j)^\top g_{j\ell} \right| \right) \right] \\
&\le c_3 d^{3/2} \left(\max_{l=1,\ldots,r, j=1,\ldots,d} \left| \xi^\top g_{jl} \right| \right) \left(\frac{1}{\lambda_{\min}(\frac{1}{n}\Gamma^\top \Gamma)^{1/2}} \left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V \log n \right) \right),
\end{aligned} \tag{131}$$

where the first inequality is based on (125) and last inequality is based on (128).

Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) \middle| \left\{ X^i \right\}_{i=1}^n \right] \\
&\le \frac{c_3 d^{3/2}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top \Gamma)^{1/2}} \left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V \log n \right) \left(\sum_{l=1,\ldots,r} \mathbb{E}\left[\max_{j=1,\ldots,d} \left| \xi^\top g_{jl} \right| \middle| \left\{ X^i \right\}_{i=1}^n \right] \right) \\
&\le \frac{c_3 d^{3/2} \sqrt{\log d}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top \Gamma)^{1/2}} \left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V \log n \right).
\end{aligned} \tag{132}$$

The first inequality follows from line (131) and

$$\max_{l=1,\ldots,r, j=1,\ldots,d} \left| \xi^\top g_{jl} \right| \le \sum_{l=1,\ldots,r} \max_{j=1,\ldots,d} \left| \xi^\top g_{jl} \right|.$$

The last inequality holds since $\xi^\top g_{jl}$ are sub-Gaussian random variables with parameter 1, and then by applying (2.66) from Wainwright (2019). Thus we have

$$\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) \middle| \left\{ X^i \right\}_{i=1}^n \right] \le \frac{c_3 d^{3/2} \sqrt{\log d}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top \Gamma)^{1/2}} \left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V \log n \right). \tag{133}$$

□

Lemma E.24. Let $t > 0$ and for $\delta \in \mathcal{F}^{(r)}(V)$ with $\Delta^2(\delta) \le t^2$ and $\delta_j \in B_{TV^{(r)}}(V)$. For q defined in Definition E.13 **6** where $\Pi_Q \delta(X) = q$, define $\Delta_q^2 \coloneqq \sum_{i=1}^n \min\left\{ \left| q^i \right|, (q^i)^2 \right\}$, where q^i is defined in Definition E.13 **6**. For $j = 1, \ldots, d$, let $g_{j1}, \ldots, g_{jr}$ be the orthonormal basis for P_j, with the definitions and notations in Definition E.13. Then, it holds that

$$\Delta_q^2 \le h(t, d, n) \coloneqq \left(2t^2 + 2t^2 \gamma(t, d, n) + 4nc^2(r, \lambda_{\min}) d^3 \left(\frac{t^2 \log^2 n}{n^2} + \frac{t^4 \log^2 n}{n^2} + \frac{V^2 \log^2 n}{n} \right) \right), \tag{134}$$

with $\gamma(t, d, n) \coloneqq \frac{c_3 d^{3/2}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top \Gamma)^{1/2}} \left(\frac{t \log n}{n} + \frac{t^2 \log n}{n} + \frac{V \log n}{n^{1/2}} \right)$ as the same quantity in Lemma E.22.

Proof. Based on Lemma E.22, we have

$$\|p\|_\infty = \sum_{j=1}^{d} \|p_j\|_\infty \le \gamma(t, d, n) \coloneqq \frac{c_3 d^{3/2}}{\lambda_{\min}(\frac{1}{n}\Gamma^\top \Gamma)^{1/2}} \left(\frac{t \log n}{n} + \frac{t^2 \log n}{n} + \frac{V \log n}{n^{1/2}} \right). \tag{135}$$

Based on triangle inequality, we have

$$\|q - \delta\|_\infty = \|p\|_\infty \le \gamma(t, d, n). \tag{136}$$

Also we find that

$$\begin{aligned}\Delta_q^2 &= \sum_{i=1}^n \min\left\{\left|q^i\right|, (q^i)^2\right\} \\ &\le \sum_{i=1}^n |q_i|\, \mathbb{I}\{\left|\delta(X^i)\right| > 1\} + \sum_{i=1}^n q_i^2 \mathbb{I}\{\left|\delta(X^i)\right| \le 1\}.\end{aligned} \tag{137}$$

Then we have

$$\begin{aligned}\Delta_q^2 &\le \sum_{i=1}^n (\left|\delta(X^i)\right| + \gamma(t,d,n))\mathbb{I}\{\left|\delta(X^i)\right| > 1\} + \sum_{i=1}^n (2\delta(X^i)^2 + 2(\gamma(t,d,n)^2))\mathbb{I}\{\left|\delta(X^i)\right| \le 1\} \\ &\le t^2(\gamma(t,d,n)) + \sum_{i=1}^n (\left|\delta(X^i)\right|)\mathbb{I}\{\left|\delta(X^i)\right| > 1\} + 2\sum_{i=1}^n \delta_i^2 \mathbb{I}\{\left|\delta(X^i)\right| \le 1\} + 2\sum_{i=1}^n (\gamma(t,d,n)^2)\mathbb{I}\{\left|\delta(X^i)\right| \le 1\} \\ &\le 2t^2 + 2t^2\gamma(t,d,n) + 2n\gamma(t,d,n)^2 \\ &\le \left(2t^2 + 2t^2\gamma(t,d,n) + 6nc^2(r,\lambda_{\min})d^3\left(\frac{t^2\log^2 n}{n^2} + \frac{t^4\log^2 n}{n^2} + \frac{V^2\log^2 n}{n}\right)\right).\end{aligned} \tag{138}$$

Where the first inequality follows since $\|q-\delta\|_\infty \le \gamma(t,d,n)$ and $a^2+b^2 \ge 2ab$, where the second inequality follows from the fact that

$$\left|\left\{i \in \{1,\dots,n\} : \left|\delta(X^i)\right| > 1\right\}\right| \le t^2, \tag{139}$$

which holds because $\Delta^2(\delta) \le t^2$. The third inequality follows by the definition of $\Delta^2(\delta) \coloneqq \sum_{i=1}^n \min\left\{\left|\delta(X^i)\right|, \delta(X^i)^2\right\} = \sum_{i=1}^n (\left|\delta(X^i)\right| \mathbb{I}\{\left|\delta(X^i)\right| > 1\}) + \sum_{i=1}^n (\delta(X^i)^2 \mathbb{I}\{\left|\delta(X^i)\right| \le 1\})$. □

Lemma E.25 (Bounding T_2)**.** Let $t \asymp n^{-r/(2r+1)}V^{1/(2r+1)}\max\left\{1, V^{(2r-1)/(4r+2)}\right\}$, for $\delta \in \mathcal{F}^{(r)}(V)$ with $\Delta^2(\delta) \le t^2$, and with the definitions and notations in Definition E.13. For a positive constant c_4 , it holds that

$$\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)} \frac{1}{n}\xi^\top \Pi_Q \delta(X) | \left\{X^i\right\}_{i=1}^n\right] \le c_4 (d)^{1/2+1/2r} n^{-1/2} V^{1/2r} m^{1-1/(2r)}, \tag{140}$$

with m is a quantity depends on $t, n, d, r, \lambda_{\min}$.

Proof. For q defined in Definition E.13 **6** where $\Pi_Q\delta(X) = q$, define $\Delta_q^2 \coloneqq \sum_{i=1}^n \min\left\{\left|q^i\right|, (q^i)^2\right\}$, where q^i is defined in Definition E.13 **6**. First, by Lemma E.16, we have

$$\begin{aligned}\|q\|_n &\le \max\left\{\|q\|_\infty^{1/2}, 1\right\}\Delta_q n^{-1/2} \\ &\le \max\left\{d^{1/2}\max_{j=1\dots d}\|q_j\|_\infty^{1/2}, 1\right\}\Delta_q n^{-1/2} \\ &\le c(r,\lambda_{\min},d)\max\left\{V^{1/2},1\right\}\left(t + d^{3/4}\left(\frac{V^{1/2}\sqrt{\log n}}{n^{1/4}} + \frac{t^{1/2}\sqrt{\log n}}{n^{1/2}} + \frac{t\log n}{n^{1/2}}\right)t\right. \\ &\qquad \left. + d^{3/2}\left(\frac{t\log n}{n^{1/2}} + \frac{t^2\log n}{n^{1/2}} + V\log n\right)\right)n^{-1/2} \\ &\le c(r,\lambda_{\min},d)\max\left\{V^{1/2},1\right\}\left(\left(1 + \frac{t^{1/2}\sqrt{\log n}}{n^{1/2}} + \frac{t\log n}{n^{1/2}}\right)t + \frac{t^2\log n}{n^{1/2}} + V\log n\right)n^{-1/2} \\ &=: m,\end{aligned} \tag{141}$$

where the second line we use triangle inequality, where in the third line we used Lemma E.20 to control $\|q_j\|_\infty$, Lemma E.24 to control Δ_q, and $\sqrt{a_1^2+\cdots+a_n^2} \le a_1+\cdots+a_n$ for positive numbers $a_1,\dots,a_n$, and $c(r,\lambda_{\min},d)$ is introduced that depends on $d, r, \lambda_{\min}$, where in the fourth line we assume $t \ge 1$ and $t \asymp n^{-r/(2r+1)}V^{1/(2r+1)}\max\left\{1, V^{(2r-1)/(4r+2)}\right\}$ and combine some lower order terms.

Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)|\left\{X^i\right\}_{i=1}^n\right]\\
&\leq\mathbb{E}\left[\sup_{g\in\mathcal{F}^{(r)}(2V)\cap B_n(m)}\frac{1}{n}\sum_{i=1}^n\xi^i g(X^i)|\left\{X^i\right\}_{i=1}^n\right]\\
&\leq c_{\mathrm{Dud}}\frac{1}{\sqrt{n}}\int_0^m\sqrt{\log N(\epsilon,\|\cdot\|_n,\mathcal{F}^{(r)}(2V))}d\epsilon\\
&\leq c_4 n^{-1/2}(d)^{1/2+1/2r}V^{1/2r}\int_0^m\epsilon^{-1/2r}d\epsilon\\
&=c_4 n^{-1/2}(d)^{1/2+1/2r}V^{1/2r}m^{1-1/(2r)}.
\end{aligned}
\tag{142}
$$

The second line follows from (141), the third line applies Dudley's entropy integral Dudley (1967), with c_{Dud} a positive constant, and the fourth line follows from Lemma 15 in Sadhanala and Tibshirani (2019), specifically from the middle derivations and the third-to-last displayed equation on page 45, where c_4 is a positive constant.

□

F PROOF OF THEOREM 3.1

Proof. For $t>1$, by Theorem E.11,

$$
\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\Delta^2(\widehat{f}-f_0)>t^2\mid\left\{X^i\right\}_{i=1}^n\right)&\leq\frac{c}{t^2}\mathcal{R}_n\left(\left\{f-f_0:f\in\mathcal{F}^{(r)}(V)\right\}\cap\left\{f-f_0:\Delta^2(f-f_0)\leq nt^2\right\}\right)\\
&=\frac{c}{t^2}\,\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(V)-f_0:\Delta^2(\delta)\leq nt^2}\frac{1}{n}\xi^\top\delta(X)|\left\{X^i\right\}_{i=1}^n\right],
\end{aligned}
\tag{143}
$$

where c is a positive constant that depends on L and $\underline{f}$ in Assumption D.4. We denote the numerator of the right-hand side as $T_n(t)$ that depends on t and n. By (78) in Lemma E.15, we decompose the $T_n(t)$ into two terms. Equation (130) in Lemma E.23 gives the bound for the first term, and Equation (142) in Lemma E.25 gives the bound for the second term. Then we have the upper bound of $T_n(t)$ as

$$
\begin{aligned}
T_n(t)&\leq cc_4 n^{-1/2}(d)^{1/2+1/2r}V^{1/2r}m_1(t)^{1-1/(2r)}\\
&\leq C_1 n^{-r/(2r+1)}V^{1/(2r+1)}m_1(t)+C_1 n^{-2r/(2r+1)}V^{2/(2r+1)},
\end{aligned}
\tag{144}
$$

where in the first inequality, we used (142), which is of higher order compared to (130). The second inequality follows from Lemma F.1, where we set $a=n^{-1/2}d^{1/2+1/2r}V^{1/2r}$, $b=m_1(t)$, and $w=1/r$. Here, C_1 is a constant that depends on d, c and c_1, and $m_1(t)$ depends on $t,n,d,r,\lambda_{\min}$, as derived in (141). Specifically,

$$
m_1(t)=c(r,\lambda_{\min},d)\max\left\{V^{1/2},1\right\}\left(\left(1+t^{1/2}\sqrt{\log n}+t\log n\right)t+t^2\log^2 n+\log nVn^{-1/2}\right),
\tag{145}
$$

where $c(r,\lambda_{\min},d)$ depends on $r,\lambda_{\min},d$.

Define the function

$$
g_n(t)=\frac{T_n(t)}{t^2},
\tag{146}
$$

setting

$$
t\asymp n^{-r/(2r+1)}V^{1/(2r+1)}\max\left\{1,V^{(2r-1)/(4r+2)}\right\},
\tag{147}
$$

we obtain the result of Theorem 3.1, by verifying that

$$\begin{aligned}
\lim_{c_1\to\infty}\sup_{n\geq 1} g_n(c_1 t) &= \lim_{c_1\to\infty}\sup_{n\geq 1}\frac{T_n(c_1 t)}{c_1^2 t^2} \\
&\leq \lim_{c_1\to\infty}\sup_{n\geq 1}\frac{C_1 n^{-r/(2r+1)} V^{1/(2r+1)} m_1(c_1 t)}{c_1^2 t^2} \\
&\quad + \lim_{c_1\to\infty}\sup_{n\geq 1}\frac{C_1 n^{-2r/(2r+1)} V^{2/(2r+1)}}{c_1^2 t^2} \\
&\leq \lim_{c_1\to\infty}\sup_{n\geq 1} c(r,\lambda_{\min},d)\Big(c_1^{-1} + t^{1/2}c_1^{-1}(\log n)^{1/2} + tc_1^{-1}\log n \\
&\quad + tc_1^{-1}(\log n)^2 + c_1^{-1} V n^{-1/2} t^{-1}\Big) \\
&= 0,
\end{aligned} \tag{148}$$

with such a choice of t. □

Lemma F.1 (Lemma 11 in Sadhanala and Tibshirani (2019))**.** For any $a, b \geq 0$, and any w,

$$ab^{1-w/2} \leq a^{1/(1+w/2)}b + a^{2/(1+w/2)}.$$

Proof. Note that either $ab^{1-w/2} \leq a^{1/(1+w/2)}b$ or $ab^{1-w/2} \geq a^{1/(1+w/2)}b$, and in the latter case we get $b \leq a^{1/(1+w/2)}$, so $ab^{1-w/2} \leq a^{2/(1+w/2)}$. □

G DISCUSSION OF INFLUENCE OF τ FOR THEOREM 3.1

Note that the influence of τ on the $O_{pr}\left(n^{-2r/(2r+1)}V^{2/(2r+1)}\max\left\{1, V^{(2r-1)/(2r+1)}\right\}\right)$ rate is solely through a constant $\tilde{c}$ hiden in the symbbol O_{pr}. The constant $\tilde{c}$ satisfies $\tilde{c} = O((L\underline{f})^{-1})$, where L and $\underline{f}$ are defined as in Assumption D.4, with $F_{Y^i|X^i}(f_0^i) = \tau$. Please also refer to Lemma E.6, where c is given by $\tilde{c}^{-1}$. Hence, in principle, we could let $\tau \to 0$ or $\tau \to 1$, but in that case, the rate in (15) would have to be inflated by the factor of $(L\underline{f})^{-1}$.

H ASSUMPTION FOR THEOREM 3.3

Furthermore, to produce an error rate linearly dependent on the growing dimension d, we include the assumption below. This appeared as Assumption A3 in Sadhanala and Tibshirani (2019).

Assumption H.1. The input points $X^i, i = 1, \ldots, n$ are i.i.d. from a continuous distribution $\mathscr{F}$ supported on $[0,1]^d$, that decomposes as $\mathscr{F} = \mathscr{F}_1 \times \cdots \times \mathscr{F}_d$, where the density of each $\mathscr{F}_j$ is lower and upper bounded by positive constants b_1 and b_2, respectively, for $j = 1, \ldots, d$.

Remark H.2. Assumption H.1 appeared as Assumption A.3 in Sadhanala and Tibshirani (2019). As mentioned in Sadhanala and Tibshirani (2019), this condition is fairly restrictive, since it requires the input distribution $\mathscr{F}$ to have independent coordinates. The reason we use this assumption: when $\mathscr{F} = \mathscr{F}_1 \times \cdots \times \mathscr{F}_d$, additive functions enjoy a key decomposability property in terms of the (squared) L_2 norm defined with respect to $\mathscr{F}$. In particular, if $f = \sum_{j=1}^d f_j$ has components with L_2 mean zero, denoted by

$$\bar{f}_j = \int_0^1 f_j(x_j) d\mathscr{F}_j(x_j) = 0, \quad j = 1, \ldots, d,$$

then we have

$$\left\| \sum_{j=1}^d f_j \right\|_{L_2}^2 = \sum_{j=1}^d \|f_j\|_{L_2}^2. \tag{149}$$

This is explained by the fact that each pair of components f_j, f_l with $j \neq l$ are orthogonal with respect to the

L_2 inner product, since

$$\langle f_j, f_l \rangle_{L_2} = \int_{[0,1]^2} f_j(x_j) f_l(x_l) d\mathscr{F}_j(x_j) d\mathscr{F}_l(x_l) = \bar{f}_j \bar{f}_l = 0.$$

The above orthogonality, and thus the decomposability property in (149), is only true because of the product form $\mathscr{F} = \mathscr{F}_1 \times \cdots \times \mathscr{F}_d$. Such decomposability is not generally possible with the empirical norm (the inner products between components do not vanish even if all empirical means are zero). In the proof of Theorem 3.3, we move from considering the empirical norm of the error vector to the L_2 norm, in order to leverage the property in (149), which eventually leads to an error rate that has a polynomial dependence on the dimension d.

I SUPPLEMENTARY LEMMAS FOR THEOREM 3.3

Lemma I.1. Let $f : [0,1]^d \to \mathbb{R}$ with $f = \sum_{j=1}^d f_j$. Let Assumption H.1 holds, furthermore, if

$$\int_0^1 f_j(x_j) d\mathscr{F}_j(x_j) = 0, \tag{150}$$

for $j = 1, \ldots, d$, then it holds that

$$\langle f, f \rangle_{L_2} = \sum_{j=1}^d \langle f_j, f_j \rangle_{L_2}. \tag{151}$$

Proof. By the definition of L_2 inner product, and Assumption H.1,we have

$$\begin{aligned}
\langle f, f \rangle_{L_2} &= \int_{[0,1]^d} f(x)^2 d\mathscr{F}(x) \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^d f_j(x_j) \right)^2 d\mathscr{F}_1(X^1) \times \cdots \times d\mathscr{F}_d(x_j) \\
&= \sum_{j=1}^d \int_{[0,1]} f_j(x_j)^2 d\mathscr{F}_j(x_j) + 2 \sum_{j \neq k} \int_{[0,1]^2} f_j(x_j) f_k(x_k) d\mathscr{F}_j(x_j) d\mathscr{F}_k(x_k) \\
&= \sum_{j=1}^d \int_{[0,1]} f_j(x_j)^2 d\mathscr{F}_j(x_j) + 2 \sum_{j \neq k} \int_{[0,1]} f_j(x_j) d\mathscr{F}_j(x_j) \int_{[0,1]} f_k(x_k) d\mathscr{F}_k(x_k) \\
&= \sum_{j=1}^d \langle f_j, f_j \rangle_{L_2}.
\end{aligned} \tag{152}$$

Thus we conclude that $\langle f_j, f_k \rangle_2 = 0$ for $j \neq k$ and

$$\left\| \sum_{j=1}^d f_j \right\|_{L_2}^2 = \sum_{j=1}^d \|f_j\|_{L_2}^2 . \tag{153}$$

□

Lemma I.2 (Bounding the T_1 with growing d)**.** For $\delta \in \mathcal{F}^{(r)}(V)$, and let Π_P follow the definitions and notations in Definition E.13. it holds that

$$\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) | \left\{ X^i \right\}_{i=1}^n \right] \leq \frac{c_3 d^{3/2} \sqrt{\log d}}{\lambda_{\min}(\frac{1}{n} \Gamma^\top \Gamma)^{1/2}} \left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V \log n \right), \tag{154}$$

which goes to zero for large n.

Proof. The proof is similar to the proof of Lemma E.23, thus we omit that. □

Lemma I.3. Let $t_n = c\sqrt{d}n^{-r/(2r+1)}V^{1/(2r+1)}$, where c is a positive constant. This value is chosen such that, if we set $m = t_n$, then

$$\frac{\mathcal{R}_n(B^d_{TV^{(r)}}(2V) \cap B_{L_2}(m))}{m} \leq \frac{m}{c} \tag{155}$$

holds with high probability.

Proof. Let β be a d dimensional new variable satisfying $\|\beta\|_2 \leq m$. We then use it to control the radius of the L_2 ball. Let first use the decomposability property in L_2 norm by Assumption H.1, which is

$$\left\| \sum_{j=1}^d f_j \right\|^2_{L_2} = \sum_{j=1}^d \|f_j\|^2_{L_2}. \tag{156}$$

We let m to denote the upper bound such that

$$\left\| \sum_{j=1}^d f_j \right\|^2_{L_2} \leq m^2 \implies \sum_{j=1}^d \|f_j\|^2_{L_2} \leq m^2. \tag{157}$$

Then we have two equivalent function spaces:

$$\left\{ f_j : \sum_{j=1}^d \|f_j\|^2_{L_2} \leq m^2 \right\} = \left\{ f_j : \|f_j\|_{L_2} \leq |\beta_j|, \|\beta\|_2 \leq m \right\}. \tag{158}$$

To get the local critical radius of $B^d_{TV^{(r)}}(2V)$, by L_2 orthogonality of the components of functions in $B^d_{TV^{(r)}}(2V)$, we first have

$$\begin{aligned} \sup_{g \in B^d_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n} \sum_{i=1}^n \xi^i g(X^i) &\leq \sup_{\|\beta\|_2 \leq m} \sup_{\substack{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|) \\ j=1,\ldots,d}} \frac{1}{n} \sum_{i=1}^n \xi^i \sum_{j=1}^d g_j(X^i_j) \\ &\leq \sup_{\|\beta\|_2 \leq m} \sum_{j=1}^d \sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)} \frac{1}{n} \sum_{i=1}^n \xi^i g_j(X^i_j). \end{aligned} \tag{159}$$

The first inequality follows from (158).

Then we have with probability at least $1 - 1/n^2$, it holds that

$$\begin{aligned} &\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)} \frac{1}{n} \sum_{i=1}^n \xi^i g_j(X^i_j) \\ &\leq c\mathcal{R}_n(B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)) + c\sqrt{\frac{\log n}{n}} \left(\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)} \|g_j\|_n \right) \\ &\leq c \left(\mathcal{R}(B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)) + \frac{\log n}{n} + \sqrt{\frac{\log n}{n}} \left(\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)} \|g_j\|_n \right) \right) \\ &\leq c \left(\mathcal{R}(B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)) + \frac{\log n}{n} + \sqrt{\frac{\log n}{n}} \sqrt{2} \left(\max\left\{ |\beta_j|, r_{nj} \right\} \right) \right) \\ &\leq c \frac{V^{1/2r} |\beta_j|^{1-2/r}}{\sqrt{n}} + c\sqrt{\frac{\log n}{n}} \left(\max\left\{ |\beta_j|, r_{nj} \right\} \right). \end{aligned} \tag{160}$$

The first line holds by Theorem 3.6 in Wainwright (2019), where $\mathcal{R}_n$ is defined in (E.2), the second line holds by Lemma A.4 in Bartlett et al. (2005), and third inequality holds by Lemma 3.6 in Bartlett et al. (2005), where r_{nj} is the critical radius of the function class $B_{TV^{(r)}}(2V)$, the smallest b such that

$$\frac{\mathcal{R}(B_{TV^{(r)}}(2V) \cap B_{L_2}(b))}{b} \leq \frac{b}{c}. \tag{161}$$

By Lemma I.5, we have $r_{nj} = n^{-r/(2r+1)} V^{1/(2r+1)}$, and the last inequality in (160) follows by Lemma I.5, also uses $\frac{\log n}{n} \leq r_{nj}\sqrt{\frac{\log n}{n}}$ for n sufficiently large. Call an event based on the result of (160) that simultaneously holds for all $j = 1, \ldots, d$

$$\mathcal{E} = \left\{ \sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(|\beta_j|)} \frac{1}{n} \sum_{i=1}^{n} \xi^i g_j(X_j^i) \leq c \frac{V^{1/2r} |\beta_j|^{1-2/r}}{\sqrt{n}} + c\sqrt{\frac{\log n}{n}} \left(\max\left\{|\beta_j|, r_{nj}\right\}\right) \text{ for } j = 1, \ldots, d \right\}. \tag{162}$$

By a union bound, $\mathbb{P}(\mathcal{E}) \geq 1 - d/n^2$.

Meanwhile, on $\mathcal{E}^c$, by Lemma E.20, we have $\left\|g_j(X_j^i)\right\|_\infty \leq V \log n$.

Thus, back to (159), we have

$$\begin{aligned} \sup_{g \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n} \sum_{i=1}^{n} \xi^i g(X^i) &\leq \sup_{\|\beta\|_2 \leq m} \sum_{j=1}^{d} \left(c \frac{V^{1/2r} |\beta_j|^{1-1/2r}}{\sqrt{n}} + \sqrt{\frac{\log n}{n}} \left(\max |\beta_j|, r_{nj}\right) \right) + \frac{dV \log n}{n^2} \\ &\leq c \left(\frac{V^{1/2r} d^{(2r+1)/(4r)} m^{1-1/2r}}{\sqrt{n}} + \sqrt{\frac{d \log n}{n}} m + d r_{nj}^2 \right) + \frac{dV \log n}{n^2}. \end{aligned} \tag{163}$$

In the second line, we use Holder's inequality $a^\top b \leq \|a\|_p \|b\|_q$ for the first term, with $p = 4/(2 + 1/r)$, and $q = 4/(2 - 1/r)$; for the second term, we use $\max\{a, b\} \leq a + b$ for $a > 0$ and $b > 0$. We also use the fact in (158) to get the bound $\|\beta\|_1 \leq \sqrt{d} m$, and the fact that $r_{nj}\sqrt{\log n / n} \leq r_{nj}^2$ for large enough n. Thus, we have

$$\begin{aligned} \mathcal{R}_n(B_{TV^{(r)}}(2V) \cap B_{L_2}(m)) &\leq \mathbb{E}\left[\sup_{g \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n} \sum_{i=1}^{n} \xi^i g(X^i) \middle| \left\{X^i\right\}_{i=1}^n \right] \\ &\leq c \left(\frac{V^{1/2r} d^{(2r+1)/(4r)} m^{1-1/2r}}{\sqrt{n}} + \sqrt{\frac{d \log n}{n}} m + d r_{nj}^2 \right) + \frac{dV \log n}{n^2} \\ &\lesssim c \left(\frac{V^{1/2r} d^{(2r+1)/(4r)} m^{1-1/2r}}{\sqrt{n}} + \sqrt{\frac{d \log n}{n}} m + d r_{nj}^2 \right), \end{aligned} \tag{164}$$

the third line follows by seeing the $\frac{dV \log n}{n^2}$ is a lower order term given $V = \Omega(1)$. Denoting the right-hand side of first line of (164) as T_r, we have

$$\begin{aligned} T_r &\leq c \left(n^{-1/2} (d)^{1/2 + 1/4r} V^{1/2r} m^{1-1/(2r)} + \sqrt{\frac{d \log n}{n}} m + d r_{nj}^2 \right) \\ &\leq c n^{-r/(2r+1)} d^{1/2} V^{1/(2r+1)} m + \sqrt{\frac{d \log n}{n}} m + c d^{(2+2/r)/(2+1/r)} n^{-2r/(2r+1)} V^{2/(2r+1)}, \end{aligned} \tag{165}$$

where in the second inequality, we have used the similar bound in Equation (144), it can be verified that for $m = c\sqrt{d} n^{-r/(2r+1)} V^{1/(2r+1)}$, the upper bound in (165) is at most m^2/c. Therefore, this is an upper bound on the critical radius of $B_{TV^{(r)}}(2V)$, which completes the proof. □

Lemma I.4 (Bounding T_2 with growing d)**.** Let

$$t \asymp d^{(r+1)/(2r+1)} n^{-r/(2r+1)} V^{1/(2r+1)} \max\left\{1, V^{(2r-1)/(4r+2)}\right\} \log n,$$

and for $\delta \in \mathcal{F}^{(r)}(V)$ with $\Delta^2(\delta) \le t^2$, and let p_j, q_j, p, q, P_j, Q_j, Π_{P_j}, Π_{Q_j}, Π_P, and Π_Q follow the definitions and notations in Definition E.13. For a positive constant c, it holds that

$$\begin{aligned}&\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)|\left\{X^i\right\}_{i=1}^n\right]\\&\le cd^{(2r+2)/(2r+1)}n^{-2r/(2r+1)}V^{2/(2r+1)}\max\left\{1,V^{(2r-1)/(2r+1)}\right\}\log n.\end{aligned}\tag{166}$$

Proof. Let t_n be defined as a value of t satisfying

$$\frac{\mathcal{R}_n(B^d_{TV^{(r)}}(2V)\cap B_{L_2}(t))}{t}\le\frac{t}{c},\tag{167}$$

where $B^d_{TV^{(r)}}$ is given in (23). We write l as

$$\sqrt{2}m\vee\left(t_n+\frac{\log n}{n}\right),\tag{168}$$

where m is defined in Equation (141). Let d dimensional β be a new variable satisfying $\|\beta\|_2\le l$. We then use it to control the radius of the L_2 ball and we aim to have control over it. Then we have two equivalent function spaces:

$$\left\{f_j:\sum_{j=1}^d\|f_j\|_{L_2}^2\le l^2\right\}=\left\{f_j:\|f_j\|_{L_2}\le|\beta_j|,\|\beta\|_2\le l\right\}.\tag{169}$$

Define event $\mathcal{E}_1\coloneqq\left\{B_n(m)\subseteq B_{L_2}(\sqrt{2}m\bigvee(t_n+\frac{\log n}{n}))\right\}$, and $\mathcal{E}_1^c$ is its complement.

First, we can upper bound $\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)$, by splitting into two parts. Therefore, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)|\left\{X^i\right\}_{i=1}^n\right]\\&\le\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)\mathbb{I}\{\mathcal{E}_1\}|\left\{X^i\right\}_{i=1}^n\right]+\mathbb{E}\left[\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)\mathbb{I}\{\mathcal{E}_1^c\}|\left\{X^i\right\}_{i=1}^n\right]\\&\le\mathbb{E}\left[\sup_{g\in\mathcal{F}^{(r)}(2V)\cap B_n(m)}\frac{1}{n}\sum_{i=1}^n\xi^i g(X^i)\mathbb{I}\{\mathcal{E}_1\}|\left\{X^i\right\}_{i=1}^n\right]+\frac{c_1dV\log n}{n}\\&\le\mathbb{E}\left[\sup_{g\in\mathcal{F}^{(r)}(2V)\cap B_{L_2}(\sqrt{2}m\bigvee(t_n+\frac{\log n}{n}))}\frac{1}{n}\sum_{i=1}^n\xi^i g(X^i)|\left\{X^i\right\}_{i=1}^n\right]+\frac{c_1dV\log n}{n}\\&\le\mathbb{E}\left[\sup_{\|\beta\|_2\le l}\sup_{\substack{g_j\in B_{TV^{(r)}}(2V)\cap B_{L_2}(|\beta_j|)\\j=1,\dots,d}}\frac{1}{n}\sum_{i=1}^n\xi^i\sum_{j=1}^d g_j(X_j^i)|\left\{X^i\right\}_{i=1}^n\right]+\frac{c_1dV\log n}{n}\\&\le\mathbb{E}\left[\sup_{\|\beta\|_2\le l}\sum_{j=1}^d\sup_{g_j\in B_{TV^{(r)}}(2V)\cap B_{L_2}(|\beta_j|)}\frac{1}{n}\sum_{i=1}^n\xi^i g_j(X_j^i)|\left\{X^i\right\}_{i=1}^n\right]+\frac{c_1dV\log n}{n}.\end{aligned}\tag{170}$$

The first inequality follows the same as the proof we have shown in Lemma E.25, see (141). Lemma 3.6 in Bartlett et al. (2005) gives that $\mathcal{E}_1$ holds with probability at least $1-1/n$. On the event $\mathcal{E}_1^c$, by Lemma E.20, we have $\|q_j\|_\infty\le V\log n$, then we have the following bound,

$$\sup_{\delta\in\mathcal{F}^{(r)}(2V)\cap\mathcal{D}(t^2)}\frac{1}{n}\xi^\top\Pi_Q\delta(X)\mathbb{I}\{\mathcal{E}_1^c\}\le c_1dV\log n.$$

And the second inequality follows since Lemma 3.6 in Bartlett et al. (2005), which gives the probability of $\mathcal{E}^c$ as $1/n$.

And the third inequality follows since the decomposability property in L_2 norm by Assumption H.1, which is

$$\left\|\sum_{j=1}^{d} f_j\right\|_{L_2}^2 = \sum_{j=1}^{d} \|f_j\|_{L_2}^2 . \tag{171}$$

then we have

$$\left\|\sum_{j=1}^{d} f_j\right\|_{L_2}^2 \leq l^2 \implies \sum_{j=1}^{d} \|f_j\|_{L_2}^2 \leq l^2, \tag{172}$$

where l is introduced in (168).

Thus the third inequality holds.

The last inequality follows by $\sup(\sum_{j=1}^{d} a_j) \leq \sum_{j=1}^{d} \sup a_j$.

Following the proof steps that lead to Equation (165) in Lemma I.3, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \frac{1}{n} \xi^\top \Pi_Q \delta(X) | \left\{X^i\right\}_{i=1}^n\right] \\
&\leq c n^{-r/(2r+1)} d^{(1+1/r)/(2+1/r)} V^{1/(2r+1)} l + \sqrt{\frac{d \log n}{n}} l + c d^{(2+2/r)/(2+1/r)} n^{-2r/(2r+1)} V^{2/(2r+1)} \\
&= S_1 + S_2 + S_3.
\end{aligned} \tag{173}$$

Based on the proof of Lemma E.25, for q defined in Definition E.13 **6** where $\Pi_Q \delta(X) = q$, we have

$$\|q\|_n \leq m. \tag{174}$$

Then we let

$$\begin{aligned}
m = \max\left\{d^{1/2} V^{1/2}, 1\right\} \Bigg(& t + d^{3/4} \left(\frac{V^{1/2} \sqrt{\log n}}{n^{1/4}} + \frac{t^{1/2} \sqrt{\log n}}{n^{1/2}} + \frac{t \log n}{n^{1/2}} \right) t \\
& + d^{3/2} \left(\frac{t \log n}{n^{1/2}} + \frac{t^2 \log n}{n^{1/2}} + V \log n \right) \Bigg) n^{-1/2}.
\end{aligned} \tag{175}$$

We also get $t_n = c\sqrt{d} n^{-r/(2r+1)} V^{1/(2r+1)}$ from Lemma I.3 and $l = \sqrt{2} m \vee \left(t_n + \frac{\log n}{n}\right)$.

For the first term S_1 in (173), we have

$$\begin{aligned}
S_1 \lesssim \max \Bigg(& V^{\frac{2}{2r+1}} c d^{\frac{2r+\frac{3}{2}}{2r+1}} n^{-\frac{2r}{2r+1}}, \\
& C_2 V^{\frac{1}{2r+1}} d^{\frac{r+1}{2r+1}} n^{-\frac{r}{2r+1}} \max(1, V^{1/2}) \Bigg(\frac{V d^2}{n^{1/2}} + \frac{\sqrt{d} t^{3/2} \sqrt{\log n}}{n^{1/4}} + \sqrt{d} t^2 \log^2 n \\
& + \sqrt{d} t^2 \log n + \frac{d^2 t \log n}{n^{1/2}} \Bigg) \Bigg).
\end{aligned} \tag{176}$$

which follows by the definition of l in (168). For the second term S_2, we have

$$\begin{aligned}
S_2 \lesssim \max \Bigg(& C_2 \max(1, V^{1/2}) \Bigg(\frac{V d^{5/2} \sqrt{\log n}}{n} + \frac{d t^{3/2} \log n}{n^{3/4}} + \frac{d t^2 \log n^{2/5}}{n^{1/2}} \\
& + \frac{d t^2 \log n^{2/3}}{n^{1/2}} + \frac{d^{5/2} t \log n^{2/3}}{n} \Bigg), \quad \frac{V^{1/(2r+1)} d^{(2r+1.5)/(2r+1)} n^{-r/(2r+1)} \sqrt{\log n}}{n^{1/2}} \Bigg).
\end{aligned} \tag{177}$$

which follows by the definition of l in (168). For the third term S_3, we have

$$S_3 = cd^{(2r+2)/(2r+1)}n^{-2r/(2r+1)}V^{2/(2r+1)}. \tag{178}$$

Thus, if we choose

$$t \asymp d^{(r+1)/(2r+1)}n^{-r/(2r+1)}V^{1/(2r+1)} \max\left\{1, V^{(2r-1)/(4r+2)}\right\} \log n, \tag{179}$$

then we have the upper bound of $S_1 + S_2 + S_3$ is all $O(t^2)$.

Using the bounds in (176), (177), and (178), then we come back to (173), get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \frac{1}{n}\xi^\top \Pi_Q \delta(X) | \left\{X^i\right\}_{i=1}^n\right] \\
&\leq cd^{(2r+2)/(2r+1)}n^{-2r/(2r+1)}V^{2/(2r+1)} \max\left\{1, V^{(2r-1)/(2r+1)}\right\} \log n + \frac{c_1 dV \log n}{n} \\
&\lesssim cd^{(2r+2)/(2r+1)}n^{-2r/(2r+1)}V^{2/(2r+1)} \max\left\{1, V^{(2r-1)/(2r+1)}\right\} \log n.
\end{aligned} \tag{180}$$

Since the second term $\frac{c_1 dV \log n}{n}$ is linearly dependent on d, and we see that it is also a lower order term compared to the first term, so we omit that in the third line.

□

Lemma I.5. Let Q_j be defined in Definition E.13. For a positive constant c, let $m = cn^{-r/(2r+1)}V^{1/(2r+1)}$, then the local Rademacher complexity of function class $B_{TV^{(r)}}(2V) \cap Q_j \cap B_{L_2}(m)$ satisfies that

$$\mathcal{R}(B_{TV^{(r)}}(2V) \cap Q_j \cap B_{L_2}(m)) \leq cn^{-2r/(2r+1)}V^{2/(2r+1)}. \tag{181}$$

Proof. Consider the empirical local Rademacher complexity,

$$\mathcal{R}_n(B_{TV^{(r)}}(2V) \cap B_{L_2}(m) = \mathbb{E}\left[\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n}\sum_{i=1}^n \xi^i g_j(X_j^i) | \left\{X^i\right\}_{i=1}^n\right]. \tag{182}$$

Define event $\mathcal{E}_2 \coloneqq \left\{B_{TV^{(r)}}(2V) \cap B_{L_2}(m) \subseteq B_{TV^{(r)}}(2V) \cap B_n(\sqrt{2}m)\right\}$, and let $\mathcal{E}_2^c$ be its complement.

Corollary 2.2 of Bartlett et al. (2005) gives $\mathcal{E}_2$ holds with high probability $1-\eta$, with η is a small positive number. Thus, we have

$$\begin{aligned}
&\mathcal{R}_n(B_{TV^{(r)}}(2V) \cap B_{L_2}(m)) \\
&\leq \mathbb{E}\left[\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n}\sum_{i=1}^n \xi^i g_j(X_j^i)\mathbb{I}\{\mathcal{E}_2\} | \left\{X^i\right\}_{i=1}^n\right] + \mathbb{E}\left[\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n}\sum_{i=1}^n \xi^i g_j(X_j^i)\mathbb{I}\{\mathcal{E}_2^c\} | \left\{X^i\right\}_{i=1}^n\right].
\end{aligned}$$

To bound the first term, notice that

$$\begin{aligned}
\mathbb{E}\left[\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n}\sum_{i=1}^n \xi^i g_j(X_j^i)\mathbb{I}\{\mathcal{E}_2\} | \left\{X^i\right\}_{i=1}^n\right] &\leq \mathbb{E}\left[\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_n(\sqrt{2}m)} \frac{1}{n}\sum_{i=1}^n \xi^i g_j(X_j^i) | \left\{X^i\right\}_{i=1}^n\right] \\
&\leq c_{\mathrm{Dud}} \frac{1}{\sqrt{n}} \int_0^{\sqrt{2}m} \sqrt{\log N(\epsilon, \|\cdot\|_n, B_{TV^{(r)}}(2V))} d\epsilon \\
&\leq c_4 n^{-1/2} V^{1/2r} \int_0^m \epsilon^{-1/2r} d\epsilon \\
&= c_4 n^{-1/2} V^{1/2r} m^{1-1/(2r)},
\end{aligned} \tag{183}$$

where the second line applies Dudley's entropy integral Dudley (1967), with c_{Dud} as a positive constant, and the third line follows from Lemma 15 in Sadhanala and Tibshirani (2019), specifically from the middle derivations and the third-to-last displayed equation on page 45, which uses the fact of TV ball, where $B_{TV^{(r)}}$ is defined in Definition E.1. And c_4 is a positive constant. Also, on the event $\mathcal{E}_2^c$, by Lemma E.20, we have $\|q_j\|_\infty \leq V \log n$, then we have the following bound,

$$\mathbb{E}\left[\sup_{g_j \in B_{TV^{(r)}}(2V) \cap B_{L_2}(m)} \frac{1}{n} \sum_{i=1}^{n} \xi^i g_j(X_j^i) \mathbb{I}\{\mathcal{E}_2^c\} \middle| \left\{X^i\right\}_{i=1}^n\right] \leq cV\eta \log n. \tag{184}$$

Thus, we have

$$\mathcal{R}_n(B_{TV^{(r)}}(2V) \cap B_{L_2}(m)) \leq c_5 n^{-1/2} V^{1/2r} m^{1-1/(2r)}. \tag{185}$$

Therefore, we can upper bound the local Rademacher complexity, splitting the expectation over two events,

$$\mathcal{R}(B_{TV^{(r)}}(2V) \cap B_{L_2}(m)) = \mathbb{E}\left[\mathcal{R}_n(B_{TV^{(r)}}(2V) \cap B_{L_2}(m))\right] \leq c_5 n^{-1/2} V^{1/2r} m^{1-1/(2r)}, \tag{186}$$

where c_5 is a positive constant. Therefore, we have an upper bound on the critical radius r_{nj} is thus given by the solution of

$$\frac{c_5 n^{-1/2} V^{1/2r} m^{1-1/(2r)}}{m} = \frac{m}{c}, \tag{187}$$

for m, which is $m = cn^{-r/(2r+1)} V^{1/(2r+1)}$, this completes the proof.

□

J PROOF OF THEOREM 3.3

Proof. For $t > 1$, by Theorem E.11,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n}\Delta^2(\widehat{f} - f_0) > t^2\right) &\leq \frac{c}{t^2} \mathcal{R}_n\left(\left\{f - f_0 : f \in \mathcal{F}^{(r)}(V)\right\} \cap \left\{f : \Delta^2(f - f_0) \leq nt^2\right\}\right) \\ &= \frac{c}{t^2}\, \mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) \middle| \left\{X^i\right\}_{i=1}^n\right], \end{aligned} \tag{188}$$

we then have

$$\begin{aligned} \sup_{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) \leq & \underbrace{\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X)}_{T_1} \\ & + \underbrace{\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X)}_{T_2}. \end{aligned} \tag{189}$$

Taking expectation conditioned on $\left\{X^i\right\}_{i=1}^n$ on both sides, we have

$$\begin{aligned} &\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(V) - f_0 : \Delta^2(\delta) \leq t^2} \xi^\top \delta(X) \middle| \left\{X^i\right\}_{i=1}^n\right] \\ &\leq \underbrace{\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_P \delta(X) \middle| \left\{X^i\right\}_{i=1}^n\right]}_{T_1} + \underbrace{\mathbb{E}\left[\sup_{\delta \in \mathcal{F}^{(r)}(2V) \cap \mathcal{D}(t^2)} \xi^\top \Pi_Q \delta(X) \middle| \left\{X^i\right\}_{i=1}^n\right]}_{T_2}, \end{aligned} \tag{190}$$

the result follows by bounds of each two terms above based on Lemma I.2 and Lemma I.4. We see that by choosing

$$t \asymp d^{(r+1)/(2r+1)} n^{-r/(2r+1)} V^{1/(2r+1)} \max\left\{1, V^{(2r-1)/(4r+2)}\right\} \log n, \tag{191}$$

we have

$$T_1 = O\left(c_3 d^{3/2}\left(\frac{t}{n^{1/2}} + \frac{t^2}{n^{1/2}} + V\log n\right)\right)$$
$$= O\left(c_3 d^{3/2} d^{(2r+2)/(2r+1)} n^{-(2r+1/2)/(2r+1)} V^{2/(2r+1)} \max\left\{1, V^{(2r-1)/(2r+1)}\right\}\log n\right),$$

and

$$T_2 = O\left(c d^{(2r+2)/(2r+1)} n^{-2r/(2r+1)} V^{2/(2r+1)} \max\left\{1, V^{(2r-1)/(2r+1)}\right\}\log n\right).$$

If d does not grow too quickly, T_1 remains a lower order term relative to T_2 due to its dependence on n. Thus, our analysis only focuses on T_2.

Define the function

$$g_n(t) = \frac{G_n(t)}{t^2}, \tag{192}$$

where $G_n(t)$ equals to the numerator of the right-hand side of (188) that depends on t and n. Then we see that given an small positive ϵ, there is a positive constant c_1 that depends on ϵ such that

$$\begin{aligned}\lim_{c_1\to\infty}\sup_{n\geq 1} g_n(c_1 t) &= \lim_{c_1\to\infty}\sup_{n\geq 1}\frac{T_n(c_1 t)}{c_1^2 t^2}\\ &\leq \lim_{c_1\to\infty}\sup_{n\geq 1}\frac{C_1 c_1 d^{(2r+2)/(2r+1)} n^{-2r/(2r+1)} V^{2/(2r+1)}\max\left\{1, V^{(2r-1)/(2r+1)}\right\}\log n}{c_1^2 t^2}\\ &= O\left(\frac{1}{c_1}\right) < \epsilon,\end{aligned} \tag{193}$$

by setting

$$t^2 \asymp d^{\frac{2r+2}{2r+1}} n^{-\frac{2r}{2r+1}} V^{\frac{2}{2r+1}} \max\left\{1, V^{\frac{2r-1}{2r+1}}\right\}.$$

Thus, we conclude that

$$\Delta_n^2(\widehat{f} - f_0) = O_{pr}\left(d^{\frac{2r+2}{2r+1}} n^{-\frac{2r}{2r+1}} V^{\frac{2}{2r+1}} \cdot \max\left\{1, V^{\frac{2r-1}{2r+1}}\right\}\right).$$

which is the conclusion in Theorem 3.3.

□

K SOLVING PROBLEM (16)

K.1 Reformat of Problem (16)

We reformulate the optimization problem (16) as

$$\begin{aligned}\min_{\theta_j\in\mathbb{R}^n, z\in\mathbb{R}^n} \quad & \sum_{i=1}^n \rho_\tau(u_j^i - \theta_j^i) + \lambda\left\|D_n^{(X_j,r)} S_j z_j\right\|_1,\\ \text{subject to} \quad & z_j = \theta_j, \quad \mathbb{1}^\top\theta_j = 0,\end{aligned} \tag{194}$$

and the augmented Lagrangian can then be written as

$$L(\theta_j, z_j, s_j, \nu_j) = \sum_{i=1}^n \rho_\tau(u_j^i - \theta_j^i) + \lambda\left\|D_n^{(X_j,r)} S_j z_j\right\|_1 + \frac{\eta}{2}\|\theta_j - z_j + s_j\|^2 + \frac{\omega}{2}(\mathbb{1}^\top\theta_j + \nu_j)^2, \tag{195}$$

where η and ω is the penalty parameter that controls step size in the update.

Thus we initialize the variables $\theta_j^{(0)} = 0$ and $z_j^{(0)} = 0$ for $j = 1, \ldots, d$. We can solve (194) iteratively until convergence, with mth iteration, we have:

$$\theta_j^{(m)} = \arg\min_{\theta_j \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \rho_\tau(u_j^i - \theta_j^i) + \frac{\eta}{2}\|\theta_j - z_j^{(m-1)} + s_j^{(m-1)}\|^2 + \frac{\omega}{2}(\mathbb{1}^\top \theta_j + \nu_j^{(m-1)})^2 \right\}, \tag{196}$$

$$z_j^{(m)} = \arg\min_{z_j \in \mathbb{R}^n} \left\{ \frac{1}{2}\|\theta_j^{(m)} + s_j^{(m-1)} - z_j\|^2 + \frac{\lambda}{\eta}\|D_n^{(X_j,r)} S_j z_j\|_1 \right\}, \tag{197}$$

$$s_j^{(m)} \leftarrow s_j^{(m-1)} + \theta_j^{(m)} - z_j^{(m)}, \tag{198}$$

$$\nu_j^{(m)} \leftarrow \nu_j^{(m-1)} + \mathbb{1}^\top \theta_j^{(m)}. \tag{199}$$

To solve (196), we first set

$$u_j^i \leftarrow Y^i - \sum_{l<j} \theta_l^{i(m)} - \sum_{l>j} \theta_l^{i(m-1)}, \tag{200}$$

then we do the following update,

$$\theta_j^{i(m)} \leftarrow \begin{cases} \frac{1}{\eta+\omega}\left(\tau - \omega \sum_{l \neq j} \theta_j^{l(m-1)} - \omega\nu^{(m-1)} + \eta(z_j^{i(m-1)} - s_j^{i(m-1)})\right), \\ \quad \text{if } u_j^i > \frac{1}{\eta+\omega}\left(\tau - \omega \sum_{l \neq j} \theta_j^{l(m-1)} - \omega\nu^{(m-1)} + \eta(z_j^{i(m-1)} - s_j^{i(m-1)})\right), \\ \frac{1}{\eta+\omega}\left(\tau - 1 - \omega \sum_{l \neq j} \theta_j^{l(m-1)} - \omega\nu^{(m-1)} + \eta(z_j^{i(m-1)} - s_j^{i(m-1)})\right), \\ \quad \text{if } u_j^i < \frac{1}{\eta+\omega}\left(\tau - 1 - \omega \sum_{l \neq j} \theta_j^{l(m-1)} - \omega\nu^{(m-1)} + \eta(z_j^{i(m-1)} - s_j^{i(m-1)})\right), \\ u_j^i, \quad \text{otherwise.} \end{cases} \tag{201}$$

To solve (197), we do two cases. For $r = 1$, we use dynamic programming by Johnson (2013). For $r > 1$, we use ADMM algorithm from (Ramdas and Tibshirani, 2016), see Below K.2 for more details.

K.2 Solve Problem (197) by ADMM

The minimization Problem (197) is a univariate trend filtering problem, on unevenly spaced inputs. For $r > 1$, the solution to this problem has been well studied, and we solve it using methods described in Ramdas and Tibshirani (2016), implemented in the `trendfilter` function in the `glmgen` R package.

The standard ADMM approach (e.g., Boyd et al. (2011)) is based on rewriting problem (197) as

$$\min_{z_j \in \mathbb{R}^n,\, \alpha_j \in \mathbb{R}^{n-k-1}} \frac{1}{2}\|\theta_j + s_j - z_j\|_2^2 + \lambda/\eta\|\alpha_j\|_1 \quad \text{subject to} \quad \alpha_j = D_n^{(X_j,r)} S_j z_j. \tag{202}$$

The augmented Lagrangian can then be written as

$$g(z_j, \alpha_j, u_j) = \frac{1}{2}\|\theta_j + s_j - z_j\|_2^2 + \lambda/\eta\|\alpha_j\|_1 + \frac{\rho}{2}\|\alpha_j - D_n^{(X_j,r)} S_j z_j + u_j\|_2^2 - \frac{\rho}{2}\|u_j\|_2^2,$$

with updates as

$$z_j^{(k)} \leftarrow \left(I + \rho(D_n^{(X_j,r)})^\top D_n^{(X_j,r)}\right)^{-1}\left(\theta_j^{(k)} + s_j^{(k-1)} + \rho(D_n^{(X_j,r)})^\top(\alpha_j^{(k-1)} + u_j^{(k-1)})\right), \tag{203}$$

$$\alpha_j^{(k)} \leftarrow S_{\lambda/(\eta\rho)}\left(D_n^{(X_j,r)} S_j z_j^{(k)} - u_j^{(k-1)}\right), \tag{204}$$

$$u_j^{(k)} \leftarrow u_j^{(k-1)} + \alpha_j^{(k)} - D_n^{(X_j,r)} S_j z_j^{(k)}, \tag{205}$$

The α-update, where $S_{\lambda/(\eta\rho)}$ denotes coordinate-wise soft-thresholding at the level $\lambda/(\eta\rho)$,

K.3 Computational Complexity

In Algorithm 1, the backfitting process is conducted, involving d distinct fits of quantile trend filtering. For each fit of quantile trend filtering. The ADMM procedure involves three updates (primal, dual, and slack). The total number of ADMM iterations is denoted by m. For each update, updating the primal variable θ_j in (201) in $O(n)$ time and computing u_j in (200) in $O(n)$ time. The slack variable z_j, introduced in Equation (194), is updated via dynamic programming in $O(n)$ time for $r = 1$. For $r > 1$, z_j can be updated in $O(n(r+2)^2)$ time using another ADMM approach. The update of auxiliary variable takes time $O(n - r - 1)$. The dual update taking time $O(n(r+2))$. The update of s_j in (198) takes time $O(1)$. The update of ν_j in (199) takes time $o(n)$. There are d components. And therefore one full iteration of standard backfitting ADMM updates can be done in linear time $O(dmn)$(considering r as a constant).

L CONVERGENCE OF THE UPDATE IN (12)

We demonstrate the convergence of the update in Equation (201) through the following theorem.

Theorem L.1. Let $\{\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})\}_{t=0,1,\ldots}$ represent the parameter updates at the tth iteration in BCD in Algorithm 1. Under the conditions of Theorem 3.3, it holds that every cluster point of the sequence $\theta^{(t)}$ generated by the BCD method is a coordinate minimum point of Equation (12).

We have the problem (12)

$$\begin{aligned} &\min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \sum_{i=1}^{n} \rho_\tau(Y^i - \sum_{j=1}^{d} \theta_j^i) + \lambda \sum_{j=1}^{d} \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1 \\ &\text{subject to} \quad \mathbb{1}^\top \theta_j = 0, \quad j = 1, \ldots, d. \end{aligned}$$

Define the following functions,

$$f(\theta_1, \ldots, \theta_d; \lambda) \coloneqq f_0(\theta_1, \ldots, \theta_d) + \sum_{j=1}^{d} f_j(\theta_j).$$

$$f(\theta_1, \ldots, \theta_d; \lambda) \coloneqq \min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \sum_{i=1}^{n} \rho_\tau(Y^i - \sum_{j=1}^{d} \theta_j^i) + \lambda \sum_{j=1}^{d} \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1 .$$

$$f_0(\theta_1, \ldots, \theta_d) \coloneqq \min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \sum_{i=1}^{n} \rho_\tau(Y^i - \sum_{j=1}^{d} \theta_j^i),$$

$$f_j(\theta_j) \coloneqq \left\| D_n^{(X_j,r)} S_j \theta_j \right\|_1 .$$

We observe that

(1) f_0 is continuous on dom f_0.

(2) For each $j \in \{1, \ldots, d\}$ and $(\theta_j)_{j \neq k}$, the function $\theta_j \mapsto f(\theta^1, \ldots, \theta_d)$ is quasiconvex and hemivariate.

(3) $f_0, f_1, \ldots, f_d$ are lower semicontinuous .

(4) dom $f_0 = Y_1 \times \cdots \times Y_d$, for some $Y_j \subseteq \mathbb{R}^n$, $j = 1, \ldots, d$.

Then by Theorem 5.1 of Tseng (2001), we have the every cluster point is a coordinatewise minimum point of f.

M RELATED WORKS

M.1 Review of Additive Models

Since introduced, additive models have been extensively studied in various contexts of statistics and machine learning, including Cox regression (Cox, 1972), logistic regression (Hastie and Tibshirani, 1987), exponential family data distributions (Hastie, 2017), neural networks (Thielmann et al., 2024; Agarwal et al., 2021; Shen et al., 2021; Jo and Kim, 2023), and online machine learning (Abbasi-Yadkori et al., 2011; Li et al., 2017; Ding et al., 2022; Kang et al., 2022). Due to their model structure, additive models effectively address the curse of dimensionality (Breiman and Friedman, 1985; Hastie, 2017), allowing for more accurate and interpretable predictions.

M.2 Review of Trend Filtering

In the field of image processing, total variation smoothing penalties were introduced by (Rudin et al., 1992). These penalties enable edge detection and permit sharp breaks in gradients, surpassing the limitations of conventional Sobolev penalties.

Proposed independently by Rudin et al. (1992); Kim et al. (2009), trend filtering, which functions as a total variation smoothing penalty, is a relatively new approach to univariate nonparametric regression. Notably, the work by Sadhanala and Tibshirani (2019) studies additive trend filtering estimates, involving regularizing each component function based on the total variation of its rth order discrete derivative. Overall, trend filtering is favored for its favorable theoretical and computational properties, largely due to the localized nature of the total variation regularization it employs.

M.3 Review of Quantile Regression

Unlike classical mean regression, quantile regression estimates conditional quantiles of the response variable, making it robust to outliers and providing a more comprehensive view of the relationship between the response and covariates (Koenker, 2005).

The quantile trend filtering method, as proposed by Brantley et al. (2020) and studied in Madrid Padilla and Chatterjee (2022), produces a trend filtering estimator with smooth structures by imposing a penalty consisting of the total variation of the rth order discrete derivatives. The quantile fused lasso is a special case of quantile trend filtering when $r = 0$. The risk bound for quantile trend filtering has been studied by Madrid Padilla and Chatterjee (2022), which established the minimax optimality of univariate trend filtering under the quantile loss with minor assumptions of the data generation mechanism.

N IMPORTANT DEFINITIONS AND EXAMPLES

N.1 Total Variation

The total variation of the $(r-1)$th weak derivative of a function f, denoted as $TV^{(r)}(f)$, is a key concept in this formulation. It is well-known that the total variation $TV^{(r)}(f)$ is equivalent to the Riemann approximation of the integral $\int_{[0,1]} |f^{(r)}(t)|\, dt$ if f is r times differentiable. Moreover, for $(r-1)$th order polynomials, the rth derivative is zero, making $TV^{(r)}(g) = 0$ for all $g(x) = x^l$ where $l = 0, \ldots, r-1$.

N.2 Discrete Difference Operator for Trend Filtering

The rth order trend filtering estimate is derived from the following penalized least squares optimization problem:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \left(\frac{1}{2} \|y - \theta\|_2^2 + \frac{n^{r-1}}{(r-1)!} \cdot \lambda \left\| D^{(r)} \theta \right\|_1 \right),$$

as described by Kim et al. (2009) and Tibshirani (2014). Here, $D^{(r)}$ represents the discrete difference operator of order r, which is a banded matrix with bandwidth $r+1$. The operator $D^{(r)}$ can be understood as the discrete analogue of the rth order derivative operator, with the penalty term enforcing smoothness by penalizing the discrete rth derivative of the vector $\theta \in \mathbb{R}^n$.

N.2.1 Examples of the Discrete Difference Operator

To further clarify the structure of $D_n^{(X,r)}$, we consider specific examples for $r = 1$ and $r = 2$, which approximate the first and second derivatives, respectively.

Example 1 ($r = 1$)**.** Given four points (X^1, X^2, X^3, X^4) and its sorted ones $0 < X^{(1)} < X^{(2)} < X^{(3)} < X^{(4)} < 1$ within the interval $[0, 1]$, the first-order operator $D_4^{(X,1)}$ is:

$$D_4^{(X,1)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 4}, \quad D_4^{(X,1)}(\theta) = (\theta^2 - \theta^1, \theta^3 - \theta^2, \theta^4 - \theta^3).$$

Example 2 ($r = 2$)**.** For the second-order operator $D_4^{(X,2)}$, using the same four points as before, we have:

$$D_4^{(X,2)} = D_3^{(X,1)} \operatorname{diag}\left(\frac{1}{X^{(2)} - X^{(1)}}, \frac{1}{X^{(3)} - X^{(2)}}, \frac{1}{X^{(4)} - X^{(3)}}\right) D_4^{(X,1)},$$

which explicitly becomes:

$$D_4^{(X,2)} = \begin{pmatrix} \frac{1}{X^{(2)}-X^{(1)}} & \frac{-1}{X^{(2)}-X^{(1)}} + \frac{-1}{X^{(3)}-X^{(2)}} & \frac{1}{X^{(3)}-X^{(2)}} & 0 \\ 0 & \frac{1}{X^{(3)}-X^{(2)}} & \frac{-1}{X^{(3)}-X^{(2)}} + \frac{-1}{X^{(4)}-X^{(3)}} & \frac{1}{X^{(4)}-X^{(3)}} \end{pmatrix}.$$

This operator acts as a second-order difference operator, adjusted for the non-uniform spacing of the points.

Special Case ($r = 0$)**.** For $r = 0$, the operator $D_n^{(X,0)}$ is simply the identity matrix:

$$D_n^{(X,0)} = I, \quad D_n^{(X,0)}(\theta) = \theta.$$

O EXAMPLES USING ALGORITHM 1

Example O.1. In Figure 1, we present the true quantile signal alongside the fitted values for each component in the model. The black curves represent the functions $f_{0j}(x) = a_j g_j(x) - b_j$ for $j = 1, \ldots, 4$, where a_j and b_j are chosen such that $f_{0j}(X_j)$ has an empirical mean zero and an empirical norm $\|f_{0j}\|_n = 1$. Specifically:

- $g_1(x) = -\frac{1}{2}x^2$
- $g_2(x) = \frac{3}{2}\sin(4\pi x) + \mathbb{1}_{x \le \frac{1}{2}} \cdot \sin(16\pi x)$
- g_3 is a dummy dimension (where only one randomly assigned point takes a non-zero value)
- $g_4(x) = e^{3x}\sin(4\pi x)$.

The blue curves in the figure plot the fitted $\widehat{f}_j$ values for $j = 1, \ldots, 4$. The model is fitted on 1000 noisy data points, constructed similarly to our simulated experiments in Section 5. Here, the noise is generated from independent draws from a t distribution with 3 degrees of freedom denoted as $t(3)$. This figure demonstrates that quantile additive trend filtering effectively captures varying levels of smoothness both within and between component functions, even under heavy-tailed error conditions.

Example O.2. Figure 2 provides a 3D visualization of the true quantile signal $\sum_{j=1}^2 f_{0j}(X_j)$, the noisy data y, and the reconstructed signal $\sum_{j=1}^2 \widehat{f}_j$. These components, f_{0j}, y, and $\widehat{f}_j$ for $j = 1, \ldots, 2$, are defined similarly to those in Example 1, and $X_j = (X_j^1 \ldots X_j^{2000})$ constructed as described in our experiments in Section 5. The underlying component functions here are:

- $g_1(x) = \frac{1}{2}\cos(6\pi x) + 0.1$
- $g_2(x) = -(x - \frac{1}{2})^2$.

With only two input dimensions, the model's output can be directly plotted along its inputs, allowing us to visually assess the model's effectiveness and the performance of our back-fitting algorithm.

P COMPUTATION SETTING

In the main paper, we have provided detailed information on data reproduction in the experimental section referenced as 5. All experiments were conducted on a Linux-based system equipped with an Intel(R) Xeon(R) Platinum 8160 CPU running at 2.10 GHz. The system had 24 processor cores and a total memory capacity of 260 GB. The experiments were performed using R version 4.2.2.

Q ADDITIONAL EXPERIMENTS RESULTS

Table 3: Average mean squared error, $\frac{1}{n}\sum_{i=1}^{n}(f_0^i - \widehat{f^i})^2$, averaging over 50 Monte Carlo simulations for the different methods considered. The best MSE is listed with bold text.

n	Scenario	d	τ	QATF1	QATF0	QS	ATF1	ATF0
500	1	10	0.2	**1.3440**	1.4921	1.3603	NA	NA
1000	1	10	0.2	**0.7628**	1.1676	0.9103	NA	NA
2500	1	10	0.2	**0.4034**	0.5036	0.5405	NA	NA
500	1	10	0.8	**1.3648**	1.6115	1.5026	NA	NA
1000	1	10	0.8	**0.8058**	1.1484	0.9332	NA	NA
2500	1	10	0.8	**0.3812**	0.5585	0.5304	NA	NA
500	2	10	0.2	3.1722	**3.1520**	3.3013	NA	NA
1000	2	10	0.2	**3.0346**	3.2824	3.1760	NA	NA
2500	2	10	0.2	**1.3725**	1.6200	1.6323	NA	NA
500	2	10	0.8	**4.4045**	4.5964	4.6682	NA	NA
1000	2	10	0.8	**2.7262**	3.1071	2.7387	NA	NA
2500	2	10	0.8	**1.4714**	1.6283	1.7447	NA	NA
500	3	10	0.2	**1.3607**	1.9141	1.5422	NA	NA
1000	3	10	0.2	**0.6426**	1.1087	0.8326	NA	NA
2500	3	10	0.2	**0.2518**	0.4391	0.3849	NA	NA
500	3	10	0.8	**2.2376**	2.4855	2.4760	NA	NA
1000	3	10	0.8	**1.7452**	2.0362	1.9083	NA	NA
2500	3	10	0.8	**1.1270**	1.2700	1.2928	NA	NA
500	5	10	0.2	2.4657	**2.3672**	2.7830	NA	NA
1000	5	10	0.2	1.6413	**1.4640**	1.7123	NA	NA
2500	5	10	0.2	0.9482	**0.6189**	0.9951	NA	NA
500	5	10	0.8	2.7530	**2.5349**	2.8937	NA	NA
1000	5	10	0.8	1.5717	**1.5360**	1.6207	NA	NA
2500	5	10	0.8	0.9286	**0.6175**	0.9613	NA	NA
500	6	10	0.2	**3.4553**	**3.4553**	3.8648	NA	NA
1000	6	10	0.2	**2.7782**	3.3125	2.9650	NA	NA
2500	6	10	0.2	**2.2229**	2.4699	2.3843	NA	NA
500	6	10	0.8	**0.0245**	0.0345	2.6176	NA	NA
1000	6	10	0.8	**0.0243**	0.0267	1.7869	NA	NA
2500	6	10	0.8	**0.0245**	0.0271	1.4223	NA	NA

R WORLD HAPPINESS COMPONENT PLOTS

Provided in this section are the component plots for all variables considered in the statistical inference study on the World Happiness Data.

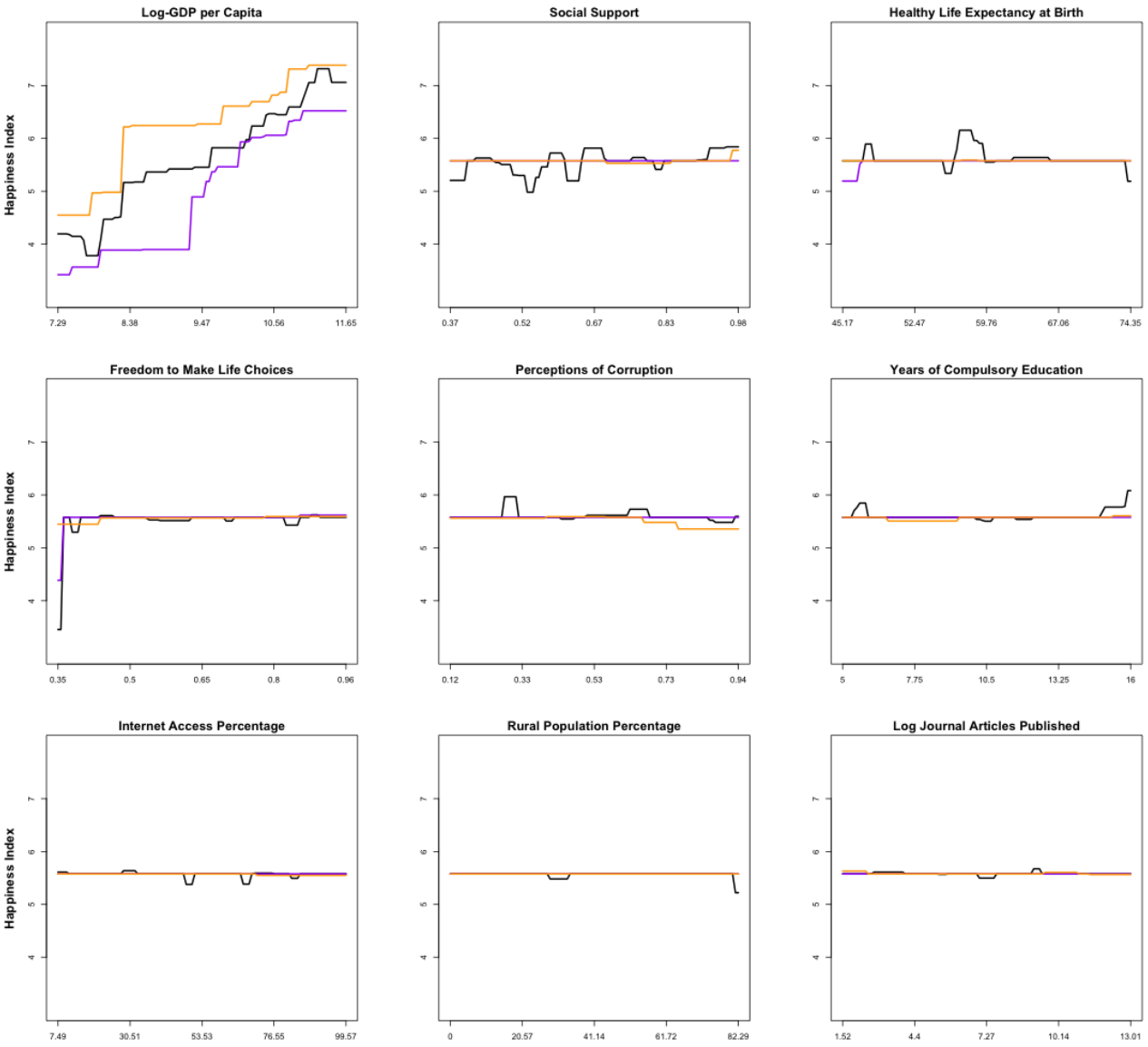


Figure 4: Quantile Additive Trend Filtering component estimations with $\tau = 0.1, 0.5,$ and 0.9, plotted in purple, black, and orange, respectively, for most relevant components.