Nonparametric Methods with Total Variation Type Regularization

Veeranjaneyulu Sadhanala

May 2019
CMU-ML-19-104

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Ryan Tibshirani, Chair
Aarti Singh
Dejan Slepcev
Larry Wasserman
James Sharpnack (UC Davis)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

This research was sponsored by the National Science Foundation grant DMS1554123, the National Institutes of Health grant R01GM093156 and a gift from Adobe.

Keywords: nonparametric regression, total variation, higher order total variation, image denoising, additive models, trend filtering, lattice graphs, hypothesis testing, Kolmogorov Smirnov

Abstract

We consider two special cases of the classical nonparametric regression problem of estimating a function $f : \mathbb{R}^d \to \mathbb{R}$ given n noisy observations at inputs $x_1, \cdots, x_n \in \mathbb{R}^d$.

In the first case, the input points correspond to nodes on a *d-dimensional regular lattice* (having equal side lengths $n^{1/d}$). We define two new higher-order TV classes with signals which are allowed to exhibit different amounts of smoothness at different regions in the lattice, and derive lower bounds on their minimax estimation errors. We analyze two naturally associated estimators and show that they are rate optimal over the appropriate classes. Linear smoothers are shown to be minimax suboptimal over these TV classes.

In the second case, we consider *additive models* built with kth order trend filters resulting in kth degree piecewise polynomial components. We derive fast error rates for additive trend filtering and prove that these rates are minimax optimal when the underlying function is itself additive and has component functions whose derivatives have bounded kth order TV. We show that such rates are unattainable by additive models built from linear smoothers.

We also study an extension of the Kolmogorov-Smirnov (KS) two-sample test, which can be more sensitive to differences in the tails. Our test statistic is an integral probability metric (IPM) defined over a higher-order total variation ball, recovering the original KS test as its simplest case.

Acknowledgements

I will be forever in debt of my advisor Ryan. Most of the good things I learned at CMU are from him. It was inspiring to see his optimistic and creative approach when we were confronted with tricky research problems. He was very kind and supportive while I was going through a rough patch personally. I was truly fortunate to be mentored by him. I thank Aarti, Dejan, Larry and James for serving on my committee and directing me with interesting questions. I would like to thank Dejan for additionally helping us with his insights on some problems. I also thank Eric Xing for his guidance in my first year at CMU and my undergraduate advisor S. Sudarshan.

I was lucky to work with Yu-Xiang Wang, James Sharpnack, Aaditya Ramdas, Suvrit Sra, Alnur Ali, Wei Dai and Willie Neiswanger. In particular, working with Yu-Xiang Wang was a learning experience that I will cherish. I thank Alex Smola for his guidance in some projects. I am grateful to Andrew Price for giving me an opportunity to work with him at Amazon.

Diane has been very kind to me from the time I first visited CMU for open house. Many thanks to her for that. If everybody were like her, the world would be a much happier place.

Several friends helped me during my stay at Pittsburgh. Sashank Reddi in particular guided me several times on professional and personal issues. I will fondly remember the lighter conversations with him, Chaitanya Amdekar, Vikram Kamath, Yash Puranik, Nisarg Shah and Avinava Dubey. Thanks to Sangwon Hyun for teaching me how to drive and also for some cheerful conversations over coffee with Alnur.

Finally, I would not have been here without the strong support from my family. My father Rangarao and mother Samrajyam always believed in me, and encouraged me to pursue my studies. My wife Supraja left her job in India and moved to USA for me. She wholeheartedly supported my career decisions and had to bear with my irregular schedules. I am grateful to my grandfather Subbaiah who initiated my schooling and my *mavayya* Anjaneyulu for guiding me throughout my early years of education. I thank my brothers, sisters and maternal uncles who have always been there for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We study the classic nonparametric regression problem where the responses $Y^i, i = 1, \ldots, n \in \mathbb{R}$ and the inputs $X^i, i = 1, \ldots, n \in \mathbb{R}^d$ are related by

$$Y^i = f(X^i) + \epsilon^i, \quad i = 1, \ldots, n \tag{1.1}$$

where f is a smooth function we intend to estimate and ϵ^i are i.i.d. subgaussian with mean zero. Compared to parametric regression methods, where the regression function f belongs to a class of functions which can be parameterized by a fixed number of variables (such as linear functions, polynomials up to a degree less than a fixed positive integer), nonparametric regression methods allow for a lot of flexibility in the estimate. They are very useful tools when we do not have much information about the structure of the regression function or when we do not want to restrict the regression function class much. There are many popular nonparametric regression methods such as kernel smoothing, k-nearest neighbors, local polynomial regression, Gaussian process regression, RKHS regression, smoothing splines and wavelet shrinkage. Thousands of papers and several dozens of books have been published on these methods over the past century; hence we do not attempt to list all the works and refer the reader to the books Gyorfi et al. (2002), Tsybakov (2009) and Green & Silverman (1994). Many standard software packages provide implementations of these methods.

We are interested in estimating *heterogeneously* smooth functions f, which are allowed to exhibit different amounts of smoothness at different regions in the grid. Such heterogeneity eludes classical measures of smoothness from nonparametric statistics, such as Holder smoothness. Total variation (TV) smoothness classes allow for heterogeneity. Locally adaptive regression splines (Mammen & van de Geer 1997) automatically adapt to the local level of smoothness of the true function in 1d. Trend filtering(Steidl et al. 2006, Kim et al. 2009, Tibshirani 2014) can be thought of as a computationally fast version of locally adaptive regression splines with no loss in statistical properties.

As the dimension d of the input space grows, nonparametric regression turns into a notoriously difficult problem. We study the following two specific settings of the problem with a focus on estimating heterogeneously smooth functions:

(a) Lattices (grids) in d-dimensions. X^i correspond to nodes in a *d-dimensional grid graph* and Y^i are observations at the nodes. f belongs to a class of functions whose kth weak derivative has a bounded total variation in a sense that is described later.

(b) Additive model. f is an *additive* function

$$f(X_1, \ldots, X_d) = \sum_{j=1}^{d} f_j(X_j)$$

where the univariate component functions f_j are smooth in the sense that their kth weak derivative has a bounded total variation.

We derive minimax optimal rates for the two settings and show that our trend filtering estimators achieve these rates upto $\log n$ factors. Importantly, no linear smoother (including popular methods such as kernel smoothers, Gaussian process regression, Laplacian smoothing, Laplacian eigenmaps on graphs) can achieve the minimax optimal rate. In fact, in the case of grids, when the smoothness level is low, no linear smoother is even consistent! This extends fundamental findings of Donoho & Johnstone (1998) in 1-dimensional total variational spaces to higher dimensions.

We also study a related *two-sample test* with the test statistic

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(X^i) - \frac{1}{n} \sum_{j=1}^{n} f(Y^j) \right|$$

where $X^1, \ldots, X^m \in \mathbb{R}$ and $Y^1, \ldots, Y^n \in \mathbb{R}$ are samples from two distributions and $\mathcal{F}$ is a certain set of functions whose kth weak derivative is bounded in total variation. This reduces to the classic Kolmogorov-Smirnov (KS) test if $k = 0$. Therefore our test may be seen as a higher-order version of the KS test.

We now review univariate trend filtering which is the foundation to many of our methods.

1.1 Trend filtering in 1d

The kth order *locally adaptive regression splines* (LAR splines) (Mammen & van de Geer 1997) defined by

$$\hat{f} = \operatorname*{argmin}_{f} \frac{1}{n} \sum_{i=1}^{n} (Y^i - f(X^i))^2 + \lambda \mathrm{TV}(f^{(k)}) \tag{1.2}$$

where $\lambda \geq 0$ is a tuning parameter, $\mathrm{TV}(f)$ denotes the total variation of a univariate real-valued function and $f^{(k)}$ denotes the kth weak derivative of a function f. LAR splines automatically adapt to the local level of smoothness of the true function. They achieve the minimax optimal rate of $n^{-(2k+2)/(2k+3)}$ over the kth order total variation space

$$\mathcal{F}_k(C_n) = \left\{ f : [0,1] \to \mathbb{R} \mid \mathrm{TV}(f^{(k)}) \leq C_n \right\} \tag{1.3}$$

where C_n is a sequence of positive numbers. Popular nonparametric regression methods such as local polynomial kernel regression, smoothing splines and Gaussian process regression cannot achieve the minimax optimal rate over this class. A difficulty with LAR splines is that the defining optimization problem (1.2) is computationally hard.

Proposed independently by Steidl et al. (2006) and Kim et al. (2009), trend filtering preserves the local adaptivity property of LARS while being computationally efficient. As

explained in Tibshirani (2014), it can be seen as a discrete-time analog of LARS. Denoting by $X = (X^1, \ldots, X^n) \in \mathbb{R}^n$ the vector of univariate input points with $X^1 < \ldots < X^n$, the trend filtering estimate of order $k \geq 0$ is defined as the solution of the optimization problem

$$\min_{\theta \in \mathbb{R}^n} \ \frac{1}{2}\|Y - \theta\|_2^2 + \lambda \|D^{(X,k+1)}\theta\|_1, \tag{1.4}$$

where $\lambda \geq 0$ is a tuning parameter, and $D^{(X,k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is a kth order difference operator, constructed based on X. These difference operators can be defined recursively, as in

$$D^{(X,1)} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}, \tag{1.5}$$

$$D^{(X,k+1)} = D^{(X,1)} \cdot \mathrm{diag}\left(\frac{k}{X^k - X^1}, \ldots, \frac{k}{X^n - X^{n-k+1}}\right) \cdot D^{(X,k)} \in \mathbb{R}^{(n-k-1)\times n}, \quad k \geq 1. \tag{1.6}$$

(The leading matrix $D^{(X,1)}$ in (1.6) is the $(n-k-1) \times (n-k)$ version of the difference operator in (1.5).) Intuitively, the interpretation is that the problem (1.4) penalizes the sum of absolute $(k+1)$st order discrete derivatives of $\theta^1, \ldots, \theta^n$ across the input points $X^1, \ldots, X^n$. Thus, at optimality, the coordinates of the trend filtering solution $\hat{\theta}^1, \ldots, \hat{\theta}^n$ obey a kth order piecewise polynomial form. See Figure 1.1.

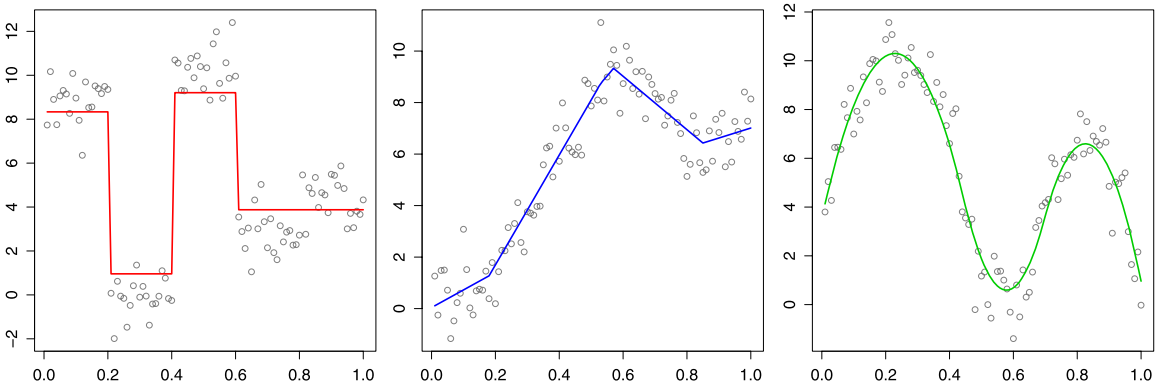

Figure 1.1: *(From Tibshirani (2014)) Piecewise kth degree polynomial structure of trend filtering estimates for $k = 0, 1, 2$ from left to right.*

This intuition is formalized in Tibshirani (2014) and Wang et al. (2014), where it is shown that the components of the kth order trend filtering estimate $\hat{\theta}$ are precisely the evaluations of a fitted kth order piecewise polynomial function across the inputs, and that the trend filtering and locally adaptive regression spline estimates of the same order k are asymptotically equivalent. When $k = 0$ or $k = 1$, in fact, there is no need for asymptotics, and the equivalence between trend filtering and locally adaptive regression spline estimates is exact in finite samples. It is also worth pointing out that when $k = 0$, the trend filtering

estimate reduces to the 1d fused lasso estimate (Tibshirani et al. 2005), which is known as 1d total variation denoising in signal processing (Rudin et al. 1992).

Over the kth order total variation function class defined in (1.3), Tibshirani (2014), Wang et al. (2014) prove that kth order trend filtering achieves the minimax optimal $n^{-(2k+2)/(2k+3)}$ error rate, just like kth order locally adaptive regression splines. Another important property, as developed by Steidl et al. (2006), Kim et al. (2009), Tibshirani (2014), Ramdas & Tibshirani (2016), is that trend filtering estimates are relatively cheap to compute—much cheaper than locally adaptive regression spline estimates—owing to the bandedness of the difference operators in (1.5), (1.6), which means that specially implemented convex programming routines can solve (1.4) in an efficient manner.

Falling factorial basis. Tibshirani (2014), Wang et al. (2014) establish a connection between univariate trend filtering and the falling factorial functions, and show that the trend filtering problem can be interpreted as a sparse basis regression problem using these functions. Given knot points $t^1 < \ldots < t^n \in \mathbb{R}$, the kth order falling factorial basis functions $h_1, \ldots, h_n$ are defined by

$$\begin{aligned} h_i(t) &= \prod_{\ell=1}^{i-1} (t - t^\ell), \quad i = 1, \ldots, k+1, \\ h_{i+k+1}(t) &= \prod_{\ell=1}^{k} (t - t^{i+\ell}) \cdot 1\{t > t^{i+k}\}, \quad i = 1, \ldots, n-k-1. \end{aligned} \tag{1.7}$$

Our convention is to define the empty product to be 1, so that $h_1(t) = 1$. The functions $h_1, \ldots, h_n$ are piecewise polynomial functions of order k, and appear very similar in form to the well-known kth order truncated power basis functions defined as follows:

$$\begin{aligned} g_1(x) = 1, g_2(x) = x, \ldots, g_{k+1}(x) = x^k, \\ g_{k+1+j}(x) = (x - t^j)^k \cdot 1(x \geq t^j), \quad j = 1, \ldots, n-k-1. \end{aligned} \tag{1.8}$$

See Figure 1.2. In fact, when $k = 0$ or $k = 1$, the two bases are exactly equivalent (meaning that they have the same span). Similar to an expansion in the truncated power basis, an expansion in the falling factorial basis,

$$g = \sum_{i=1}^{n} \alpha^i h_i$$

is a continuous piecewise polynomial function, having a global polynomial structure determined by $\alpha^1, \ldots, \alpha^{k+1}$, and exhibiting a knot—i.e., a change in its kth derivative—at the location t^{i+k} when $\alpha^{i+k+1} \neq 0$. But, unlike the truncated power functions, the falling factorial functions in (1.7) are not splines, and when g (as defined above) has a knot at a particular location, it displays a change not only in its kth derivative at this location, but also in all lower order derivatives (i.e., all derivatives of orders $1, \ldots, k-1$).

Define the kth order falling factorial basis matrix $H^{(k)}$ and truncated power basis matrix $G^{(k)}$ by:

$$H^{(k)}_{ij} = h_j(t^i), \quad G^{(k)}_{ij} = g_j(t^i).$$

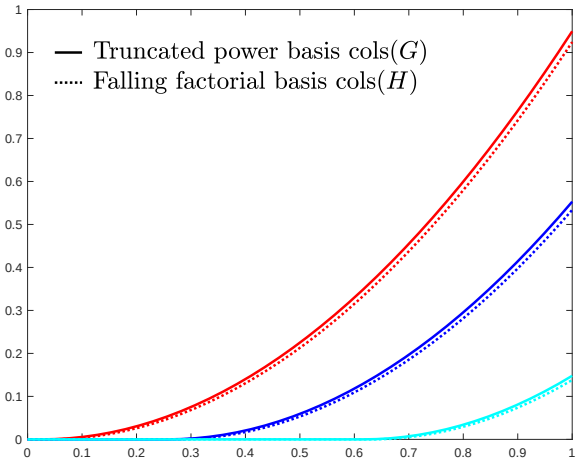

Figure 1.2: *Some falling factorial basis and truncated power basis functions for* $k = 2$.

Tibshirani (2014), Wang et al. (2014) show that (1.4) may be written in lasso form as

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2}\|Y - H^{(k)}\alpha\|_2^2 + \lambda \sum_{i=k+2}^{n} |\alpha_i| \tag{1.9}$$

where the solutions of (1.9) and (1.4) are related by $\hat{\theta} = H^{(k)}\hat{\alpha}$. This also suggests the variational formulation (see Wang et al. (2014))

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(Y^i - f(X^i))^2 + \lambda \sum_{i=k+2}^{n} \mathrm{TV}(f^{(k)}) \tag{1.10}$$

where $\mathcal{H} = \left\{ \sum_{i=1}^{n} \alpha_i h_i \right\}$ is the span of falling factorial basis functions determined by X. A natural interpolant from trend filtering solution would be

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i h_i(x).$$

1.2 Review: Minimax rates for kth order TV classes in 1d

As mentioned above, a key focus of our work is to study minimax optimal rates for certain total variation type classes and analyzing estimators that achieve this optimal rate. We review the results for kth order TV classes in univariate setting. We need to introduce some notation to explain the results.

Notation. For deterministic sequences a_n, b_n we write $a_n = O(b_n)$ to denote that a_n/b_n is upper bounded for large enough n, and $a_n \asymp b_n$ to denote that both $a_n = O(b_n)$ and $a_n^{-1} = O(b_n^{-1})$. For random sequences A_n, B_n, we write $A_n = O_{\mathbb{P}}(B_n)$ to denote that A_n/B_n is bounded in probability. For integers $n \geq 0$, let $[n]$ denote the set of positive integers not larger than n.

Suppose the data is distributed as in (1.1) with $d = 1$, that is,

$$y_i \sim N(f(x_i), \sigma^2), \quad \text{independently, for } i = 1, \dots, n, \tag{1.11}$$

where $\sigma > 0$ and $x_i = i/n, i \in [n]$. Given an estimator $\hat{f}$ of f in (1.11), the quantity

$$\mathrm{MSE}(\hat{f}, f) = \frac{1}{n} \sum_{i=1}^{n} \big(\hat{f}(x_i) - f(x_i)\big)^2$$

is called the mean squared error (MSE) of $\hat{f}$; we will also call $\sup_{f \in \mathcal{F}} \mathbb{E}[\mathrm{MSE}(\hat{f}, f)]$ the risk of $\hat{f}$. The minimax risk and minimax linear risk over a class of functions $\mathcal{F}$ are

$$R(\mathcal{F}) = \inf_{\hat{\theta}} \sup_{f \in \mathcal{F}} \mathbb{E}\big[\mathrm{MSE}(\hat{f}, f)\big] \quad \text{and} \quad R_L(\mathcal{F}) = \inf_{\hat{f} \text{ linear}} \sup_{f \in \mathcal{F}} \mathbb{E}\big[\mathrm{MSE}(\hat{f}, f)\big]$$

where the infimum is taken over all estimators $\hat{f}$ for minimax risk and over only estimators linear in y for minimax linear risk.

The classical nonparametric statistics literature (Donoho et al. 1990, Donoho & Johnstone 1998, Mammen & van de Geer 1997) provides a more or less complete story for estimation under total variation constraints in 1d. See also Tibshirani (2014) for a translation of these results to a setting more consistent (notationally) to that in the current document. For the kth order TV space in (1.3), the results in Donoho & Johnstone (1998) imply that, for C_n that is not too small and not too large,

$$R(\mathcal{F}_k(C_n)) \asymp C_n^{\frac{2}{2k+3}} n^{-\frac{2k+2}{2k+3}}. \tag{1.12}$$

Further, Mammen & van de Geer (1997) showed that the locally adaptive regression splines $\hat{f}$ in (1.2) , with $\lambda \asymp C_n^{-(2k+1)/(2k+3)} n^{1/(2k+3)}$, satisfies

$$\mathrm{MSE}(\hat{f}, f) = O_{\mathbb{P}}(C_n^{\frac{2}{2k+3}} n^{-\frac{2k+2}{2k+3}}), \tag{1.13}$$

for all $f \in \mathcal{F}_k(C_n)$, and is thus minimax rate optimal over $\mathcal{F}_k(C_n)$. (In assessing rates here and throughout, we do not distinguish between convergence in expectation versus convergence in probability.) Wavelet denoising, under various choices of wavelet bases, also achieves the minimax rate. However, many simpler estimators do not. To be more precise, it is shown in Donoho & Johnstone (1998) that

$$R_L(\mathcal{F}_k(C_n)) \asymp C_n^{\frac{1}{k+1}} n^{-\frac{2k+1}{2k+2}}. \tag{1.14}$$

Therefore, a substantial number of commonly used nonparametric estimators—such as running mean estimators, smoothing splines, kernel smoothing, Laplacian smoothing, and Laplacian eigenmaps, which are all linear smoothers—have a major deficiency when it comes to estimating functions of bounded variation. Roughly speaking, they will require many more samples to estimate f within the same degree of accuracy as an optimal method like TV or wavelet denoising (on the order of $\epsilon^{-1/2}$ times more samples to achieve an MSE of ϵ in the TV denoising case $k = 0$). Further theory and empirical examples (e.g., Donoho & Johnstone (1994*b*, 1998), Tibshirani (2014)) offer the following perspective: linear smoothers

cannot cope with functions in $\mathcal{F}_k(C)$ that have spatially inhomogeneous smoothness, i.e., that vary smoothly at some locations and vary wildly at others. Linear smoothers can only produce estimates that are smooth throughout, or wiggly throughout, but not a mix of the two. They can hence perform well over smaller, more homogeneous function classes like Sobolev or Holder classes, but not larger ones like total variation classes (or more generally, Besov and Triebel classes), and for these, one must use more sophisticated, nonlinear techniques. A motivating question: does such a gap persist in higher dimensions, between optimal nonlinear and linear estimators, and if so, how big is it? How about the same question in additive models? We attempt to answer these questions in the rest of the document.

1.3 Preview, Summary and Outline

We extend the ideas developed in univariate trend filtering to multivariate data and hypothesis testing.

1.3.1 Trend filtering on Grids

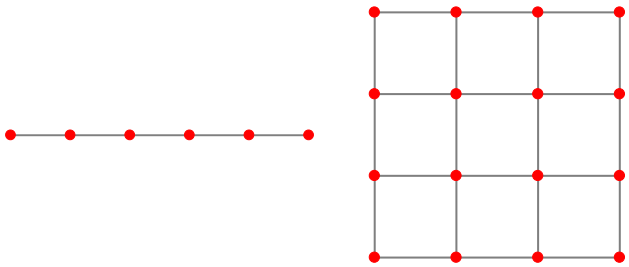

Figure 1.3: *Grid graphs in one dimension and two dimensions*

In our first extension, we estimate a mean parameter defined over the nodes of a d-dimensional grid graph $G = (V, E)$, with equal side lengths $N = n^{1/d}$; see Figure 1.3. Let us enumerate $V = \{1, \dots, n\}$ and $E = \{e_1, \dots, e_m\}$, and consider data $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ observed over V, distributed as

$$y_i \sim N(\theta_{0,i}, \sigma^2), \quad \text{independently, for } i \in [n], \tag{1.15}$$

where $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n}) \in \mathbb{R}^n$ is the mean parameter to be estimated, and $\sigma^2 > 0$ the common noise variance. We will assume that θ_0 displays some kind of regularity or smoothness over G, and are specifically interested in notions of regularity built around graph total variation. Image denoising problems can be directly formulated in this setup. In Figure 1.4, the left figure shows the familiar cameraman image with added Gaussian noise and the right figure shows the image obtained by graph TV denoising.

Although TV denoising has local adaptivity property, it exhibits "staircasing" artifacts, as it is restricted to fitting piecewise constant functions. See Figure 1.5. Higher-order TV regularization methods, which, loosely speaking, consider the TV of derivatives of the

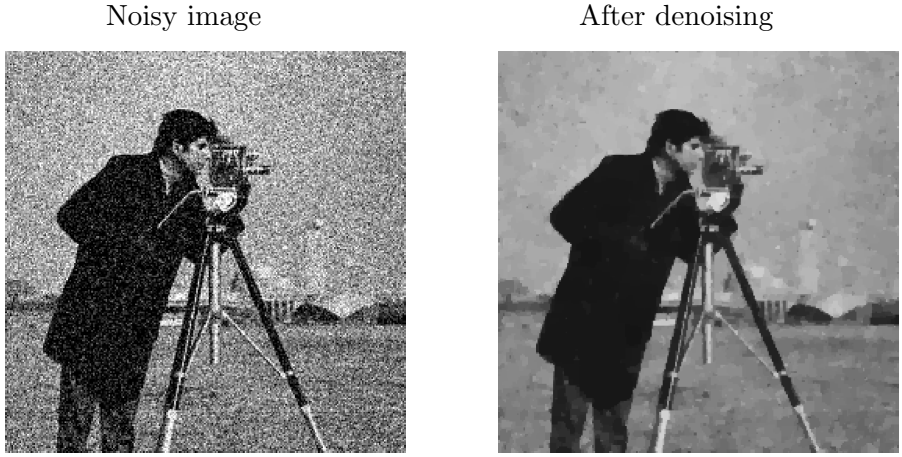

Figure 1.4: *Noisy "cameraman" image and its denoised version.*

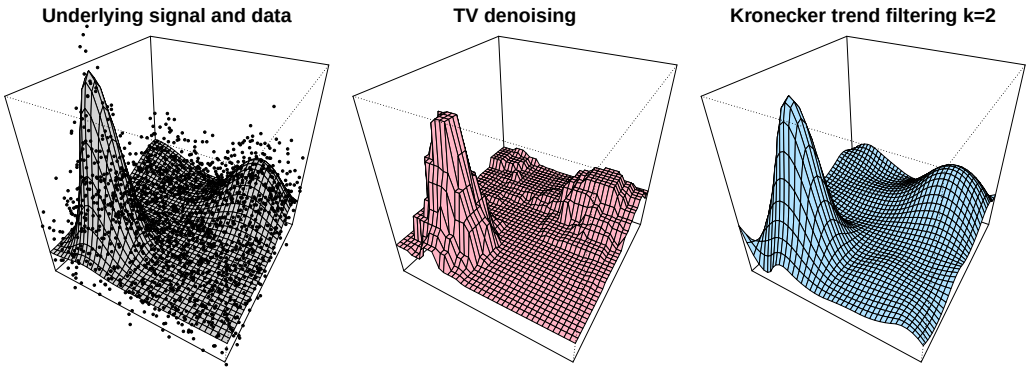

Figure 1.5: *Left: an underlying signal θ_0 and associated data y (shown as black points). Middle and right: TV denoising (2.3), and Kronecker trend filtering (2.7) of order $k = 2$ fit to y, respectively with penalty operator as described in Section 2.1.*

parameter avoid such artifacts while maintaining local adaptivity. We study two such estimators — Graph Trend Filtering (GTF) (Wang et al. 2016) and Kronecker Trend Filtering (KTF) (analyzed in Sadhanala et al. (2017)) — which are of the form

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|\Delta\theta\|_1,$$

for a matrix $\Delta \in \mathbb{R}^{r \times n}$, referred to as the penalty operator for an appropriate integer r. The penalty operator in the estimators captures different notions of kth order (discrete) TV on grids. In 1-dimension, kth order KTF reduces to trend filtering of the same order in (1.4). Under appropriate scaling of $\|\Delta\theta_0\|_1$ where θ_0 is the true signal, we show that these estimators achieve a minimax optimal rate of $n^{-(2k+2)/(d+\max\{d,2k+2\})}$ modulo $\log n$ factors on d-dimensional grids. Extending classic results in 1d from Donoho & Johnstone (1998), we show that no linear smoother can achieve this rate. Remarkably, in the low-smoothness regime of $2k + 2 \leq d$, we see that no linear smoother is consistent!

1.3.2 Additive model with Trend filtering

A classical way to tackle the curse of dimensionality is to consider an additive model for the regression function. We study additive models whose components are built from univariate trend filtering. In Figure 1.6, additive trend filtering displays better fitting of to inhomogeneous functions than additive smoothing splines; the details of the simulation are in Chapter 3.

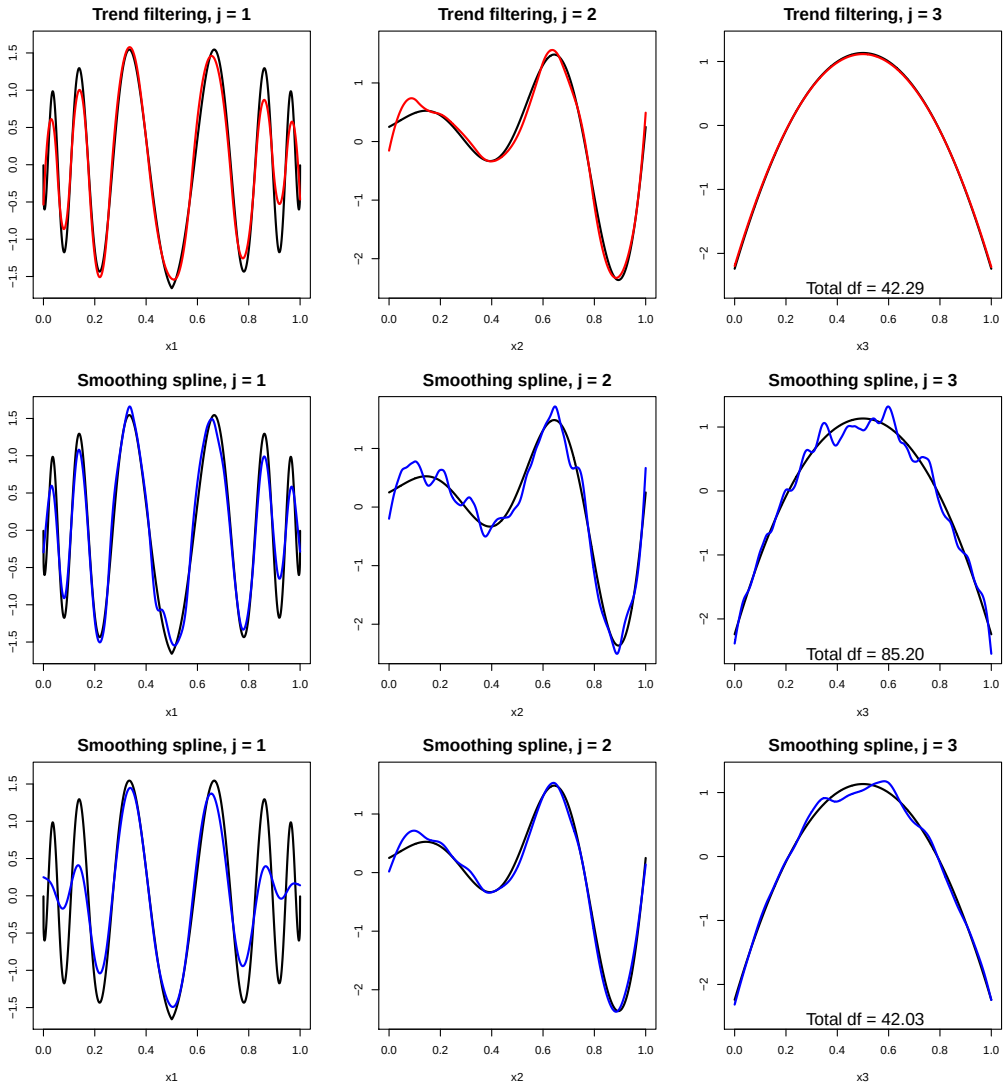

Figure 1.6: *Comparing estimates from additive trend filtering* (3.4) *(of quadratic order) and additive smoothing splines* (3.1) *(of cubic order), for a simulation with* $n = 3000$ *and* $d = 3$*, as described in Section 3.1.4. In each row, the underlying component functions are plotted in black.*

We study the properties of our additive trend filtering estimator

$$\hat{\theta} \in \underset{\theta_1,\ldots,\theta_d \in \mathbb{R}^n}{\operatorname{argmin}} \quad \frac{1}{2}\Big\|Y - \bar{Y}\mathbb{1} - \sum_{j=1}^{d}\theta_j\Big\|_2^2 + \lambda\sum_{j=1}^{d}\big\|D^{(X_j,k+1)}S_j\theta_j\big\|_1$$
$$\text{subject to} \quad \mathbb{1}^T\theta_j = 0, \quad j \in [d]$$

where $Y - \bar{Y}\mathbb{1}$ is the centered response vector, $\lambda \geq 0$ is a tuning regularization parameter, S_j is a permutation matrix such that it sorts the vector X_j in increasing order. In Chapter 3, we derive error bounds for this estimator. Assuming that the underlying regression function is additive, denoted by $f_0 = \sum_{j=1}^{d} f_{0j}$, and that $\mathrm{TV}(f_{0j}^{(k)})$ is bounded, for $j \in [d]$, we prove that the kth order additive trend filtering estimator converges to f_0 at the rate $dn^{-(2k+2)/(2k+3)}$. Note that this is d times the optimal univariate error rate. We prove that this rate is optimal in a minimax sense, and also show that additive models built from linear smoothers of any kind are suboptimal. Also, we devise a *new parallel backfitting algorithm* by looking at the alternating projections scheme in the *dual* of the additive trend filtering problem above.

1.3.3 Higher-order KS test

It is well-known that the general purpose classical KS test is not very sensitive to tail differences. To remedy this, we propose and study a *kth-order KS test* defined by the test statistic

$$\sup_{f \in \mathcal{F}_k}\left|\frac{1}{m}\sum_{i=1}^{m} f(X^i) - \frac{1}{n}\sum_{j=1}^{n} f(Y^j)\right|$$

where $X^1, \ldots, X^m \in \mathbb{R}$ and $Y^1, \ldots, Y^n \in \mathbb{R}$ are samples from two distributions and

$$\begin{aligned}\mathcal{F}_k = \big\{ f : \mathrm{TV}(f^{(k)}) \leq 1,& \\ f^{(j)}(0) = 0,\ j \in \{0\} \cup [k-1],& \\ f^{(k)}(0+) = 0 \text{ or } f^{(k)}(0-) = 0\big\}.&\end{aligned}$$

Here $f^{(k)}(0+)$ and $f^{(k)}(0-)$ denote one-sided limits at 0 from above and below, respectively. The zero derivative conditions at 0 ensure that the functions in $\mathcal{F}_k$ do not grow faster than $x \mapsto |x|^k/k!$. With $k = 0$, it is well-known that this test statistic reduces to the KS statistic.

Time complexity for computing the statistic is $O((m+n)\log(m+n))$ for $k \leq 5$ – same as the time complexity for sorting the joint sample. For $k \geq 6$, we can approximate the statistic to ϵ accuracy in an additional $O((m+n)\log\frac{1}{\epsilon})$ time. We derive asymptotic null distribution of the test statistic and also concentration bounds on test statistic in the alternative case. We empirically show that the test has superior power compared to KS and other familiar two-sample tests in some cases with heavy tails.

1.3.4 Outline

We discuss the problem of nonparametric regression on grids in Chapter 2. We define the total variation based estimators and show our results from Sadhanala et al. (2016) and Sadhanala et al. (2017) and also a few new results. The additive model with trend filtering

is discussed in Chapter 3. We state the results from Sadhanala & Tibshirani (2017) and discuss some natural ways of extending the work. In Chapter 4, we discuss the higher-order extension to Kolmogorov-Smirnov two sample test from Sadhanala et al. (2019). Most of the proofs and other details are given in appendices.

Chapter 2

Trend filtering on Grids

In this chapter, we focus on estimation of a mean parameter defined over the nodes of a d-dimensional grid graph $G = (V, E)$, with equal side lengths $N = n^{1/d}$. Let us enumerate $V = \{1, \dots, n\}$ and $E = \{e_1, \dots, e_m\}$, and consider data $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ observed over V, distributed as

$$y_i \sim N(\theta_{0,i}, \sigma^2), \quad \text{independently, for } i = 1, \dots, n, \tag{2.1}$$

where $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n}) \in \mathbb{R}^n$ is the mean parameter to be estimated, and $\sigma^2 > 0$ the common noise variance. We will assume that θ_0 displays some kind of regularity or smoothness over G, and are specifically interested in notions of regularity built around on the *total variation (TV)* operator

$$\|D\theta\|_1 = \sum_{(i,j)\in E} |\theta_i - \theta_j|, \tag{2.2}$$

defined with respect to G, where $D \in \mathbb{R}^{m\times n}$ is the edge incidence matrix of G, which has ℓth row $D_\ell = (0, \dots, -1, \dots, 1, \dots, 0)$, with -1 in location i and 1 in location j, provided that the ℓth edge is $e_\ell = (i, j)$ with $i < j$. There is an extensive literature on estimators based on TV regularization, both in Euclidean spaces and over graphs. Higher-order TV regularization, which, loosely speaking, considers the TV of derivatives of the parameter, is much less understood, especially over graphs. We develop statistical theory for higher-order TV smoothness classes, and we analyze associated trend filtering methods, which are seen to achieve the minimax optimal estimation error rate over such classes.

Motivation. TV denoising over grid graphs, specifically 1d and 2d grid graphs, is a well-studied problem in signal processing, statistics, and machine learning, some key references being Rudin et al. (1992), Chambolle & Lions (1997), Tibshirani et al. (2005). Given data $y \in \mathbb{R}^n$ as per the setup described above, the *TV denoising* or *fused lasso* estimator over the grid G is defined as

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^n} \frac{1}{2}\|y - \theta\|_2^2 + \lambda\|D\theta\|_1, \tag{2.3}$$

where $\lambda \geq 0$ is a tuning parameter. The TV denoising estimator generalizes seamlessly to arbitrary graphs. The problem of denoising over grids, the setting we focus on, is of

particular relevance to a number of important applications, e.g., in time series analysis, and image and video processing.

A strength of the nonlinear TV denoising estimator in (2.3)—where by "nonlinear", we mean that $\hat{\theta}$ is nonlinear as a function of y—is that it can adapt to heterogeneity in the local level of smoothness of the underlying signal θ_0. Moreover, it adapts to such heterogeneity at an extent that is beyond what linear estimators are capable of capturing. This principle is widely evident in practice and has been championed by many authors in the signal processing literature. It is also backed by statistical theory, i.e., Donoho & Johnstone (1998), Mammen & van de Geer (1997), Tibshirani (2014) in the 1d setting.

Another pair of methods that we refer to are *Laplacian smoothing* and *Laplacian eigenmaps*, which are most commonly seen in the context of clustering, dimensionality reduction, and semi-supervised learning, but are also useful tools for estimation in a regression setting like ours (e.g., Belkin & Niyogi (2002, 2003), Smola & Kondor (2003), Zhu et al. (2003), Belkin & Niyogi (2004), Zhou et al. (2005), Belkin et al. (2005), Belkin & Niyogi (2005), Ando & Zhang (2006), Sharpnack & Singh (2010)). The Laplacian smoothing estimator is given by

$$\hat{\theta}^{\mathrm{LS}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \|y - \theta\|_2^2 + \lambda \|D\theta\|_2^2, \quad \text{i.e.,} \quad \hat{\theta}^{\mathrm{LS}} = (I + \lambda L)^{-1} y, \tag{2.4}$$

for a tuning parameter $\lambda \geq 0$, where in the second expression we have written $\hat{\theta}^{\mathrm{LS}}$ in closed-form (this is possible since it is the minimizer of a convex quadratic). For Laplacian eigenmaps, we must introduce the eigendecomposition of the graph Laplacian, $L = V \Sigma V^T$, where $\Sigma = \mathrm{diag}(\rho_1, \ldots, \rho_n)$ with $0 = \rho_1 < \rho_2 \leq \ldots \leq \rho_n$, and where $V = [V_1, V_2, \ldots, V_n] \in \mathbb{R}^{n \times n}$ has orthonormal columns. The Laplacian eigenmaps estimator is

$$\hat{\theta}^{\mathrm{LE}} = V_{[k]} V_{[k]}^T y, \quad \text{where} \quad V_{[k]} = [V_1, V_2, \ldots, V_k] \in \mathbb{R}^{n \times k}, \tag{2.5}$$

where now $k \in \{1, \ldots, n\}$ acts as a tuning parameter.

Linear smoothers such as Laplacian smoothing and Laplacian eigenmaps are appealing because they are (relatively) simple: they are just linear transformations of the data y. Indeed, as we are considering G to be a grid, both estimators in (2.4), (2.5) can be computed very quickly, in nearly $O(n)$ time, since the columns of V here are discrete cosine transform (DCT) basis vectors when $d = 1$, or Kronecker products thereof, when $d \geq 2$ (e.g., Conte & de Boor (1980), Godunov & Ryabenkii (1987), Kunsch (1994), Ng et al. (1999), Wang et al. (2008)). The TV denoising estimator in (2.3), on the other hand, cannot be expressed in closed-form, and is much more difficult to compute, especially when $d \geq 2$, though several advances have been made over the years (see the references above, and in particular Barbero & Sra (2018) for an efficient operator-splitting algorithm and nice literature survey). Importantly, these computational difficulties are often worth it: TV denoising often practically outperforms ℓ_2-regularized estimators like Laplacian smoothing (and also Laplacian eigenmaps) in image denoising tasks, as it is able to better preserve sharp edges and object boundaries (this is now widely accepted, early references are, e.g., Acar & Vogel (1994), Dobson & Santosa (1996), Chambolle & Lions (1997)). See Figure 2.1 for an example, using the often-studied "cameraman" image.

In the 1d setting, classical theory from nonparametric statistics draws a clear distinction between the performance of TV denoising and estimators like Laplacian smoothing and Laplacian eigenmaps. Perhaps surprisingly, this theory has not yet been fully developed

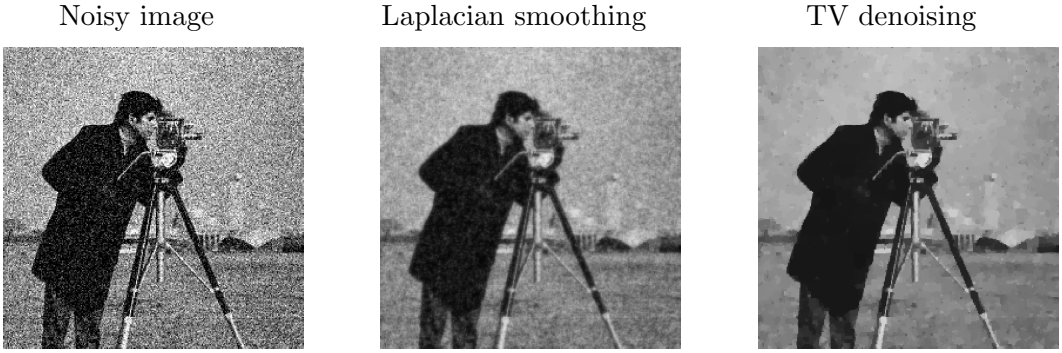

Figure 2.1: *Comparison of Laplacian smoothing and TV denoising for the common "cameraman" image. TV denoising provides a more visually appealing result, and also achieves about a 35% reduction in MSE compared to Laplacian smoothing (MSE being measured to the original image). Both methods were tuned optimally.*

in dimensions $d \geq 2$. Arguably, the comparison between TV denoising and Laplacian smoothing and Laplacian eigenmaps is even more interesting in higher dimensions, because the computational gap between the methods is even larger (the former method being much more expensive, say in 2d and 3d, than the latter two). Shortly, we review the 1d theory, and what is known in d-dimensions, for $d \geq 2$.

Note that the TV denoising estimator $\hat{\theta}$ in (2.3) takes a *piecewise constant* structure by design, i.e., at many adjacent pairs $(i, j) \in E$ we will have $\hat{\theta}_i = \hat{\theta}_j$, and this will be generally more common for larger λ. For some problems, this structure may not be ideal and we might instead seek a *piecewise smooth* estimator, that is still able to cope with local changes in the underlying level of smoothness, but offers a richer structure (beyond a simple constant structure) for the base trend. In a 1d setting, this is accomplished by trend filtering methods, which move from piecewise constant to *piecewise polynomial* structure, via TV regularization of discrete derivatives of the parameter Steidl et al. (2006), Kim et al. (2009), Tibshirani (2014). An extension of trend filtering to general graphs was developed in Wang et al. (2016). In what follows, we study the statistical properties of this graph trend filtering method over grids, and we propose and analyze a more specialized trend filtering estimator for grids based on the idea that something like a Euclidean coordinate system is available at any (interior) node. See Figure 2.2 for a motivating illustration.

Related work. The literature on TV denoising is enormous and we cannot give a comprehensive review, but only some brief highlights. Important methodological and computational contributions are found in Rudin et al. (1992), Chambolle & Lions (1997), Tibshirani et al. (2005), Chambolle & Darbon (2009), Hoefling (2010), Chambolle & Pock (2011), Tibshirani & Taylor (2011), Kovac & Smith (2011), Condat (2012), Johnson (2013), Barbero & Sra (2018), Tansey & Scott (2015), and notable theoretical contributions are found in Mammen & van de Geer (1997), Rinaldo (2009), Harchaoui & Levy-Leduc (2010), Sharpnack et al. (2012), Hutter & Rigollet (2016), Padilla et al. (2016). The literature on higher-order TV-based methods is more sparse and more concentrated on the 1d setting. Trend filtering methods in 1d were pioneered in Steidl et al. (2006), Kim et al. (2009), and analyzed statistically in Tibshirani (2014), where they were also shown to be asymptotically equivalent

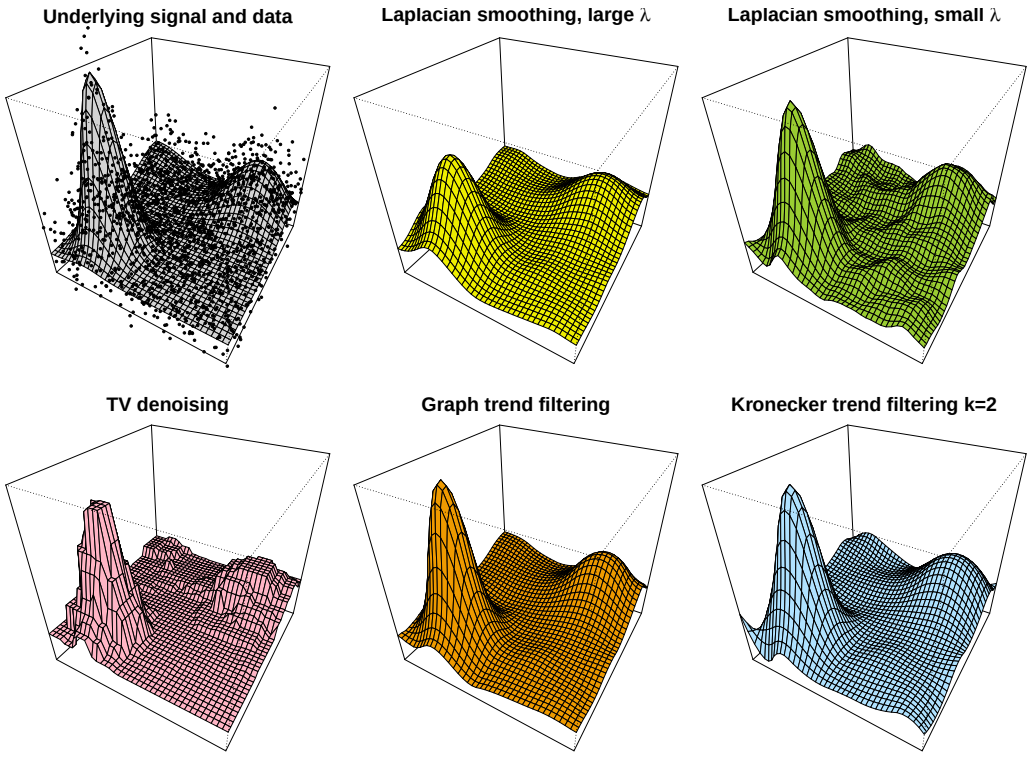

Figure 2.2: *Top left: an underlying signal θ_0 and associated data y (shown as black points). Top middle and top right: Laplacian smoothing fit to y, at large and small tuning parameter values, respectively. Bottom left, middle, and right: TV denoising (2.3), graph trend filtering (2.7), and Kronecker trend filtering (2.7) fit to y, respectively (the latter two are of order $k = 2$, with penalty operators as described in Section 2.1). In order to capture the larger of the two peaks, Laplacian smoothing must significantly undersmooth throughout; with more regularization, it oversmooths throughout. TV denoising is able to adapt to heterogeneity in the smoothness of the underlying signal, but exhibits "staircasing" artifacts, as it is restricted to fitting piecewise constant functions. Graph and Kronecker trend filtering overcome this, while maintaining local adaptivity.*

to locally adaptive regression splines of Mammen & van de Geer (1997). A generalization of trend filtering that operates over an arbitrary graph structure was given in Wang et al. (2016). Trend filtering is not the only avenue for higher-order TV regularization: the signal processing community has also studied higher-order variants of TV, see, e.g., Poschl & Scherzer (2008), Bredies et al. (2010). The construction of the discrete versions of these higher-order TV operators is somewhat similar to that in Wang et al. (2016) as well our Kronecker trend filtering proposal, however, the focus of the work is quite different.

Summary of contributions. An overview of our contributions is given below.

- We propose a new method for trend filtering over grid graphs that we call *Kronecker trend filtering* (KTF), and compare its properties to the more general graph trend filtering (GTF) proposal of Wang et al. (2016).

- For d-dimensional grids, we derive estimation error rates for GTF and KTF, each of these rates being a function of the regularizer evaluated at the mean θ_0.

- Again for d-dimensional grids, we derive minimax lower bounds for estimation over two higher-order TV classes defined using the operators from GTF and KTF. These lower bounds match the upper bounds in rate (apart from log factors) derived for GTF and KTF, ensuring that each method is minimax rate optimal (modulo log factors) for its own notion of regularity. Also, the KTF class contains a Holder class of an appropriate order, and KTF is seen to be rate optimal (modulo log factors) for this more homogeneous class as well.

- We also derive minimax linear rates over these higher-order TV classes and show that linear smoothers cannot achieve the minimax optimal rate when $2k + 2 < d$.

Notation. Given a d-dimensional grid $G = (V, E)$, where $V = \{1, \dots, n\}$, as before, we will sometimes index a parameter $\theta \in \mathbb{R}^n$ defined over the nodes in the following convenient way. Letting $N = n^{1/d}$ and $Z_d = \{(i_1/N, \dots, i_d/N) : i_1, \dots, i_d \in \{1, \dots, N\}\} \subseteq [0, 1]^d$, we will index the components of θ by their lattice positions, denoted $\theta(x)$, $x \in Z_d$. Further, for each $j = 1, \dots, d$, we will define the discrete derivative of θ in the jth coordinate direction at a location x by

$$(D_{x_j}\theta)(x) = \begin{cases} \theta(x + e_j/N) - \theta(x) & \text{if } x, x + e_j/N \in Z_d, \\ 0 & \text{else.} \end{cases} \tag{2.6}$$

Naturally, we denote by $D_{x_j}\theta \in \mathbb{R}^n$ the vector with components $(D_{x_j}\theta)(x)$, $x \in Z_d$. Higher-order discrete derivatives are simply defined by repeated application of the above definition. We use abbreviations

$$(D_{x_j^2}\theta)(x) = (D_{x_j}(D_{x_j}\theta))(x) \text{ for } j \in [d],$$
$$(D_{x_j, x_\ell}\theta)(x) = (D_{x_j}(D_{x_\ell}\theta))(x) \text{ for } j, \ell \in [d],$$

and so on.

Given an estimator $\hat{\theta}$ of the mean parameter θ_0 in (2.1), and $\mathcal{K} \subseteq \mathbb{R}^n$, the quantity

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = \frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2$$

is called the mean squared error (MSE) of θ; we will also call $\mathbb{E}[\mathrm{MSE}(\hat{\theta}, \theta_0)]$ the risk of $\hat{\theta}$. The minimax risk and minimax linear risk over $\mathcal{K}$ are

$$R(\mathcal{K}) = \inf_{\hat{\theta}} \sup_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big] \quad \text{and} \quad R_L(\mathcal{K}) = \inf_{\hat{\theta} \text{ linear}} \sup_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big]$$

where the infimum is taken over all estimators $\hat{\theta}$ for minimax risk and over only estimators linear in y for minimax linear risk.

2.1 Trend filtering methods

Graph trend filtering. To review the family of estimators developed in Wang et al. (2016), we start by introducing a general-form estimator called the *generalized lasso* signal approximator Tibshirani & Taylor (2011),

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|\Delta \theta\|_1, \tag{2.7}$$

for a matrix $\Delta \in \mathbb{R}^{r \times n}$, referred to as the penalty operator. For an integer $k \geq 0$, Wang et al. (2016) defined the *graph trend filtering* (GTF) estimator of order k by (2.7), with the penalty operator being

$$\Delta^{(k+1)} = \begin{cases} DL^{k/2} & \text{for } k \text{ even}, \\ L^{(k+1)/2} & \text{for } k \text{ odd}. \end{cases} \tag{2.8}$$

Here, as before, we use D for the edge incidence matrix of G. We also use $L = D^T D$ for the graph Laplacian matrix of G. The intuition behind the above definition is that $\Delta^{(k+1)}\theta$ gives something roughly like the $(k+1)$st order discrete derivatives of θ over the graph G.

Note that the GTF estimator reduces to TV denoising in (2.3) when $k = 0$. For any signal $\theta \in \mathbb{R}^n$, we can write $\|\Delta^{(k+1)}\theta\|_1 = \sum_{x \in Z_d} d_x$, where at all points $x \in Z_d$ (except for those close to the boundary),

$$d_x = \begin{cases} \displaystyle\sum_{j_1=1}^{d} \Big| \sum_{j_2,\ldots,j_q=1}^{d} \Big(D_{x_{j_1}, x_{j_2}^2, \ldots, x_{j_q}^2} \theta \Big)(x) \Big| & \text{for } k \text{ even, where } q = k/2, \\ \displaystyle\Big| \sum_{j_1,\ldots,j_q=1}^{d} \Big(D_{x_{j_1}^2, x_{j_2}^2, \ldots, x_{j_q}^2} \theta \Big)(x) \Big| & \text{for } k \text{ odd, where } q = (k+1)/2. \end{cases} \tag{2.9}$$

Written in this form, it appears that the GTF operator $\Delta^{(k+1)}$ aggregates derivatives in somewhat of an unnatural way. But we must remember that for a general graph structure, only first derivatives and divergences have obvious discrete analogs—given by application of D and L, respectively.

Kronecker trend filtering. There is a natural alternative to the GTF penalty operator that takes advantage of the Euclidean-like structure available at the (interior) nodes of a grid graph. At a point $x \in Z_d$ (not close to the boundary), consider using

$$d_x = \sum_{j=1}^{d} \big| \big(D_{x_j^{k+1}} \theta \big)(x) \big| \tag{2.10}$$

as a basic building block for penalizing derivatives, rather than (2.9). This gives rise to a method we call *Kronecker trend filtering* (KTF), which for an integer order $k \geq 0$ is defined by (2.7), but now with the choice of penalty operator

$$\widetilde{\Delta}^{(k+1)} = \begin{bmatrix} D_{\text{1d}}^{(k+1)} \otimes I \otimes \cdots \otimes I \\ I \otimes D_{\text{1d}}^{(k+1)} \otimes \cdots \otimes I \\ \vdots \\ I \otimes I \otimes \cdots \otimes D_{\text{1d}}^{(k+1)} \end{bmatrix}. \tag{2.11}$$

Here, $D_{\text{1d}}^{(k+1)} \in \mathbb{R}^{(N-k-1)\times N}$ is the 1d discrete derivative operator of order $k+1$, $I \in \mathbb{R}^{N\times N}$ is the identity matrix, and $A \otimes B$ is the Kronecker product of matrices A, B. Each group of rows in (2.11) features a total of $d-1$ Kronecker products.

KTF and GTF can be solved using a standard quadratic programming solver on the dual of (2.7) with appropriate penalty operators. The regularization parameter λ is chosen by cross-validation in practice. KTF reduces to TV denoising in (2.3) when $k = 0$, and thus also to GTF with $k = 0$. But for $k \geq 1$, GTF and KTF are different estimators. A look at the action of their penalty operators, as displayed in (2.9), (2.10) reveals some of their differences. For example, we see that GTF considers mixed derivatives of total order $k+1$, but KTF only considers directional derivatives of order $k+1$ that are parallel to the coordinate axes. Also, GTF penalizes sums of derivatives, whereas KTF penalizes individual ones.

More differences between GTF and KTF have to do with the structure of their estimates, as we discuss next. Another subtle difference lies in how the GTF and KTF operators (2.8), (2.11) relate to more classical notions of smoothness, particularly, Holder smoothness. This is covered in Section 2.4.

2.2 Structure of estimates

It is straightforward to see that the GTF operator (2.8) has a 1-dimensional null space, spanned by $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^n$. This means that GTF lets constant signals pass through unpenalized, but nothing else; or, in other words, it preserves the projection of y onto the space of constant signals, $\bar{y}\mathbb{1}$, but nothing else. The KTF operator, meanwhile, has a much richer null space.

Lemma 2.1. The null space of the KTF operator (2.11) has dimension $(k+1)^d$, and it is spanned by a polynomial basis made up of elements

$$p(x) = x_1^{a_1} x_2^{a_2} \cdots x_d^{a_d}, \quad x \in Z_d,$$

where $a_1, \ldots, a_d \in \{0, \ldots, k\}$.

The lemma shows that KTF preserves the projection of y onto the space of polynomials of max degree k, i.e., lets much more than just constant signals pass through unpenalized. The proofs of all results in this chapter including this lemma are in Appendix A.

Beyond the differences in these base trends (represented by their null spaces), GTF and KTF admit estimates with similar but generally different structures. KTF has the advantage that this structure is more transparent: its estimates are piecewise polynomial functions of max degree k, with generally fewer pieces for larger λ. This is demonstrated by a functional representation for KTF, given next.

Lemma 2.2. Let $h_i : [0,1] \to \mathbb{R}$, $i = 1, \ldots, N$ be the (univariate) falling factorial functions (1.7) of order k, defined over knots $1/N, 2/N, \ldots, 1$. Let H_d be the space spanned by all d-wise tensor products of falling factorial functions, i.e., H_d contains $f : [0,1]^d \to \mathbb{R}$ of the form

$$f(x) = \sum_{i_1,\ldots,i_d=1}^{N} \alpha_{i_1,\ldots,i_d} h_{i_1}(x) h_{i_2}(x_2) \cdots h_{i_d}(x_d), \quad x \in [0,1]^d,$$

for coefficients $\alpha \in \mathbb{R}^n$ (whose components we index by $\alpha_{i_1,\ldots,i_d}$, for $i_1, \ldots, i_d = 1, \ldots, N$). Then the KTF estimator defined in (2.7), (2.11) is equivalent to the optimization problem

$$\hat{f} = \operatorname*{argmin}_{f \in H_d} \frac{1}{2} \sum_{x \in Z_d} \big(y(x) - f(x)\big)^2 + \lambda \sum_{j=1}^{d} \sum_{x_{-j} \in Z_{d-1}} \mathrm{TV}\bigg(\frac{\partial^k f(\cdot, x_{-j})}{\partial x_j^k}\bigg), \tag{2.12}$$

where $f(\cdot, x_{-j})$ denotes f as function of the jth dimension with all other dimensions fixed at x_{-j}, $\partial^k / \partial x_j^k(\cdot)$ denotes the kth partial weak derivative operator with respect to x_j, for $j = 1, \ldots, d$. The discrete (2.7), (2.11) and functional (2.12) representations are equivalent in that $\hat{f}$ and $\hat{\theta}$ match at all grid locations $x \in Z_d$.

Aside from shedding light on the structure of KTF solutions, the functional optimization problem in (2.12) is of practical importance: the function $\hat{f}$ is defined over all of $[0,1]^d$ (as opposed to $\hat{\theta}$, which is of course only defined on the grid Z_d) and thus we may use it to interpolate the KTF estimate to non-grid locations. It is not clear to us that a functional representation as in (2.12) (or even a sensible interpolation strategy) is available for GTF on d-dimensional grids.

2.3 Upper bounds on estimation error

In this section, we derive upper bounds on the estimation error of GTF and KTF for d-dimensional grids where $d \geq 2$. Upper bounds for generalized lasso estimators were studied in Wang et al. (2016), and we will leverage one of their key results, which is based on what these authors call *incoherence* of the left singular vectors of the penalty operator Δ. A slightly refined version of this result is stated below.

Theorem 2.1 (Theorem 6 in Wang et al. (2016))**.** Suppose that $\Delta \in \mathbb{R}^{r \times n}$ has rank q, and denote by $\xi_1 \leq \ldots \leq \xi_q$ its nonzero singular values. Also let $u_1, \ldots, u_q$ be the corresponding left singular vectors. Assume that these vectors, except for the first i_0, are incoherent, meaning that for a constant $\mu \geq 1$,

$$\|u_i\|_\infty \leq \mu / \sqrt{n}, \quad i = i_0 + 1, \ldots, q,$$

Then for $\lambda \asymp \mu \sqrt{(\log r / n) \sum_{i=i_0+1}^{q} \xi_i^{-2}}$, the generalized lasso estimator $\hat{\theta}$ in (2.7) satisfies

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = O_{\mathbb{P}}\bigg(\frac{\mathrm{nullity}(\Delta)}{n} + \frac{i_0}{n} + \frac{\mu}{n} \sqrt{\frac{\log r}{n} \sum_{i=i_0+1}^{q} \frac{1}{\xi_i^2}} \cdot \|\Delta \theta_0\|_1 \bigg).$$

For GTF and KTF, we will apply this result, balancing an appropriate choice of i_0 with the partial sum of reciprocal squared singular values $\sum_{i=i_0+1}^{q} \xi_i^{-2}$. The main challenge is in establishing incoherence of the singular vectors.

2.3.1 Error bounds for graph trend filtering

The authors in Wang et al. (2016) have already used Theorem 2.1 (their Theorem 6) in order to derive error rates for GTF on 2d grids. However, their results (specifically, their

Corollary 8) can be refined using a tighter upper bound for the partial sum term $\sum_{i=i_0+1}^{q} \xi_i^{-2}$ as we show in Sadhanala et al. (2017). We give a more general error rate that applies that applies to all $d \geq 2$ and all $k \geq 0$. No real further tightening is possible, since, as we show later, the results below match the minimax lower bound in rate, up to log factors.

Theorem 2.2. Assume that $d \geq 1$. For non-negative integers k, denote $C_n = \|\Delta^{(k+1)}\theta_0\|_1$ where $\Delta^{(k+1)}$ is the GTF operator defined in (2.8). Then GTF estimator in (2.7), (2.8) satisfies

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\lambda}{n} C_n\right).$$

with

$$\lambda \asymp \begin{cases} \sqrt{\log n} & 2k+2 < d \\ \log n & 2k+2 = d \\ (\log n)^{\frac{d}{2k+2+d}} \left(\frac{n}{C_n}\right)^{\frac{2k+2-d}{2k+2+d}} & 2k+2 > d. \end{cases}$$

The result for the TV denoising case $d \geq 2, k = 0$ in Theorem 2.2 was already established by Hutter & Rigollet (2016). In Sadhanala et al. (2017) we establish the result for the case $d = 2, k \geq 1$. The above result is general, in the sense that it is applicable for all $d \geq 1, k \geq 0$. However, in the 1d case, the above bound is weaker than known results from Mammen & van de Geer (1997), Tibshirani (2014) by a factor of $(\log n)^{\frac{1}{2k+3}}$.

Remark 2.1. It is interesting to note that the case $2k+2 \leq d$ appears to be quite special, in that the GTF estimator is adaptive to the underlying smoothness parameter C_n (the prescribed choice of tuning parameter $\lambda \asymp \sqrt{\log n}$ when $2k+2 < d$ and $\lambda \asymp \log n$ when $2k+2 = d$ does not depend on C_n).

With canonical scaling of C_n, we see the following error bound.

Corollary 2.1. With canonical scaling $C_n = n^{1-\frac{k+1}{d}}$, the GTF estimator with λ scaling as in Theorem 2.2 satisfies

$$\sup_{\theta_0 \in \mathcal{T}_d^k(C_n)} \mathrm{MSE}(\hat{\theta}, \theta_0) = \begin{cases} O_{\mathbb{P}}\left(n^{-\frac{k+1}{d}} \sqrt{\log n}\right) & 2k+2 < d \\ O_{\mathbb{P}}\left(n^{-\frac{k+1}{d}} \log n\right) & 2k+2 = d \\ O_{\mathbb{P}}\left(n^{-\frac{2k+2}{2k+2+d}} (\log n)^{\frac{d}{2k+2+d}}\right) & 2k+2 > d. \end{cases}$$

The technique for upper bounding $\sum_{i=i_0+1}^{q} \xi_i^{-2}$ in the proof of Theorem 2.2 can be roughly explained as follows. The GTF operator $\Delta^{(k+1)}$ on a d-dimensional grid has squared singular values:

$$\left(\sum_{j=1}^{d} 4 \sin^2 \frac{\pi(i_j - 1)}{2N}\right)^{k+1}, \quad i_1, \ldots, i_d \in [N].$$

We can upper bound the sum of squared reciprocal singular values with an integral over $[0,1]^2$, make use of the identity $\sin x \geq x/2$ for small enough x, and then switch to polar coordinates to calculate the integral (similar to Hutter & Rigollet (2016), in analyzing TV denoising). The arguments to verify incoherence of the left singular vectors of $\Delta^{(k+1)}$ are themselves somewhat delicate, but were already given in Wang et al. (2016) in the case of 2d grids. We generalize this incoherence result for $d \geq 3$. For details, see Appendix A.3.1.

2.3.2 Error bounds for Kronecker trend filtering

In comparison to the GTF case, the application of Theorem 2.1 to KTF is a much more difficult task, because (to the best of our knowledge) the KTF operator $\widetilde{\Delta}^{(k+1)}$ does not admit closed-form expressions for its singular values and vectors. This is true in any dimension (i.e., even for $d = 1$, where KTF reduces to univariate trend filtering). As it turns out, the singular values can be handled with a relatively straightforward application of the Cauchy interlacing theorem. It is establishing the incoherence of the singular vectors that proves to be the real challenge. This is accomplished by leveraging specialized approximation bounds for the eigenvectors of Toeplitz matrices from Bogoya et al. (2016).

Theorem 2.3. Assume that $d \geq 1$. For non-negative integers k, denote $C_n = \|\widetilde{\Delta}^{(k+1)}\theta_0\|_1$. where $\widetilde{\Delta}^{(k+1)}$ is the KTF operator defined in (2.11). Then KTF estimator in (2.7), (2.11) satisfies

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = O_{\mathbb{P}}\bigg(\frac{1}{n} + \frac{\lambda}{n} C_n\bigg).$$

with

$$\lambda \asymp \begin{cases} \sqrt{\log n} & 2k+2 < d \\ \log n & 2k+2 = d \\ (\log n)^{\frac{d}{2k+2+d}} \big(\frac{n}{C_n}\big)^{\frac{2k+2-d}{2k+2+d}} & 2k+2 > d. \end{cases}$$

Again, with canonical scaling of C_n, the error bound is as follows.

Corollary 2.2. With canonical scaling $C_n = n^{1-\frac{k+1}{d}}$, the KTF estimator with λ scaling as in Theorem 2.3 satisfies

$$\sup_{\theta_0 \in \mathcal{T}_k^n(C_n)} \mathrm{MSE}(\hat{\theta}, \theta_0) = \begin{cases} O_{\mathbb{P}}\Big(n^{-\frac{k+1}{d}} \sqrt{\log n}\Big) & 2k+2 < d \\ O_{\mathbb{P}}\Big(n^{-\frac{k+1}{d}} \log n\Big) & 2k+2 = d \\ O_{\mathbb{P}}\Big(n^{-\frac{2k+2}{2k+2+d}} (\log n)^{\frac{d}{2k+2+d}}\Big) & 2k+2 > d. \end{cases}$$

The proof of Theorem 2.3 is in Appendix A.3.2. The results in Theorems 2.2 and 2.3 match, in terms of their dependence on n, C_n. As we will see in the next section, the smoothness classes defined by the GTF and KTF operators are similar, though not exactly the same, and each GTF and KTF is minimax rate optimal with respect to its own smoothness class, up to log factors. The remarks about GTF following Theorem 2.2 are applicable to KTF as well, so we do not repeat them here.

2.4 Lower bounds on estimation error

We present lower bounds on the minimax estimation error over smoothness classes defined by the operators from GTF (2.8) and KTF (2.11), denoted

$$\mathcal{T}_d^k(C_n) = \{\theta \in \mathbb{R}^n : \|\Delta^{(k+1)}\theta\|_1 \leq C_n\}, \tag{2.13}$$

$$\widetilde{\mathcal{T}}_d^k(C_n) = \{\theta \in \mathbb{R}^n : \|\widetilde{\Delta}^{(k+1)}\theta\|_1 \leq C_n\}, \tag{2.14}$$

respectively (where the subscripts mark the dependence on the dimension d of the underlying grid graph). Before we derive such lower bounds, we examine embeddings of (the discretization of) the class of Holder smooth functions into the GTF and KTF classes, both to understand the nature of these new classes, and to define what we call a "canonical" scaling for the radius parameter C_n.

Embedding of Holder spaces and canonical scaling. Given an integer $k \geq 0$ and $L > 0$, recall that the *Holder class* $H(k+1, L; [0,1]^d)$ contains k times differentiable functions $f : [0,1]^d \to \mathbb{R}$, such that for all integers $\alpha_1, \ldots, \alpha_d \geq 0$ with $\alpha_1 + \cdots + \alpha_d = k$,

$$\left| \frac{\partial^k f(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} - \frac{\partial^k f(z)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right| \leq L \|x - z\|_2, \quad \text{for all } x, z \in [0,1]^d.$$

To compare Holder smoothness with the GTF and KTF classes defined in (2.13), (2.14), we discretize the class $H(k+1, L; [0,1]^d)$ by considering function evaluations over the grid Z_d, defining

$$\mathcal{H}_d^{k+1}(L) = \big\{\theta \in \mathbb{R}^n : \theta(x) = f(x),\ x \in Z_d,\ \text{for some } f \in H(k+1, L; [0,1]^d)\big\}. \tag{2.15}$$

Now we examine how the (discretized) Holder class in (2.15) compares to the GTF and KTF classes in (2.13), (2.14).

Beginning with a comparison to KTF, fix $\theta \in \mathcal{H}_d^{k+1}(L)$, corresponding to evaluations of $f \in H(k+1, L; [0,1]^d)$, and consider a point $x \in Z_d$ that is away from the boundary. Then the KTF penalty at x is

$$\begin{aligned} \big|\big(D_{x_j^{k+1}}\theta\big)(x)\big| &= \big|\big(D_{x_j^k}\theta\big)(x + e_j/N) - \big(D_{x_j^k}\theta\big)(x)\big| \\ &\leq N^k \left| \frac{\partial^k}{\partial x_j^k} f(x + e_j/N) - \frac{\partial^k}{\partial x_j^k} f(x) \right| + N^k \delta(N) \\ &\leq LN^{k-1} + cLN^{k-1}. \end{aligned} \tag{2.16}$$

In the second line above, we define $\delta(N)$ to be the sum of absolute errors in the discrete approximations to the partial derivatives (i.e., the error in approximating $\partial^k f(x)/\partial x_j^k$ by $(D_{x_j^k}\theta)(x)/N^k$, and similarly at $x + e_j/N$). In the third line, we use Holder smoothness to upper bound the first term, and we use standard numerical analysis (details in Appendix A) for the second term to ensure that $\delta(N) \leq cL/N$ for a constant $c > 0$ depending only on k. Summing the bound in (2.16) over $x \in Z_d$ as appropriate gives a uniform bound on the KTF penalty at θ, and leads to the next result.

Lemma 2.3. For any integers $k \geq 0$, $d \geq 1$, the (discretized) Holder and KTF classes defined in (2.15), (2.14) satisfy $\mathcal{H}_d^{k+1}(L) \subseteq \widetilde{T}_d^k(cLn^{1-(k+1)/d})$, where $c > 0$ is a constant depending only on k.

This lemma has three purposes. First, it provides some supporting evidence that the KTF class is an interesting smoothness class to study, as it shows the KTF class contains (discretizations of) Holder smooth functions, which are a cornerstone of classical nonparametric regression theory. In fact, this containment is strict and the KTF class contains

more heterogeneous functions in it as well. Second, it leads us to define what we call the *canonical scaling* $C_n \asymp n^{1-(k+1)/d}$ for the radius of the KTF class (2.14). This will be helpful for interpreting our minimax lower bounds in what follows; at this scaling, note that we have $\mathcal{H}_d^{k+1}(1) \subseteq \widetilde{\mathcal{T}}_d^k(C_n)$. Third and finally, it gives us an easy way to establish lower bounds on the minimax estimation error over KTF classes, by invoking well-known results on minimax rates for Holder classes. This will be described shortly.

As for GTF, due to the lower order discrete derivatives for x on the boundary of the grid Z_d, the discretized Holder class is not contained in a similarly defined GTF class.

Lemma 2.4. For any integers $k, d \geq 1$, there are elements in the (discretized) Holder class $\mathcal{H}_d^{k+1}(1)$ in (2.15) that do not lie in the GTF class $\mathcal{T}_d^k(C_n)$ in (2.13) for arbitrarily large C_n.

The fact that GTF classes do not contain (discretized) Holder classes makes them seem less natural (and perhaps, in a sense, less interesting) than KTF classes. In addition, it means that we cannot use standard minimax theory for Holder classes to establish lower bounds for the estimation error over GTF classes. However, as we will see next, we can construct lower bounds for GTF classes via another (more purely geometric) embedding strategy; interestingly, the resulting rates match the Holder rates, suggesting that, while GTF classes do not contain all (discretized) Holder functions, they do contain "enough" of these functions to admit the same lower bound rates.

We have the following lower bounds on the minimax rates over KTF and GTF classes.

Theorem 2.4. For any integers $k \geq 0$, $d \geq 1$, the minimax estimation error for GTF class defined in (2.13) satisfies

$$R\big(\mathcal{T}_d^k(C_n)\big) = \Omega\bigg(\frac{\sigma^2}{n} + \frac{C_n}{n} + \bigg(\frac{C_n}{n}\bigg)^{\frac{2d}{2k+2+d}}\bigg).$$

Theorem 2.5. For any integers $k \geq 0$, $d \geq 1$, the minimax estimation error for KTF class defined in (2.14) satisfies

$$R\big(\widetilde{\mathcal{T}}_d^k(C_n)\big) = \Omega\bigg(\frac{(k+1)^d\sigma^2}{n} + \frac{C_n}{n} + \bigg(\frac{C_n}{n}\bigg)^{\frac{2d}{2k+2+d}}\bigg).$$

The first terms in the two lower bounds are due to the nullity of the GTF and KTF operators. We get the second terms by embedding ℓ_1-balls of appropriate size in the classes $\mathcal{T}_d^k(C_n)$ and $\mathcal{T}_d^k(C_n)$, and using the lower bound results from Birge & Massart (2001) on ℓ_1-balls. We derive the third terms in the two lower bounds in different ways. For KTF classes, it follows from the Holder class embedding from Lemma 2.3 and classical minimax theory for Holder classes Korostelev & Tsybakov (2003), Tsybakov (2009). For GTF classes, we do not have this Holder class embedding; however, we can embed an ellipse, then rotate the parameter space and embed a hypercube. The proofs of these theorems are in Appendices A.5, A.5.2. Several remarks are in order.

Remark 2.2. For all $d \geq 2$ and $k \geq 0$, the lower bounds in Theorems 2.4,2.5, certify that the upper bound rates in Theorems 2.2,2.3 are tight, modulo a $\log n$ factor.

Remark 2.3. Plugging in the canonical scaling $C_n \asymp n^{1-(k+1)/d}$ in Theorems 2.5 and taking the dominant terms, we see that

$$R(\widetilde{\mathcal{T}}_d^k(C_n)) = \begin{cases} \Omega\big(n^{-\frac{k+1}{d}}\big) & 2k+2 \leq d \\ \Omega\big(n^{-\frac{2k+2}{2k+2+d}}\big) & 2k+2 > d. \end{cases}$$

The same lower bound holds for $R(\mathcal{T}_d^k(C_n))$.

Remark 2.4. An immediate consequence of Theorem 2.3 and the Holder embedding in Lemma 2.3 is that the KTF estimator achieves a rate (ignoring log factors) of $n^{-(2k+2)/(2k+2+d)}$ over $\mathcal{H}_d^{k+1}(1)$. Since this matches the well-known lower bound rate for Holder class, we see that KTF adapts automatically to Holder smooth signals, i.e., it achieves the optimal rate (up to log factors) for the more homogeneous class $\mathcal{H}_d^{k+1}(1)$, for $d \geq 2$ and all $2k+2 \geq d$. It is not clear that GTF shares this property.

2.5 Minimax bounds restricted to linear estimators

Next, we ask the question of whether the same minimax rate on the KTF and GTF classes $\widetilde{\mathcal{T}}_d(C_n)$, $\mathcal{T}_d(C_n)$ defined in (2.14), (2.13) can be achieved by linear smoothers — a simpler class of commonly used estimators, including Laplacian eigenmaps, kernel smoothing and so on. For illustration purposes, we first give the results in the case of TV denoising (that is, $k = 0$) before stating the results for general k. Recall that for $k = 0$, $\mathcal{T}_d^k(C_n)$ is same as $\widetilde{\mathcal{T}}_d^k(C_n)$ for any d, k, C_n.

Theorem 2.6. Denote $d_{\max} = 2d$. Then

$$R_L(\mathcal{T}_d(C_n)) \geq \frac{\sigma^2 C_n^2}{C_n^2 + \sigma^2 d_{\max}^2 n} \vee \frac{\sigma^2}{n} \geq \frac{1}{2}\left(\frac{C_n^2}{d_{\max}^2 n} \wedge \sigma^2\right) \vee \frac{\sigma^2}{n}. \tag{2.17}$$

The proof relies on an elegant meta-theorem on minimax rates from Donoho et al. (1990), which uses the concept of a "quadratically convex" set, whose minimax linear risk is the same as that of its hardest rectangular subproblem.

Remark 2.5. When C_n^2 grows with n, but not too fast (scales as $\sqrt{n}$, at most), the lower bound rate in (2.17) will be C_n^2/n. Compared to the C_n/n minimax rate from Theorem 2.2 (ignoring log terms), we see a clear gap between optimal nonlinear and linear estimators. In fact, under the canonical scaling $C_n \asymp n^{1-1/d}$, for any $d \geq 2$, this gap is seemingly huge: the lower bound for the minimax linear rate will be a constant, whereas the minimax rate (ignoring log terms) will be $n^{-1/d}$. This justifies the practical success of TV denoising over linear smoothers such as Laplacian smoothing.

The lower bound in Theorem 2.6 is essentially tight, and remarkably, it is certified by analyzing two trivial linear estimators: the mean estimator and the identity estimator.

Lemma 2.5. Let M_n denote the largest column norm of $D^\dagger$ where $D^\dagger$ denotes the pseudo-inverse of the GTF penalty operator in d-dimensions for $k = 0$. For the mean estimator $\hat{\theta}^{\text{mean}} = \bar{y}\mathbb{1}$,

$$\sup_{\theta_0 \in \mathcal{T}_d(C_n)} \mathbb{E}\big[\text{MSE}(\hat{\theta}^{\text{mean}}, \theta_0)\big] \leq \frac{\sigma^2 + C_n^2 M_n^2}{n}.$$

From Proposition 4 in Hutter & Rigollet (2016), we have $M_n = O(\sqrt{\log n})$ when $d = 2$ and $M_n = O(1)$ when $d \geq 3$.

The risk of the identity estimator $\hat{\theta}^{\text{id}} = y$ is clearly σ^2. Combining this logic with Lemma 2.5 gives the upper bound $R_L(\mathcal{T}_d(C_n)) \leq (\sigma^2 + C_n^2 M_n^2)/n \wedge \sigma^2$. Comparing this with the lower bound described in Remark 2.5, we see that the two rates basically match, modulo the M_n^2 factor in the upper bound, which only provides an extra $\log n$ factor when $d = 2$. The takeaway message: in the sense of max risk, the best linear smoother does not perform much better than the trivial estimators.

Now we give the result for GTF classes of order $k \geq 0$.

Theorem 2.7. Denote $\Delta = \Delta^{(k+1)}$. Let $\kappa = \text{nullity}(\Delta) = 1$. The minimax linear rate of the GTF class satisfies

$$R_L(\mathcal{T}_d^k(C_n)) \geq \begin{cases} \min\left\{\sigma^2, \frac{\kappa\sigma^2}{n} + c\frac{C_n^2}{n}\right\}, & \text{if } 2k+2 \leq d \\ \min\left\{\sigma^2, \frac{\kappa\sigma^2}{n} + c\left(\frac{C_n^2}{n}\right)^{\frac{d}{2k+2}}\right\}, & \text{if } 2k+2 > d. \end{cases} \tag{2.18}$$

for some constant c independent of n. If $C_n = 0$, then $\hat{\theta}^{\text{mean}}$ achieves this rate. If $C_n > 0$, then the rate is attained by the linear smoother

$$\hat{y} = \left(I + \frac{m\sigma^2}{C_n^2} L^{k+1}\right)^{-1} y$$

where $m = d(n - n^{1-1/d})$ is the number of edges in the grid graph and $L^{k+1} = \Delta^T \Delta$ is graph Laplacian of the graph raised to $k+1$. Further, in the case $2k+2 \leq d$, the trivial estimator $\widehat{\theta}^{\text{mean}}$ satisfies

$$\sup_{\theta_0 \in \mathcal{T}_d^k(C_n)} \mathbb{E}\big[\text{MSE}(\hat{\theta}^{\text{mean}}, \theta_0)\big] \leq \frac{\kappa\sigma^2 + C_n^2 M_{n,k}}{n},$$

where $M_{n,k} = O(1)$ if $2k+2 < d$ and $M_{n,k} = O(\log n)$ if $2k+2 = d$.

A similar result holds for kth order KTF classes. Instead of $\hat{\theta}^{\text{mean}}$, the "trivial" estimator in this case is the polynomial projection estimator

$$\hat{\theta}^{\text{poly}} = P_{\text{null}(\widetilde{\Delta}^{(k+1)})} y \tag{2.19}$$

where $P_{\text{null}(\Delta)}$ is the matrix that projects onto the null space of Δ. If Δ is the GTF operator of any order $k \geq 0$, note that the analogous projection estimator is simply the mean estimator $\hat{\theta}^{\text{mean}}$ defined in Lemma 2.5.

Theorem 2.8. Denote $\Delta = \widetilde{\Delta}^{(k+1)}$. Let $\kappa = \text{nullity}(\Delta) = (k+1)^d$. The minimax linear rate of the KTF class is

$$R_L(\widetilde{\mathcal{T}}_d^k(C_n)) \asymp \begin{cases} \min\left\{\sigma^2, \frac{\kappa\sigma^2}{n} + \frac{C_n^2}{n}\right\}, & \text{if } 2k+2 \leq d \\ \min\left\{\sigma^2, \frac{\kappa\sigma^2}{n} + \left(\frac{C_n^2}{n}\right)^{\frac{d}{2k+2}}\right\}, & \text{if } 2k+2 > d. \end{cases} \tag{2.20}$$

If $C_n = 0$, then $\widehat{\theta}^{\text{poly}}$ in (2.19) achieves this rate. If $C_n > 0$, then the rate is attained by the linear smoother

$$\hat{y} = \left(I + \frac{m\sigma^2}{C_n^2} L^{(k+1)}\right)^{-1} y$$

where $m = d(n - n^{1-1/d})$ is the number of edges in the grid graph and $L^{(k+1)} = \Delta^T \Delta$. Further, in the case $2k + 2 \leq d$, the trivial estimator $\hat{\theta}^{\text{poly}}$ satisfies

$$\sup_{\theta_0 \in \widetilde{\mathcal{T}}_d^k(C_n)} \mathbb{E}\big[\text{MSE}(\hat{\theta}^{\text{poly}}, \theta_0)\big] \leq \frac{\kappa\sigma^2 + C_n^2 M_{n,k}}{n},$$

where $M_{n,k} = O(1)$ if $2k + 2 < d$ and $M_{n,k} = O(\log n)$ if $2k + 2 = d$.

Remark 2.6. Assume $2k + 2 \leq d$. Either the identity estimator, or the trivial estimator $\hat{\theta}^{\text{mean}}$ ($\hat{\theta}^{\text{poly}}$ for KTF) achieves the lower bound rate in (2.18) ((2.20) for KTF) up to a factor $M_{n,k}$ (which is $O(1)$ when $2k + 2 < d$ and $O(\log n)$ when $2k + 2 = d$).

Suboptimality of linear smoothers. Similar to the case $k = 0$, linear smoothers do not perform as well as KTF on KTF classes for $k \geq 1$. Consider the case $2k + 2 \leq d$. If C_n grows with n (but grows slower than $\sqrt{n}$, then C_n^2/n is the dominant lower bound term for R_L and we see that R_L/R is larger than a factor of C_n. Plugging in the canonical rate $C_n \asymp n^{1-(k+1)/d}$, we see that $R_L \asymp \sigma^2$ because $2k + 2 \leq d$. In other words, all linear smoothers fail to be consistent on KTF classes! The same remark is applicable for GTF classes as well.

Now consider the other case $2k + 2 > d$. Assume that σ is a constant. To see the ranges for C_n where the bounds in Theorems 2.8 (or equivalently 2.7) and 2.3 are non-trivial, observe that

$$\frac{1}{n} \leq \left(\frac{C_n^2}{n}\right)^{\frac{d}{2k+2}} \leq 1 \iff n^{\frac{1}{2} - \frac{k+1}{d}} \leq C_n \leq \sqrt{n}, \text{ and}$$
$$\frac{1}{n} \leq \left(\frac{C_n}{n}\right)^{\frac{2d}{2k+2+d}} \leq 1 \iff n^{\frac{1}{2} - \frac{k+1}{d}} \leq C_n \leq n.$$

Consider the non-trivial range of C_n where it grows (strictly) faster than $n^{1/2-(k+1)/d}$ (otherwise $\kappa\sigma^2/n$ will dominate the lower bound on R_L) but slower than $\sqrt{n}$ (otherwise the σ^2 term will dominate). In this range, for $\widetilde{\mathcal{T}}_d^k(C_n)$ (or $\mathcal{T}_d^k(C_n)$),

$$\frac{R_L}{R} \gtrsim \left(\frac{C_n^2}{n}\right)^{\frac{d}{2k+2}} \left(\frac{C_n}{n}\right)^{\frac{-2d}{2k+2+d}} = \left(C_n n^{\frac{k+1}{d} - \frac{1}{2}}\right)^{\frac{d^2}{(k+1)(2k+2+d)}}.$$

To simplify for illustration, suppose $C_n = n^{\alpha + \frac{1}{2} - \frac{k+1}{d}}$ with $\alpha \in \left(0, \frac{k+1}{d}\right)$ so that it is in the said non-trivial range. Then

$$\frac{R_L}{R} \gtrsim n^{\frac{\alpha d^2}{(k+1)(2k+2+d)}}$$

and so, the ratio grows unbounded. For canonical scaling $C_n \asymp n^{1-\frac{k+1}{d}}$, we can plugin $\alpha = \frac{1}{2}$ in the above display, to see that

$$\frac{R_L}{R} \gtrsim n^{\frac{d^2}{(2k+2)(2k+2+d)}}.$$

These results reveal a significant gap between linear smoothers and optimal estimators, for estimation over $\mathcal{T}_d^k(C_n)$ and $\mathcal{T}_d^k(C_n)$ in d dimensions, as long as C_n scales appropriately. Roughly speaking, the TV classes and higher-order TV classes encompass a challenging setting for estimation because they are very broad, containing a wide array of functions—both globally smooth functions, said to have homogeneous smoothness, and functions with vastly different levels of smoothness at different grid locations, said to have heterogeneous smoothness. Linear smoothers cannot handle heterogeneous smoothness, and only nonlinear methods can enjoy good estimation properties over the entirety of $\mathcal{T}_d^k(C_n)$ (or $\widetilde{\mathcal{T}}_d^k(C_n)$). To reiterate, a telling example is in 2d with $k = 0$ and canonical scaling $C_n \asymp \sqrt{n}$, where we see that TV denoising achieves the optimal $1/\sqrt{n}$ rate (up to log factors), meanwhile, the best linear smoothers have max risk that is constant over $\mathcal{T}_2^0(\sqrt{n})$. See Figure 2.3 for an illustration.

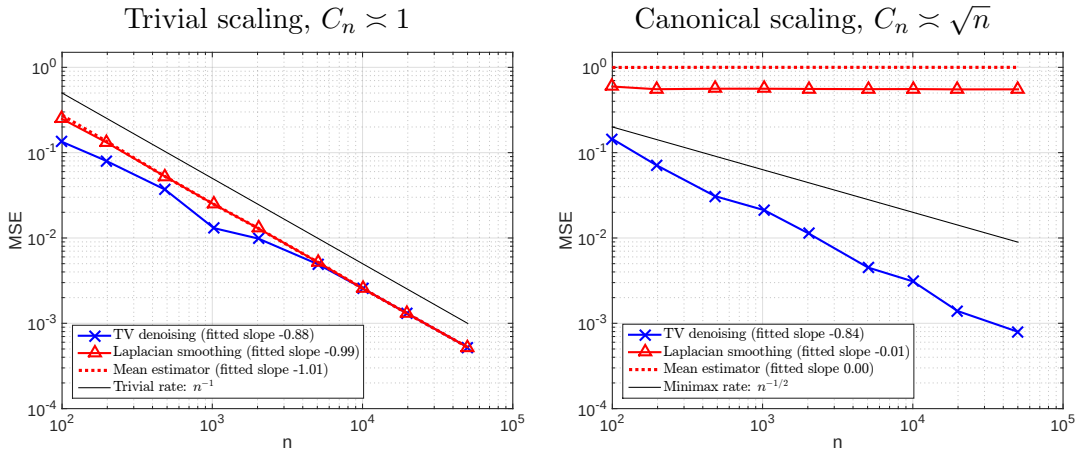

Figure 2.3: *MSE curves for estimation over a 2d grid, under two very different scalings of C_n: constant and $\sqrt{n}$. The parameter θ_0 was a "one-hot" signal, with all but one component equal to 0. For each n, the results were averaged over 5 repetitions, and Laplacian smoothing and TV denoising were tuned for optimal average MSE.*

2.6 Summary of rates

We conclude this chapter with a summary of rates over KTF and GTF smoothness classes in Table 2.1, under canonical scaling. For $\mathcal{T}_d^k(C_n)$ (and likewise for $\widetilde{\mathcal{T}}_d^k(C_n)$), the GTF estimator (KTF estimator) achieves the upper bound on the minimax error as stated in Theorem 2.2 (Theorem 2.3). The minimax lower bounds are from Theorems 2.4 and 2.5. The minimax linear rate is supported by Theorems 2.7 and 2.8.

	R Upper bound	R Lower bound	R_L
$2k+2<d$	$n^{-\frac{k+1}{d}}\sqrt{\log n}$	$n^{-\frac{k+1}{d}}$	1
$2k+2=d$	$n^{-\frac{1}{2}}\log n$	$n^{-\frac{1}{2}}$	1
$2k+2>d$	$n^{-\frac{2k+2}{2k+2+d}}(\log n)^{\frac{d}{2k+2+d}}$	$n^{-\frac{2k+2}{2k+2+d}}$	$n^{-\frac{2k+2-d}{2k+2}}$

Table 2.1: Error bounds over $\mathcal{T}_d^k(C_n)$ and $\widetilde{\mathcal{T}}_d^k(C_n)$ with canonical scaling $C_n = n^{1-\frac{k+1}{d}}$ modulo $\log n$ factors.

Chapter 3

Additive models with Trend filtering

In this chapter, we discuss results from our work Sadhanala & Tibshirani (2017) on additive models with trend filtering components.

3.1 Introduction

A common but simple approach to tackle curse of dimensionality in nonparametric regression is to assume that the regression function is additive. We consider an *additive model* for responses $Y^i \in \mathbb{R}$, $i = 1, \dots, n$ and corresponding input points $X^i = (X_1^i, \dots, X_d^i) \in \mathbb{R}^d$, $i = 1, \dots, n$, of the form

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

where $\mu \in \mathbb{R}$ is an overall mean parameter, each f_{0j} is a univariate function with $\sum_{i=1}^{n} f_{0j}(X_j^i) = 0$ for identifiability, $j = 1, \dots, d$, and the errors ϵ^i, $i = 1, \dots, n$ are i.i.d. with mean zero. A comment on notation: here and throughout, when indexing over the n samples we use superscripts, and when indexing over the d dimensions we use subscripts, so that, e.g., X_j^i denotes the jth component of the ith input point. (Exceptions will occasionally be made, but the role of the index should be clear from the context.)

Additive models are a special case of the more general *projection pursuit regression* model of Friedman & Stuetzle (1981). Additive models for the Cox regression and logistic regression settings were studied in Tibshirani (1983) and Hastie (1983), respectively. Some of the first asymptotic theory for additive models was developed in Stone (1985). Two algorithms closely related to (backfitting for) additive models are the *alternating least squares* and *alternating conditional expectations* methods, from van der Burg & de Leeuw (1983) and Breiman & Friedman (1985), respectively. The work of Buja et al. (1989) advocates for the use of additive models in combination with linear smoothers, a surprisingly simple combination that gives rise to flexible and scalable multidimensional regression tools. The book by Hastie & Tibshirani (1990) is the definitive practical guide for additive models for exponential family data distributions, i.e., generalized additive models.

More recent work on additive models is focused on high-dimensional nonparametric estimation, and here the natural goal is to induce sparsity in the component functions, so that only a few select dimensions of the input space are used in the fitted additive model. Some nice contributions are given in Lin & Zhang (2006), Ravikumar et al. (2009), Meier et al. (2009), all primarily focused on fitting splines for component functions and achieving sparsity through a group lasso type penalty. In other even more recent and interesting work sparse additive models, Lou et al. (2016) consider a semiparametric (partially linear) additive model, and Petersen et al. (2016) study componentwise fused lasso (i.e., total variation) penalization.

The literature on additive models (and by now, sparse additive models) is vast and the above is far from a complete list of references. In this paper, we examine a method for estimating additive models wherein each component is fit in a way that is *locally adaptive* to the underlying smoothness along its associated dimension of the input space. The literature on this line of work, as far as we can tell, is much less extensive. First, we review linear smoothers in additive models, motivate our general goal of local adaptivity, and then describe our specific proposal.

3.1.1 Review: additive models and linear smoothers

The influential paper by Buja et al. (1989) studies additive minimization problems of the form

$$\begin{aligned} \min_{\theta_1,\dots,\theta_d \in \mathbb{R}^n} \quad & \bigg\| Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j \bigg\|_2^2 + \lambda \sum_{j=1}^d \theta_j^T Q_j \theta_j \\ \text{subject to} \quad & \mathbb{1}^T \theta_j = 0, \quad j = 1, \dots, d, \end{aligned} \tag{3.1}$$

where $Y = (Y^1, \dots, Y^n) \in \mathbb{R}^n$ denotes the vector of responses, and $Y - \bar{Y}\mathbb{1}$ is its centered version, with $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y^i$ denoting the sample mean of Y, and $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^n$ the vector of all 1s. Each vector $\theta_j = (\theta_j^1, \dots, \theta_j^n) \in \mathbb{R}^n$ represents the evaluations of the jth component function f_j in our model, i.e., tied together by the relationship

$$\theta_j^i = f_j(X_j^i), \quad i = 1, \dots, n, \; j = 1, \dots, d.$$

In the problem (3.1), $\lambda \geq 0$ is a regularization parameter and Q_j, $j = 1, \dots, d$ are penalty matrices. As a typical example, we might consider Q_j to be the Reinsch penalty matrix for smoothing splines along the jth dimension of the input space, for $j = 1, \dots, d$. Under this choice, a backfitting (block coordinate descent) routine for (3.1) would repeatedly cycle through the updates

$$\theta_j = (I + \lambda Q_j)^{-1} \bigg(Y - \bar{Y}\mathbb{1} - \sum_{\ell \neq j} \theta_\ell \bigg), \quad j = 1, \dots, d, \tag{3.2}$$

where the jth update fits a smoothing spline to the jth partial residual, over the jth dimension of the input points, denoted by $X_j = (X_j^1, X_j^2, \dots X_j^n) \in \mathbb{R}^n$. At convergence, we arrive at an additive smoothing spline estimate, which solves (3.1).

Modeling the component functions as smoothing splines is arguably the most common formulation for additive models, and it is the standard in several statistical software

packages like the R package `gam`. However, as Buja et al. (1989) explain, the backfitting perspective suggests a more algorithmic approach to additive modeling: one can replace the operator $(I + \lambda Q_j)^{-1}$ in (3.2) by S_j, a particular (user-chosen) *linear smoother*, meaning, a linear map that performs univariate smoothing across the jth dimension of inputs X_j. The linear smoothers S_j, $j = 1, \ldots, d$ could correspond to smoothing splines, regression splines (regression using a spline basis with given knots), kernel smoothing, local polynomial smoothing, or a combination of these, across the input dimensions. In short, as argued in Buja et al. (1989), the class of linear smoothers is broad enough to offer fairly flexible, interesting mechanisms for smoothing, and simple enough to understand precisely. Most of the work following Buja et al. (1989) remains in keeping with the idea of using linear smoothers in combination with additive models.

3.1.2 The limitations of linear smoothers

The beauty of linear smoothers lies in their simplicity. However, with this simplicity comes serious limitations, in terms of their ability to adapt to varying local levels of smoothness. In the univariate setting, the seminal theoretical work by Donoho & Johnstone (1998) makes this idea precise. With $d = 1$, suppose that underlying regression function f_0 lies in the univariate function class

$$\mathcal{F}_k(C) = \{f : \mathrm{TV}(f^{(k)}) \leq C\}, \tag{3.3}$$

for a constant $C > 0$, where $\mathrm{TV}(\cdot)$ is the total variation operator, and $f^{(k)}$ the kth weak derivative of f. The class in (3.3) allows for greater fluctuation in the local level of smoothness of f_0 than, say, more typical function classes like Holder and Sobolev spaces. The results of Donoho & Johnstone (1998) (see also Section 5.1 of Tibshirani (2014)) imply that the minimax error rate for estimation over $\mathcal{F}_k(C)$ is $n^{-(2k+2)/(2k+3)}$, but the minimax error rate when we consider only linear smoothers (linear transformations of Y) is $n^{-(2k+1)/(2k+2)}$. This difference is highly nontrivial, e.g., for $k = 0$ this is a difference of $n^{-2/3}$ (optimal) versus $n^{-1/2}$ (optimal among linear smoothers) for estimating a function f_0 of bounded variation.

It is important to emphasize that this shortcoming is not just a theoretical one; it is also clearly noticeable in basic practical examples. Just as linear smoothers will struggle in the univariate setting, an additive estimate based on linear smoothers will not be able to efficiently track local changes in smoothness, across any of the input dimensions. This could lead to a loss in accuracy even if only some of the components f_{0j}, $j = 1, \ldots, d$ possesses heterogeneous smoothness across its domain.

Two well-studied univariate estimators that are locally adaptive, i.e., that attain the minimax error rate over the kth order total variation class in (1.3), are wavelet smoothing and locally adaptive regression splines, as developed by Donoho & Johnstone (1998) and Mammen & van de Geer (1997), respectively. There is a substantial literature on these methods in the univariate case (especially for wavelets), but fewer authors have considered them in the additive models context. Some notable exceptions are Zhang & Wong (2003), Sardy & Tseng (2004), Petersen et al. (2016), with the latter work especially related to our focus in this paper.

3.1.3 Additive trend filtering

We consider additive models that are constructed using *trend filtering* (instead of linear smoothers, wavelets, or locally adaptive regression splines) as their componentwise smoother. The computational efficiency, along with its capacity for local adaptivity, makes trend filtering a particularly desirable candidate to extend to the additive model setting. Specifically, we consider the *additive trend filtering* estimate of order $k \geq 0$, defined as a solution in the problem

$$\begin{aligned} \min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \quad & \frac{1}{2}\Big\|Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j\Big\|_2^2 + \lambda \sum_{j=1}^d \big\|D^{(X_j,k+1)} S_j \theta_j\big\|_1 \\ \text{subject to} \quad & \mathbb{1}^T \theta_j = 0, \quad j = 1,\ldots,d. \end{aligned} \tag{3.4}$$

As before, $Y - \bar{Y}\mathbb{1}$ is the centered response vector, $\lambda \geq 0$ is a regularization parameter, and now $S_j \in \mathbb{R}^{n\times n}$ in (3.4) is a permutation matrix that sorts the jth component of inputs $X_j = (X_j^1, X_j^2, \ldots X_j^n)$ into increasing order, i.e.,

$$S_j X_j = (X_j^{(1)}, X_j^{(2)}, \ldots, X_j^{(n)}), \quad j = 1,\ldots,d.$$

Also, $D^{(X_j,k+1)}$ in (3.4) is the $(k+1)$st order difference operator, as in (1.5), (1.6), but defined over the sorted jth dimension of inputs $S_j X_j$, for $j = 1,\ldots,d$. With backfitting (block coordinate descent), computation of a solution in (3.4) is still quite efficient, since we can leverage the efficient routines for univariate trend filtering.

3.1.4 A motivating example

Figure 3.1 shows a simulated example that compares the additive trend filtering estimates in (3.4) (of quadratic order, $k = 2$), to the additive smoothing spline estimates in (3.1) (of cubic order). In the simulation, we used $n = 3000$ and $d = 3$. We drew input points $X^i \overset{\text{i.i.d.}}{\sim} \text{Unif}[0,1]^3$, $i = 1,\ldots,3000$, and drew responses $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^3 f_{0j}(X_j^i), \sigma^2)$, $i = 1,\ldots,3000$, where $\sigma = 1.72$ was set to give a signal-to-noise ratio of about 1. The underlying component functions were defined as

$$\begin{aligned} f_{01}(t) &= \min(t, 1-t)^{0.2} \sin\left(\frac{2.85\pi}{0.3 + \min(t, 1-t)}\right), \\ f_{02}(t) &= e^{3t}\sin(4\pi t), \quad f_{03}(t) = -(t - 1/2)^2, \end{aligned}$$

so that f_{01}, f_{02}, f_{03} possess different levels of smoothness (f_{03} being the smoothest, f_{02} less smooth, and f_{01} the least smooth), and so that f_{01} itself has heteregeneous smoothness across its domain.

The first row of Figure 3.1 shows the estimated component functions from additive trend filtering, at a value of λ that minimizes the mean squared error (MSE), computed over 20 repetitions. The second row shows the estimates from additive smoothing splines, also at a value of λ that minimizes the MSE. We see that the trend filtering fits adapt well to the varying levels of smoothness, but the smoothing spline fits are undersmoothed, for the most part. In terms of effective degrees of freedom (df), the additive smoothing spline

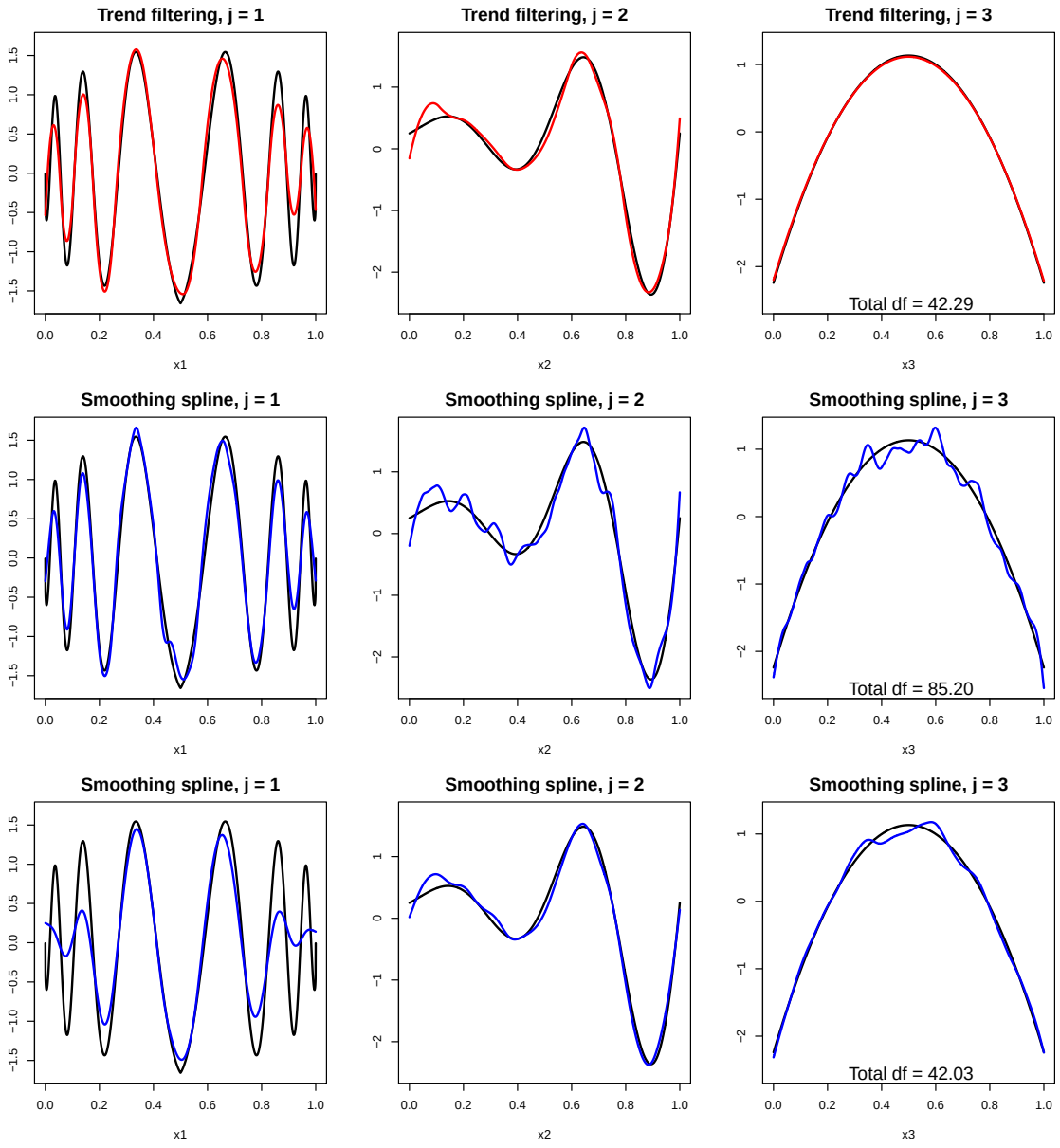

Figure 3.1: *Comparing estimates from additive trend filtering* (3.4) *(of quadratic order) and additive smoothing splines* (3.1) *(of cubic order), for a simulation with* $n = 3000$ *and* $d = 3$*, as described in Section 3.1.4. In each row, the underlying component functions are plotted in black. The first row shows the estimated component functions using additive trend filtering, in red, at a value of* λ *chosen to minimize mean squared error (MSE), computed over 20 repetitions. The second row shows the estimates from additive smoothing splines, in blue, again at a value of* λ *that minimizes MSE. The third row shows the estimates from additive smoothing splines when* λ *is tuned so that the effective degrees of freedom (df) of the fit roughly matches that of additive trend filtering in the first row.*

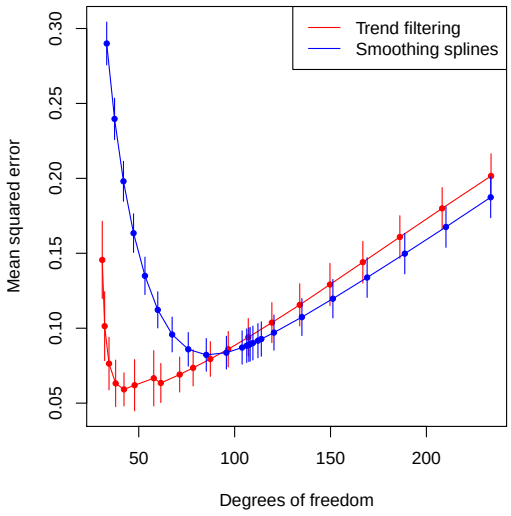

Figure 3.2: *MSE curves for additive trend filtering and additive smoothing splines, computed over 20 repetitions from the same simulation setup as in Figure 3.1. Vertical segments denote ±1 standard deviations. The MSE curves are parametrized by degrees of freedom (computed via standard Monte Carlo methods over the 20 repetitions).*

estimate is much more complex, having about 85 df (computed via Monte Carlo over the 20 repetitions); the additive trend filtering has only about 42 df. The third row of the figure shows the estimates from additive smoothing splines, when λ is chosen so that the resulting df roughly matches that of additive trend filtering in the first row. Now we see that the first component fit is oversmoothed, yet the third is still undersmoothed.

Figure 3.2 displays the MSE curves from additive trend filtering, as a function of df. We see that trend filtering achieves a lower MSE, and moreover, its MSE curve is optimized at a lower df (i.e., less complex model) than that for smoothing splines. This is analogous to what is typically seen in the univariate setting (Tibshirani 2014).

We note that this motivating example is intended to elucidate the differences in what additive smoothing splines and additive trend filtering can do with a single tuning parameter each; a serious applied statistician, in just $d = 3$ dimensions, would likely use REML or some related technique to fit a multiple tuning parameter smoothing spline model; see our later discussion on this topic in Section 3.5.2.

3.1.5 Summary of contributions

A summary of our contributions, and an outline for the rest of this paper, are given below.

- In Section 3.2, we investigate basic properties of the additive trend filtering model: an equivalent continuous-time formulation, a condition for uniqueness of component function estimates, and a simple formula for the effective degrees of freedom of the additive fit.

- In Section 3.3, we derive error bounds for additive trend filtering. Assuming that the underlying regression function is additive, denoted by $f_0 = \sum_{j=1}^d f_{0j}$, and that $\mathrm{TV}(f_{0j}^{(k)})$ is bounded, for $j = 1, \ldots, d$, we prove that the kth order additive trend filtering estimator converges to f_0 at the rate $n^{-(2k+2)/(2k+3)}$ when the dimension d is fixed (under weak assumptions), and at the rate $dn^{-(2k+2)/(2k+3)}$ when d is growing (under stronger assumptions). We prove that these rates are optimal in a minimax

sense, and also show that additive smoothing splines (generally, additive models built from linear smoothers of any kind) are suboptimal over such a class of functions f_0.

- In Section 3.4, we study the backfitting algorithm for additive trend filtering models, and give a connection between backfitting and an alternating projections scheme in the additive trend filtering dual problem. This inspires a new parallelized backfitting algorithm.

- In Section 3.5, we present empirical experiments and comparisons, and we also investigate the use of multiple tuning parameter models. In Section 3.6, we give a brief discussion.

3.2 Basic properties

In this section, we derive a number of basic properties of additive trend filtering estimates, starting with a representation for the estimates as continuous functions over $\mathbb{R}^d$ (rather than simply discrete fitted values at the input points).

3.2.1 Falling factorial representation

We may describe additive trend filtering in (3.4) as an estimation problem written in *analysis form*. The components are modeled directly by the parameters θ_j, $j = 1, \dots, d$, and the desired structure is established by regularizing the discrete derivatives of these parameters, through the penalty terms $\|D^{(X_j,k+1)} S_j \theta_j\|_1$, $j = 1, \dots, d$. Here, we present an alternative representation for (3.4) in *basis form*, where each component is expressed as a linear combination of basis functions, and regularization is applied to the coefficients in this expansion.

Tibshirani (2014), Wang et al. (2014) establish a connection between univariate trend filtering and the falling factorial functions in (1.7), and show that the trend filtering problem can be interpreted as a sparse basis regression problem using these functions. As we show next, the analogous result holds for additive trend filtering.

Lemma 3.1 (**Falling factorial representation**). For $j = 1, \dots, d$, let $h_1^{(X_j)}, \dots, h_n^{(X_j)}$ be the falling factorial basis in (1.7) with knots $(t^1, \dots, t^n) = S_j X_j$, the jth dimension of the input points, properly sorted. Then the additive trend filtering problem (3.4) is equivalent to the problem

$$\begin{aligned} \min_{\alpha_1,\dots,\alpha_d \in \mathbb{R}^n} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}(X_j^i) \Big)^2 + \lambda k! \sum_{j=1}^d \sum_{\ell=k+2}^n |\alpha_j^\ell| \\ \text{subject to} \quad & \sum_{i=1}^n \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}(X_j^i) = 0, \quad j = 1, \dots, d, \end{aligned} \tag{3.5}$$

in that, at any solutions in (3.4), (3.5), we have

$$\hat{\theta}_j^i = \sum_{\ell=1}^n \hat{\alpha}_j^\ell h_\ell^{(X_j)}(X_j^i), \quad i = 1, \dots, n, \; j = 1, \dots, d.$$

An alternative way of expressing problem (3.5) is

$$\begin{aligned} \min_{f_j \in \mathcal{H}_j,\, j=1,\ldots,d} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \Big)^2 + \lambda \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d, \end{aligned} \tag{3.6}$$

where $\mathcal{H}_j = \mathrm{span}\{h_1^{(X_j)}, \ldots, h_n^{(X_j)}\}$ is the span of the falling factorial basis over the jth dimension, and $f_j^{(k)}$ is the kth weak derivative of f_j, $j = 1, \ldots, d$. In this form, at any solutions in (3.4), (3.6),

$$\hat{\theta}_j^i = \hat{f}_j(X_j^i), \quad i = 1, \ldots, n, \; j = 1, \ldots, d.$$

Proof. For $j = 1, \ldots, d$, define the kth order falling factorial basis matrix $H^{(X_j,k)} \in \mathbb{R}^{n \times n}$ by

$$H_{i\ell}^{(X_j,k)} = h_\ell^{(X_j)}(X_j^i), \quad i = 1, \ldots, n, \; \ell = 1, \ldots, n. \tag{3.7}$$

Note that the columns of $H^{(X_j,k)}$ follow the order of the sorted inputs $S_j X_j$, but the rows do not; however, for $S_j H^{(X_j,k)}$, both its rows and columns of follow the order of $S_j X_j$. From Wang et al. (2014), we know that

$$(S_j H^{(X_j,k)})^{-1} = \begin{bmatrix} C^{(X_j,k+1)} \\ \frac{1}{k!} D^{(X_j,k+1)} \end{bmatrix},$$

for some matrix $C^{(X_j,k+1)} \in \mathbb{R}^{(k+1) \times n}$, i.e.,

$$(H^{(X_j,k)})^{-1} = \begin{bmatrix} C^{(X_j,k+1)} \\ \frac{1}{k!} D^{(X_j,k+1)} \end{bmatrix} S_j. \tag{3.8}$$

Problem (3.5) is given by reparameterizing (3.4) according to $\theta_j = H^{(X_j,k)} \alpha_j$, for $j = 1, \ldots, d$. As for (3.6), the equivalence between this and (3.5) follows by noting that for $f_j = \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}$, we have

$$f_j^{(k)}(t) = k! + k! \sum_{\ell=k+2}^n \alpha_j^\ell \cdot 1\{t > X_j^{\ell-1}\},$$

and so $\mathrm{TV}(f_j^{(k)}) = k! \sum_{\ell=k+2}^n |\alpha_j^\ell|$, for each $j = 1, \ldots, d$. □

This lemma not only provides an interesting reformulation for additive trend filtering, it is also practically useful in that it allows us to perform interpolation or extrapolation using the additive trend filtering model. That is, from a solution $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ in (3.4), we can extend each component fit $\hat{\theta}_j$ to the real line, by forming an appropriate linear combination of falling factorial functions:

$$\hat{f}_j(x_j) = \sum_{\ell=1}^n \hat{\alpha}_j^\ell h_\ell^{(X_j)}(x_j), \quad x_j \in \mathbb{R}. \tag{3.9}$$

The coefficients above are determined by the relationship $\hat{\alpha}_j = (H^{(X_j,k)})^{-1}\hat{\theta}_j$, and are easily computable given the highly structured form of $(H^{(X_j,k)})^{-1}$, as revealed in (3.8). Writing the coefficients in block form, as in $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j) \in \mathbb{R}^{(k+1)} \times \mathbb{R}^{(n-k-1)}$, we have

$$\hat{a}_j = C^{(X_j,k+1)} S_j \hat{\theta}_j, \tag{3.10}$$

$$\hat{b}_j = \frac{1}{k!} D^{(X_j,k+1)} S_j \hat{\theta}_j. \tag{3.11}$$

The first $k+1$ coefficients $\hat{a}_j$ index the pure polynomial functions $h_1^{(X_j)}, \dots, h_{k+1}^{(X_j)}$. These coefficients will be generically dense (the form of $C^{(X_j,k+1)}$ is not important here, so we omit it for simplicity, but details are given in Appendix B.1). The last $n-k-1$ coefficients $\hat{b}_j$ index the knot-producing functions $h_{k+2}^{(X_j)}, \dots, h_n^{(X_j)}$, and when $(\hat{b}_j)_\ell = \frac{1}{k!}(D^{(X_j,k+1)} S_j \hat{\theta}_j)_\ell \neq 0$, the fitted function $\hat{f}_j$ exhibits a knot at the $(\ell+k)$th sorted input point among $S_j X_j$, i.e., at $X_j^{(\ell+k)}$. Figure 3.3 gives an example, where $n = 1000$ and $d = 2$. We generated input points $X^i \overset{\text{i.i.d.}}{\sim} \text{Unif}[0,1]^2$, $i = 1, \dots, 1000$, and responses $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^2 f_{0j}(X_j^i), \sigma^2)$, $i = 1, \dots, 1000$, where $f_{01}(x_1) = \sqrt{x_1}\sin(3\pi/(x_1 + 1/2))$ and $f_{02}(x_2) = x_2(x_2 - 1/3)$, and $\sigma = 0.36$.

We note that the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$ in (3.10), (3.11) can be computed in $O(n)$ operations and $O(1)$ memory. This makes extrapolation of the jth fitted function $\hat{f}_j$ in (3.9) highly efficient. Details are given in Appendix B.1.

3.2.2 Uniqueness of component fits

It is easy to see that, for the problem (3.4), the additive fit $\sum_{j=1}^d \hat{\theta}_j$ is always uniquely determined: denoting $\sum_{j=1}^d \theta_j = T\theta$ for a linear operator T and $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^{nd}$, the loss term $\|y - T\theta\|_2^2$ is strictly convex in the variable $T\theta$, and this, along with the convexity of the problem (3.4), implies a unique additive fit $T\hat{\theta}$, no matter the choice of solution $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_d) \in \mathbb{R}^{nd}$.

On the other hand, when $d > 1$, the criterion in (3.4) is not strictly convex in θ, and hence there need not be a unique solution $\hat{\theta}$, i.e., the individual components fits $\hat{\theta}_j$, $j = 1, \dots, d$ need not be uniquely determined. We show next that uniqueness of the component fits can be guaranteed under some conditions on the input matrix $X = [X_1 \cdots X_d] \in \mathbb{R}^{n \times d}$. We will rely on the falling factorial representation for additive trend filtering, introduced in the previous subsection, and on the notion of *general position*: a matrix $A \in \mathbb{R}^{m \times p}$ is said to have columns in general position provided that, for any $\ell < \min\{m, p\}$, subset of $\ell + 1$ columns denoted $A_{i_1}, \dots, A_{i_{\ell+1}}$, and signs $s_1, \dots, s_{\ell+1} \in \{-1, 1\}$, the affine span of $\{s_1 A_{i_1}, \dots, s_{\ell+1} A_{i_{\ell+1}}\}$ does not contain any element of $\{\pm A_i : i \neq i_1, \dots, i_{\ell+1}\}$.

Lemma 3.2 (Uniqueness). For $j = 1, \dots, d$, let $H^{(X_j,k)} \in \mathbb{R}^{n \times n}$ be the falling factorial basis matrix constructed over the sorted jth dimension of inputs $S_j X_j \in \mathbb{R}^n$, as in (3.7). Decompose $H^{(X_j,k)}$ into its first $k+1$ columns $P^{(X_j,k)} \in \mathbb{R}^{n \times (k+1)}$, and its last $n-k-1$ columns $K^{(X_j,k)} \in \mathbb{R}^{n \times (n-k-1)}$. The former contains evaluations of the pure polynomials $h_1^{(X_j)}, \dots, h_{k+1}^{(X_j)}$; the latter contains evaluations of the knot-producing functions $h_{k+2}^{(X_j)}, \dots, h_n^{(X_j)}$. Also, let $\tilde{P}^{(X_j,k)}$ denote the matrix $P^{(X_j,k)}$ with its first column removed, for $j = 1, \dots, d$, and $M = I - \mathbb{1}\mathbb{1}^T/n$. Define

$$\tilde{P} = M\big[\, \tilde{P}^{(X_1,k)} \;\; \dots \;\; \tilde{P}^{(X_d,k)} \,\big] \in \mathbb{R}^{n \times dk}, \tag{3.12}$$

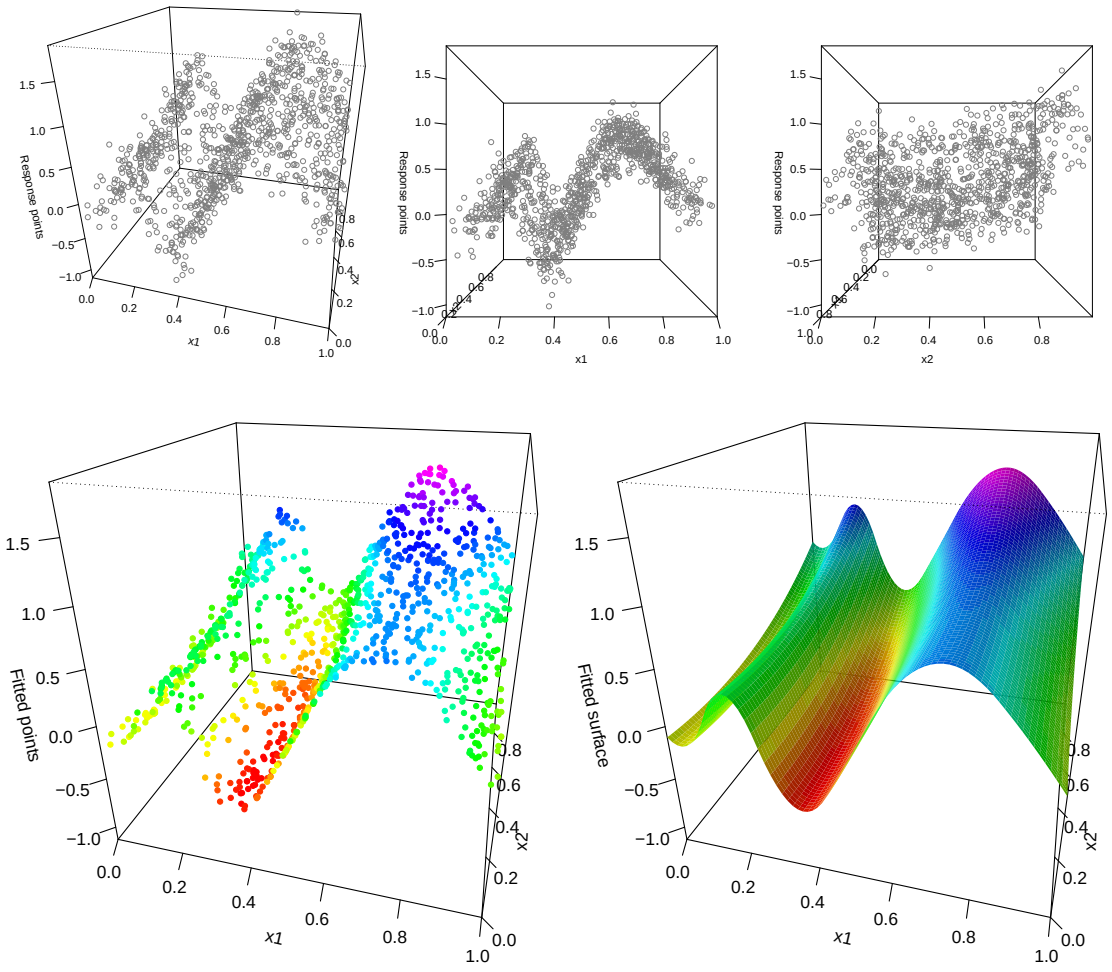

Figure 3.3: *An example of extrapolating the fitted additive trend filtering model, where* $n = 1000$ *and* $d = 2$. *The top row shows three perspectives of the data. The bottom left panel shows the fitted values from additive trend filtering* (3.4) *(with* $k = 2$ *and* $\lambda = 0.004$*), where points are colored by their depth for visualization purposes. The bottom right panel shows the 2d surface associated with the trend filtering estimate,* $\hat{f}_1(x_1) + \hat{f}_2(x_2)$ *over* $(x_1, x_2) \in [0, 1]^2$, *with each component function extrapolated as in* (3.9).

the product of M and the columnwise concatenation of $\tilde{P}^{(X_j,k)}$, $j = 1, \ldots, d$. Let UU^T denote the projection operator onto the space orthogonal to the column span of $\tilde{P}$, where $U \in \mathbb{R}^{n \times (n-kd-1)}$ has orthonormal columns, and define

$$\tilde{K} = U^T M \big[K^{(X_1,k)} \quad \ldots \quad K^{(X_d,k)} \big] \in \mathbb{R}^{(n-kd-1) \times (n-k-1)d}, \tag{3.13}$$

the product of $U^T M$ and the columnwise concatenation of $K^{(X_j,k)}$, $j = 1, \ldots, d$. A sufficient condition for uniqueness of the additive trend filtering solution in (3.4) can now be given in two parts.

1. If $\tilde{K}$ has columns in general position, then the knot-producing parts of all component fits are uniquely determined, i.e., for each $j = 1, \ldots, d$, the projection of $\hat{\theta}_j$ onto the column space of $K^{(X_j,k)}$ is unique.

2. If in addition $\tilde{P}$ has full column rank, then the polynomial parts of component fits are uniquely determined, i.e., for each $j = 1, \ldots, d$, the projection of $\hat{\theta}_j$ onto the column space of $P^{(X_j,k)}$ is unique, and thus the component fits $\hat{\theta}_j$, $j = 1, \ldots, d$ are all unique.

The proof is deferred to Appendix B.2. To rephrase, the above lemma decomposes each component of the additive trend filtering solution according to

$$\hat{\theta}_j = \hat{\theta}_j^{\text{poly}} + \hat{\theta}_j^{\text{knot}}, \quad j = 1, \ldots, d,$$

where $\hat{\theta}_j^{\text{poly}}$ exhibits a purely polynomial trend over $S_j X_j$, and $\hat{\theta}_j^{\text{knot}}$ exhibits a piecewise polynomial trend over $S_j X_j$, and hence determines the knot locations, for $j = 1, \ldots, d$. The lemma shows that the knot-producing parts $\hat{\theta}_j^{\text{knot}}$, $j = 1, \ldots, d$ are uniquely determined when the columns of $\tilde{K}$ are in general position, and the polynomial parts $\hat{\theta}_j^{\text{knot}}$, $j = 1, \ldots, d$ are unique when the columns of $\tilde{K}$ are in general position, and the columns of $\tilde{P}$ are linearly independent.

The conditions placed on $\tilde{P}, \tilde{K}$ in Lemma 3.2 are not strong. When $n > kd$, and the elements of input matrix X are drawn from a density over $\mathbb{R}^{nd}$, it is not hard to show that $\tilde{P}$ has full column rank with probability 1. We conjecture that, under the same conditions, $\tilde{K}$ will also have columns in general position with probability 1, but do not pursue a proof.

3.2.3 Dual problem

Let us abbreviate $D_j = D^{(X_j,k+1)}$, $j = 1, \ldots, d$ for the penalty matrices in the additive trend filtering problem (3.4). Basic arguments in convex analysis, deferred to Appendix B.3, show that the dual of problem (3.4) can be expressed as:

$$\begin{gathered} \min_{u \in \mathbb{R}^n} \ \|Y - \bar{Y}\mathbb{1} - u\|_2^2 \ \text{ subject to } \ u \in U = U_1 \cap \cdots \cap U_d, \\ \text{where } \ U_j = \{S_j D_j^T v_j : \|v_j\|_\infty \leq \lambda\}, \quad j = 1, \ldots, d, \end{gathered} \tag{3.14}$$

and that primal and dual solutions in (3.4), (3.14) are related by:

$$\sum_{j=1}^d \hat{\theta}_j = Y - \bar{Y}\mathbb{1} - \hat{u}. \tag{3.15}$$

From the form of (3.14), it is clear that we can write the (unique) dual solution as $\hat{u} = \Pi_U(Y - \bar{Y}\mathbb{1})$, where Π_U is the (Euclidean) projection operator onto U. Moreover, using (3.15), we can express the additive fit as $\sum_{j=1}^d \hat{\theta}_j = (\mathrm{Id} - \Pi_U)(Y - \bar{Y}\mathbb{1})$, where $\mathrm{Id} - \Pi_U$ is the operator that gives the residual from projecting onto U. These relationships will be revisited in Section 3.4, where we return to the dual perspective, and argue that the backfitting algorithm for the additive trend filtering problem (3.4) can be seen as a type of alternating projections algorithm for its dual problem (3.14).

3.2.4 Degrees of freedom

In general, given data $Y \in \mathbb{R}^n$ with $\mathbb{E}(Y) = \eta$, $\mathrm{Cov}(Y) = \sigma^2 I$, and an estimator $\hat{\eta}$ of η, recall that we define the *effective degrees of freedom* of $\hat{\eta}$ as (Efron 1986, Hastie & Tibshirani 1990):

$$\mathrm{df}(\hat{\eta}) = \frac{1}{\sigma^2} \sum_{i=1}^n \mathrm{Cov}\big(\hat{\eta}^i(Y), Y^i\big),$$

where $\hat{\eta}(Y) = (\hat{\eta}^1(y), \dots, \hat{\eta}^n(Y))$. Roughly speaking, the above definition sums the influence of the ith component Y^i on its corresponding fitted value $\hat{\eta}^i(Y)$, across $i = 1, \dots, n$. A precise understanding of degrees of freedom is useful for model comparisons (recall the x-axis in Figure 3.2), and other reasons. For linear smoothers, in which $\hat{\eta}(Y) = SY$ for some $S \in \mathbb{R}^{n \times n}$, it is clear that $\mathrm{df}(\hat{\eta}) = \mathrm{tr}(S)$, the trace of S. (This also covers additive models whose components are built from univariate linear smoothers, because in total these are still just linear smoothers: the additive fit is still just a linear function of Y.)

Of course, additive trend filtering is a not a linear smoother; however, it is a particular type of generalized lasso estimator, and degrees of freedom for such a class of estimators is well-understood (Tibshirani & Taylor 2011, 2012). The next result is a consequence of existing generalized lasso theory, proved in Appendix B.4.

Lemma 3.3 (**Degrees of freedom**). Assume the conditions of Lemma 3.2, i.e., that the matrix $\tilde{P}$ in (3.12) has full column rank, and the matrix $\tilde{K}$ in (3.13) is in general position. Assume also that the response is Gaussian, $Y \sim N(\eta, \sigma^2 I)$, and treat the input points $X^i \in \mathbb{R}^d$, $i = 1, \dots, n$ as fixed and arbitrary, as well as the tuning parameter value $\lambda \geq 0$. Then the additive trend filtering fit from (3.4) has degrees of freedom

$$\mathrm{df}\bigg(\sum_{j=1}^d \hat{\theta}_j\bigg) = \mathbb{E}\bigg(\sum_{j=1}^d (\text{number of knots in } \hat{\theta}_j)\bigg) + kd.$$

Remark 3.1 (**The effect of shrinkage**). Lemma 3.3 says that for an unbiased estimate of the degrees of freedom of the additive trend filtering fit, we count the number of knots in each component fit $\hat{\theta}_j$ (recall that this is the number of nonzeros in $D^{(X_j, k+1)}\hat{\theta}_j$, i.e., the number of changes in the discrete $(k+1)$st derivative), add them up over $j = 1, \dots, d$, and add kd. This may seem surprising, as these knot locations are chosen adaptively based on the data Y. But, such adaptivity is counterbalanced by the shrinkage induced by the ℓ_1 penalty in (3.4) (i.e., for each component fit $\hat{\theta}_j$, there is shrinkage in the differences between the attained kth derivatives on either side of a selected knot). See Tibshirani (2015) for a study of this phenomenon.

3.2.5 Two related additive spline estimators

From its equivalent formulation in (3.6), additive trend filtering is seen to be closely related to two other additive spline estimators, which we introduce here. Consider, for univariate function classes $\mathcal{S}_j$, $j = 1, \dots, d$, the problem

$$\begin{aligned} \min_{f_j \in \mathcal{S}_j \; j=1,\dots d} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \Big)^2 + \lambda \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \dots, d. \end{aligned} \tag{3.16}$$

When each $\mathcal{S}_j$, $j = 1, \dots, d$ is the set of k times weakly differentiable functions, we call the solution in (3.16) the *additive locally adaptive regression spline* of order $k \geq 0$, as it is the natural extension of the univariate estimator considered in Mammen & van de Geer (1997). Denote by $\hat{f}_j$, $j = 1, \dots, d$ this solution; the representation arguments used by these authors apply immediately to the additive setting, and imply that each $\hat{f}_j$, $j = 1, \dots, d$ is indeed a spline of degree k (justifying the choice of name). The same arguments show that, for $k = 0$ or $k = 1$, the knots of the spline $\hat{f}_j$ lie among the jth dimension of the input points $X_j^1, \dots, X_j^n$, for $j = 1, \dots, d$, but for $k \geq 2$, this need not be true, and in general the components will be splines with knots at locations other than the inputs.

We can facilitate computation by taking $\mathcal{S}_j = \mathcal{G}_j$, where $\mathcal{G}_j$ is the set of splines of degree k with knots lying among the jth dimension of inputs $X_j^1, \dots, X_j^n$, for $j = 1, \dots, d$. We call the resulting solution the *restricted additive locally adaptive regression spline* of order $k \geq 0$. More precisely, we require that the splines in $\mathcal{G}_j$ have knots in a set T_j, which, writing $t_j = S_j X_j$ for the sorted inputs along the jth dimension, is defined by

$$T_j = \begin{cases} \{t_j^{k/2+2}, \dots, t_j^{n-k/2}\} & \text{if } k \text{ is even}, \\ \{t_j^{(k+1)/2+1}, \dots, t_j^{n-(k+1)/2}\} & \text{if } k \text{ is odd}, \end{cases} \tag{3.17}$$

i.e., defined by removing $k+1$ input points at the boundaries, for $j = 1, \dots, d$. Setting $\mathcal{S}_j = \mathcal{G}_j$, $j = 1, \dots, d$ makes (3.16) a finite-dimensional problem, just like (3.6). When $k = 0$ or $k = 1$, as is evident from their form in (1.7), the falling factorial functions are simply splines, which means that $\mathcal{H}_j = \mathcal{G}_j$ for $j = 1, \dots, d$, hence additive trend filtering and restricted additive locally adaptive regression splines are the same estimator. When $k \geq 2$, this is no longer true, and they are not the same. Additive trend filtering will be much easier to compute, since $\mathrm{TV}(g^{(k)})$ does not admit a nice representation in terms of discrete derivatives for a kth order spline (and yet it does for a kth order falling factorial function, as seen in (3.4)).

To summarize, additive locally adaptive splines, restricted additive locally adaptive splines, and additive trend filtering all solve a problem of the form (3.16) for different choices of function classes $\mathcal{S}_j$, $j = 1, \dots, d$. For $k = 0$ or $k = 1$, these three estimators are equivalent. For $k \geq 2$, they will be generically different, though our intuition tells us that their differences should not be too large: the unrestricted problem admits a solution that is a spline in each component; the restricted problem simply forces these splines to have knots at the input points; and the trend filtering problem swaps splines for falling

factorial functions, which are highly similar in form. Next, we give theory that confirms this intuition, in large samples.

3.3 Error bounds

We derive error bounds for additive trend filtering and additive locally adaptive regression splines (both the unrestricted and restricted variants), when the underlying regression function is additive, and has components whose derivatives are of bounded variation. These results are actually special cases of a more general result we prove in this section, on a generic roughness-regularized additive estimator, where we assume a certain decay for the entropy of the unit ball in the roughness operator. We treat separately the settings in which the dimension d of the input space is fixed and growing. We also complement our error rates with minimax lower bounds. We start by introducing helpful notation.

3.3.1 Notation

Given a distribution Q supported on a set D, and i.i.d. samples X^i, $i = 1, \dots, n$ from Q, denote by Q_n the associated empirical distribution. We define the $L_2(Q)$ and $L_2(Q_n)$ inner products, denoted $\langle \cdot, \cdot \rangle_{L_2(Q)}$ and $\langle \cdot, \cdot \rangle_{L_2(Q_n)}$, respectively, over functions $m, r : D \to \mathbb{R}$

$$\langle m, r \rangle_{L_2(Q)} = \int_D m(x) r(x) \, dQ(x), \quad \text{and} \quad \langle m, r \rangle_{L_2(Q_n)} = \frac{1}{n} \sum_{i=1}^n m(X^i) r(X^i).$$

Definitions for the corresponding $L_2(Q)$ and $L_2(Q_n)$ norms, denoted $\| \cdot \|_{L_2(Q)}$ and $\| \cdot \|_{L_2(Q_n)}$, respectively, arise naturally from these inner products, defined by

$$\|m\|_2^2 = \langle m, m \rangle_2 = \int_D m(x)^2 \, dQ(x), \quad \text{and} \quad \|m\|_n^2 = \langle m, m \rangle_n = \frac{1}{n} \sum_{i=1}^n m(X^i)^2.$$

Henceforth, we will abbreviate subscripts when using these norms and inner products, writing $\| \cdot \|_2$ and $\| \cdot \|_n$ for the $L_2(Q)$ and $L_2(Q_n)$ norms, respectively, and similarly for the inner products. This abbreviated notation omits the underlying distribution Q; thus, unless explicitly stated otherwise, the underlying distribution should always be interpreted as the distribution of the input points. We will often call $\| \cdot \|_2$ the L_2 norm and $\| \cdot \|_n$ the empirical norm, and similarly for inner products.

In what follows, of particular interest will be the case when $D = [0, 1]^d$, and $m : [0, 1]^d \to \mathbb{R}$ is an additive function, of the form

$$m = \sum_{j=1}^d m_j,$$

which we write to mean $m(x) = \sum_{j=1}^d m_j(x_j)$. In a slight abuse of notation (overload of notation), for each $j = 1, \dots, d$, we will abbreviate the $L_2(Q_j)$ norm by $\| \cdot \|_2$, where Q_j is the jth marginal of Q, and will also abbreviate $L_2(Q_{jn})$ norm by $\| \cdot \|_n$, where Q_{jn} is the empirical distribution of X_j^i, $i = 1, \dots, n$. We will use similar abbreviations for the inner products.

A few more general definitions are in order. We denote the L_∞ norm, also called the sup norm, of a function $f : D \to \mathbb{R}$ by $\|f\|_\infty = \text{ess sup}_{z \in D} |f(z)|$. For a functional ν, acting on functions from D to $\mathbb{R}$, we write $B_\nu(\delta)$ for the ν-ball of radius $\delta > 0$, i.e., $B_\nu(\delta) = \{f : \nu(f) \leq \delta\}$. We abbreviate $B_n(\delta)$ for the $\|\cdot\|_n$-ball of radius δ, $B_2(\delta)$ for the $\|\cdot\|_2$-ball of radius δ, and $B_\infty(\delta)$ for the $\|\cdot\|_\infty$-ball of radius δ. We will use these concepts fluidly, without explicit reference to the domain D (or its dimensionality), as the meaning should be clear from the context.

Lastly, for a set S and norm $\|\cdot\|$, we define the covering number $N(\delta, \|\cdot\|, S)$ to be the smallest number of $\|\cdot\|$-balls of radius δ to cover S, and the packing number $M(\delta, \|\cdot\|, S)$ to be the largest number of disjoint $\|\cdot\|$-balls of radius δ that are contained in S. We call $\log N(\delta, \|\cdot\|, S)$ the entropy number.

3.3.2 Error bounds for a fixed dimension d

We consider error bounds for the generic roughness-penalized estimator defined as a solution of

$$\begin{aligned} \min_{f_j \in \mathcal{S}_j,\, j=1,\ldots,d} \quad & \frac{1}{2} \sum_{i=1}^n \left(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \right)^2 + \lambda \sum_{j=1}^d J(f_j) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d, \end{aligned} \tag{3.18}$$

where $\mathcal{S}_j$, $j = 1, \ldots, d$ are univariate function spaces, and J is a regularizer that acts on univariate functions. We assume in this subsection that the dimension d of the input space is fixed, i.e., it does not grow with n. Before stating our main result in this setting, we list our other assumptions, starting with our assumptions on the data generation process.

Assumption A1. The input points X^i, $i = 1, \ldots, n$ are i.i.d. from a continuous distribution Q supported on $[0, 1]^d$.

Assumption B1. The responses Y^i, $i = 1, \ldots, n$ follow the model

$$Y^i = \mu + f_0(X^i) + \epsilon^i, \quad i = 1, \ldots, n,$$

with overall mean $\mu \in \mathbb{R}$, where $\sum_{i=1}^n f_0(X^i) = 0$ for identifiability. The errors ϵ^i, $i = 1, \ldots, n$ are uniformly sub-Gaussian and have mean zero, i.e.,

$$\mathbb{E}(\epsilon) = 0, \quad \text{and} \quad \mathbb{E}[\exp(v^T \epsilon)] \leq \exp(\sigma^2 \|v\|_2^2 / 2), \quad \text{for all } v \in \mathbb{R}^n,$$

for a constant $\sigma > 0$. The errors and input points are independent.

Next, we present our assumptions on the regularizer J. We write $\|\cdot\|_{Z_n}$ for the empirical norm defined over a set of univariate points $Z_n = \{z^1, \ldots, z^n\} \subseteq [0, 1]$, i.e., $\|g\|_{Z_n}^2 = \frac{1}{n} \sum_{i=1}^n g^2(z^i)$.

Assumption C1. The regularizer J is a seminorm, and its domain is contained in the space of k times weakly differentiable functions, for an integer $k \geq 0$. Furthermore, its null space contains all kth order polynomials.

Assumption C2. There is a constant $L > 0$ such that

$$\operatorname*{ess\,sup}_{t\in[0,1]} g^{(k)}(t) - \operatorname*{ess\,inf}_{t\in[0,1]} g^{(k)}(t) \leq L, \quad \text{for } g \in B_J(1),$$

where $g^{(k)}$ is the kth weak derivative of g.

Assumption C3. There are constants $0 < w < 2$ and $K > 0$ such that

$$\sup_{Z_n=\{z^1,\ldots,z^n\}\subseteq[0,1]} \log N\big(\delta, \|\cdot\|_{Z_n}, B_J(1) \cap B_\infty(1)\big) \leq K\delta^{-w}.$$

We now state our main result in the fixed d case, which is proved in Appendix B.5, B.6.

Theorem 3.1. Assume A1, B1 on the data distribution, and assume C1, C2, C3 on the seminorm J. Also, assume that the dimension d of the input space is fixed. Let $C_n \geq 1$ be an arbitrary sequence. There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on d, σ, k, L, K, w, such that for all $c \geq c_1$, $n \geq n_0$, and tuning parameter values $\lambda \geq cn^{w/(2+w)}C_n^{-(2-w)/(2+w)}$, any solution in (3.18) satisfies

$$\bigg\|\sum_{j=1}^d \hat{f}_j - f_0\bigg\|_n^2 \leq \bigg\|\sum_{j=1}^d \tilde{f}_j - f_0\bigg\|_n^2 + \frac{6\lambda}{n}\max\bigg\{C_n, \sum_{j=1}^d J(\tilde{f}_j)\bigg\}, \tag{3.19}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$, simultaneously over all $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, feasible for the problem (3.18), such that $\|\tilde{f} - f_0\|_n \leq \max\{C_n, \sum_{j=1}^d J(\tilde{f}_j)\}$.

Remark 3.2 (Error bound for additive, J-smooth f_0). Assume $f_0 = \sum_{j=1}^d f_{0j}$, where $f_{0j} \in \mathcal{S}_j$, $j = 1, \ldots, d$, and $\sum_{j=1}^d J(f_{0j}) \leq C_n$. Letting $\tilde{f} = f_0$, the approximation error term in (3.19) (the first term on the right-hand side) is zero, and for $\lambda = cn^{w/(2+w)}C_n^{-(2-w)/(2+w)}$, the result in the theorem reads

$$\bigg\|\sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j}\bigg\|_n^2 \leq 6cn^{-2/(2+w)}C_n^{2w/(2+w)}, \tag{3.20}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$. As we will see in the minimax lower bound in Theorem 3.3 (plugging in $c_n = C_n/d$, and taking d to be a constant), the rate $n^{-2/(2+w)}C_n^{2w/(2+w)}$ is optimal for such a class of functions.

Remark 3.3 (Distance to best additive, J-smooth approximation of f_0). The arguments used to establish the oracle-type inequality (3.19) also imply a result on the empirical norm error between $\hat{f}$ and the best additive approximation of f_0. To be precise, let $(f_1^{\text{best}}, \ldots, f_d^{\text{best}})$ denote a solution in the population-level problem

$$\begin{aligned} \min_{f_j\in\mathcal{S}_j,\, j=1,\ldots,d} \quad & \frac{1}{2}\sum_{i=1}^n \bigg(f_0(X^i) - \sum_{j=1}^d f_j(X_j^i)\bigg)^2 + \frac{\lambda}{2}\sum_{j=1}^d J(f_j) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d. \end{aligned} \tag{3.21}$$

We note that (3.21) has "half" of the regularization of problem (3.18), as it uses a penalty parameter of $\lambda/2$ versus λ. We can think of $f^{\text{best}} = \sum_{j=1}^d f_j^{\text{best}}$ as the best additive, J-smooth approximation of f_0, where λ as usual controls the level of smoothness. The following is a consequence of the proof of Theorem 3.1, verified in Appendix B.7: assume that $\|f^{\text{best}} - f_0\|_n \leq \max\{C_n, \sum_{j=1}^d J(f_j^{\text{best}})\}$ almost surely (with respect to Q), for sufficiently large λ; then any solution in (3.18) satisfies for all $c \geq c_1$, $n \geq n_0$, and $\lambda \geq cn^{w/(2+w)} C_n^{-(2-w)/(2+w)}$,

$$\bigg\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_j^{\text{best}} \bigg\|_n^2 \leq \frac{6\lambda}{n} \max \bigg\{ C_n, \sum_{j=1}^d J(f_j^{\text{best}}) \bigg\}, \tag{3.22}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$, where as before $c_1, c_2, c_3, n_0 > 0$ are constants that depend only on d, σ, k, L, K, w. Notably, the right-hand side in the bound (3.22) does not depend on the approximation error; in particular, we do not even require $\|f^{\text{best}} - f_0\|_n$ to converge to zero. This is analogous to classical results from Stone (1985).

We examine a special case of the generic problem (3.18) when the regularizer is $J(g) = \text{TV}(g^{(k)})$, and derive implications of the above Theorem 3.1 for additive locally regression adaptive splines and additive trend filtering, corresponding to different choices of the function classes $\mathcal{S}_j$, $j = 1, \ldots, d$ in (3.18). We must introduce an additional (weak) assumption on the input distribution, for the results on restricted locally adaptive regression splines and trend filtering.

Assumption A2. The density of the input distribution Q is bounded below by a constant $b_0 > 0$.

Here is our result for additive locally adaptive splines and additive trend filtering. The proof is given in Appendix B.8, B.9.

Corollary 3.1. Assume A1, B1 on the data distribution. Also, assume that the dimension d of the input space is fixed, and that the underlying regression function is additive, $f_0 = \sum_{j=1}^d f_{0j}$, where the components f_{0j}, $j = 1, \ldots, d$ are k times weakly differentiable, such that $\sum_{j=1}^d \text{TV}(f_{0j}^{(k)}) \leq C_n$ for a sequence $C_n \geq 1$. For $J(g) = \text{TV}(g^{(k)})$, Assumptions C1, C2, C3 hold with $L = 1$ and $w = 1/(k+1)$. Furthermore, the following is true of the estimator defined by problem (3.18).

(a) Let $\mathcal{S}_j$ be the set of all k times weakly differentiable functions, for each $j = 1, \ldots, d$. There are constants $c_1, c_2, c_3, n_0 > 0$, depending only on d, σ, k, such that for all $c \geq c_1$ and $n \geq n_0$, any solution in the additive locally adaptive regression spline problem (3.18), with tuning parameter value $\lambda = cn^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)}$, satisfies

$$\bigg\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \bigg\|_n^2 \leq cn^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}, \tag{3.23}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$.

(b) Let $\mathcal{S}_j = \mathcal{G}_j$, the set of kth degree splines with knots in the set T_j in (3.17), for $j = 1, \ldots, d$, and assume A2 on the input density. Then there are constants $c_1, c_2, c_3, n_0 > 0$, that depend only on d, b_0, σ, k, such that for all $c \geq c_1$ and $n(\log n)^{-(1+1/k)} \geq n_0 C_n^{(2k+2)/(2k^2+2k-1)}$, any solution in the restricted additive locally adaptive spline problem (3.18), with $\lambda = cn^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)}$, satisfies the same result in (3.23), with probability at least $1 - \exp(-c_2 c) - c_3/n$.

(c) Let $\mathcal{S}_j = \mathcal{H}_j$, the set of kth degree falling factorial functions defined over X_j (the jth dimension of inputs), for $j = 1, \ldots, d$, and assume A2. Then there exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on d, b_0, σ, k, such that for all $c \geq c_1$ and $n(\log n)^{-(2k+3)} \geq n_0 C_n^{4k+4}$, any solution in the additive trend filtering problem (3.18), with $\lambda = cn^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)}$, satisfies (3.23), with probability at least $1 - \exp(-c_2 c) - c_3/n$.

Remark 3.4 (**Spline and falling factorial approximants**). For part (a) of the corollary, the approximation error (the first term on the right-hand side) in (3.20) is zero by definition, and we need only verify Assumptions C1, C2, C3 for the regularizer $J(g) = \mathrm{TV}(g^{(k)})$. Parts (b) and (c) require control over the approximation error, because the underlying regression function $f_0 = \sum_{j=1}^d f_{0j}$ need not have components that lie in the chosen function spaces $\mathcal{S}_j$, $j = 1, \ldots, d$. To be clear: for $k = 0$ or $k = 1$, as discussed in Section 3.2.5, all three problems considered in parts (a), (b), (c) are equivalent; hence parts (b) and (c) really only concern the case $k \geq 2$. For both of these parts, we control the approximation error by controlling the univariate approximation error and then applying the triangle inequality. For part (b), we use a special spline quasi-interpolant from Proposition 7 in Mammen & van de Geer (1997) (who in turn construct this using results from de Boor (1978)); for part (c), we develop a new falling factorial approximant that may be of independent interest.

3.3.3 Error bounds for a growing dimension d

In this subsection, we allow the input dimension d to grow with the sample size n. To keep our analysis as clean as possible, we consider a constrained version of the problem (3.18), namely

$$\begin{aligned} \min_{f_j \in \mathcal{S}_j,\, j=1,\ldots,d} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \Big)^2 \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \; J(f_j) \leq \delta, \quad j = 1, \ldots, d, \end{aligned} \tag{3.24}$$

for a tuning parameter $\delta > 0$. (The penalized problem (3.18) can also be analyzed in the setting of growing d, but we find that the analysis is messier and requires more assumptions in order to obtain the same results.) Instead of A1, we now use the following assumption in the input distribution.

Assumption A3. The input points X^i, $i = 1, \ldots, n$ are i.i.d. from a continuous distribution Q supported on $[0,1]^d$, that decomposes as $Q = Q_1 \times \cdots \times Q_d$, where the density of each Q_j is lower and upper bounded by constants $b_1, b_2 > 0$, for $j = 1, \ldots, d$.

Assumption A3 is fairly restrictive, since it requires the input distribution Q to be independent across dimensions of the input space. The reason we use this assumption: when $Q = Q_1 \times \cdots \times Q_d$, additive functions enjoy a key decomposability property in terms of the (squared) L_2 norm defined with respect to Q. In particular, if $m = \sum_{j=1}^d m_j$ has components with L_2 mean zero, denoted by $\bar{m}_j = \int_0^1 m_j(x_j)\, dQ_j(x_j) = 0$, $j = 1, \ldots, d$, then we have

$$\Big\| \sum_{j=1}^d m_j \Big\|_2^2 = \sum_{j=1}^d \|m_j\|_2^2. \tag{3.25}$$

This is explained by the fact that each pair of components m_j, m_ℓ with $j \neq \ell$ are orthogonal with respect to the L_2 inner product, since

$$\langle m_j, m_\ell \rangle_2 = \int_{[0,1]^2} m_j(x_j) m_\ell(x_\ell)\, dQ_j(x_j)\, dQ_\ell(x_\ell) = \bar{m}_j \bar{m}_\ell = 0.$$

The above orthogonality, and thus the decomposability property in (3.25), is only true because of the product form $Q = Q_1 \times \cdots \times Q_d$. Such decomposability is not generally possible with the empirical norm. In the proof of Theorem 3.2, we move from considering the empirical norm of the error vector to the L_2 norm, in order to leverage the property in (3.25), which eventually leads to an error rate that has a linear dependence on the dimension d. In the absence of L_2 decomposability, the same error rate can be achieved with a weaker incoherence bound, as in (3.30); see Remark 3.7 after the theorem.

We now state our main result in the growing d case, whose proof is in Appendix B.10, B.11.

Theorem 3.2. Assume A3, B1 on the data distribution, and assume C1, C2, C3 on the seminorm J. Let $\delta \geq 1$ be arbitrary. There are constants $c_1, c_2, c_3, n_0 > 0$, that depend only on $b_1, b_2, \sigma, k, L, K, w$, such that for all $c \geq c_1$ and $n \geq n_0 (d\delta)^{1+w/2}$, any solution in (3.24) satisfies both

$$\Big\| \sum_{j=1}^d \hat{f}_j - f_0 \Big\|_n^2 \leq \Big\| \sum_{j=1}^d \tilde{f}_j - f_0 \Big\|_n^2 + cdn^{-2/(2+w)}\delta, \tag{3.26}$$

$$\Big\| \sum_{j=1}^d \hat{f}_j - f_0 \Big\|_2^2 \leq 2 \Big\| \sum_{j=1}^d \tilde{f}_j - f_0 \Big\|_2^2 + 24 \Big\| \sum_{j=1}^d \tilde{f}_j - f_0 \Big\|_n^2 + cdn^{-2/(2+w)}\delta^2, \tag{3.27}$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$, simultaneously over all functions $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, feasible for the problem (3.24).

Remark 3.5 (Error bound for additive, J-smooth f_0). Assume $f_0 = \sum_{j=1}^d f_{0j}$, where $f_{0j} \in \mathcal{S}_j$ and $J(f_{0j}) \leq c_n$, $j = 1, \ldots, d$, for a sequence $c_n \geq 1$. Letting $\tilde{f} = f_0$, and $\delta = c_n$, the results in (3.26), (3.27) translate to

$$\Big\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \Big\|_n^2 \leq cdn^{-2/(2+w)} c_n, \quad \text{and} \quad \Big\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \Big\|_2^2 \leq cdn^{-2/(2+w)} c_n^2, \tag{3.28}$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$, provided that $n \geq n_0 (dc_n)^{1+w/2}$. From the minimax lower bound in Theorem 3.3, we can see that the optimal rate for such a class of

functions is in fact $dn^{-2/(2+w)}c_n^{2w/(2+w)}$, which reveals that the rates in (3.28) are tight when c_n is a constant, but not when c_n grows with n. It is worth noting that the dependence of the bounds on c_n in Theorem 3.2 (and hence in (3.28)) can be improved to have the optimal scaling of $c_n^{2w/(2+w)}$ by assuming that f_0 is sup norm bounded, and additionally placing a sup norm bound on the components in (3.24). This feels like an unnecessary restriction, so we prefer to present results without it, as in Theorem 3.2 (and (3.28)).

Remark 3.6 (Distance to best additive, J-smooth approximation of f_0). A consequence of the proof of (3.26) is a bound on the empirical norm error between $\hat{f}$ and the best additive approximation of f_0. To be precise, let $f^{\text{best}} = \sum_{j=1}^d f_j^{\text{best}}$ minimize $\|\sum_{j=1}^d \tilde{f}_j - f_0\|_n^2$ over all additive functions $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$ feasible for problem (3.24). Then following directly from (B.36) in the proof of Theorem 3.2, we have for all $c \geq c_1$ and $n \geq n_0(d\delta)^{1+w/2}$,

$$\bigg\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_j^{\text{best}} \bigg\|_n^2 \leq cdn^{-2/(2+w)}\delta, \tag{3.29}$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$, where again $c_1, c_2, c_3, n_0 > 0$ are constants that depend on $b_1, b_2, \sigma, k, L, K, w$. Just as we saw in fixed d case, the right-hand side in (3.29) does not depend on the approximation error $\|f^{\text{best}} - f_0\|_n$, which is analogous to classical results from Stone (1985).

Remark 3.7 (L_2 decomposability and incoherence). The decomposability property in (3.25) is critical in obtaining the sharp (linear) dependence on d in the error rates (3.26), (3.27). However, it is worth noting that all that is needed in the proof is in fact a lower bound of the form

$$\bigg\| \sum_{j=1}^d m_j \bigg\|_2^2 \geq \phi_0 \sum_{j=1}^d \|m_j\|_2^2, \tag{3.30}$$

for a constant $\phi_0 > 0$, rather than an equality, as in (3.25). The above is an incoherence condition that can hold for nonproduct distributions Q, over an appropriate class of functions (additive functions with smooth components), provided that the correlations between components of Q are not too large. See Meier et al. (2009), van de Geer (2014) for similar incoherence conditions.

Next we present our results for additive locally adaptive regression splines (both unresricted and restricted variants) and additive trend filtering. The proof is in Appendix B.12.

Corollary 3.2. Assume A3, B1 on the data distribution. Also, assume that the underlying regression function is additive, $f_0 = \sum_{j=1}^d f_{0j}$, where the components f_{0j}, $j = 1, \dots, d$ are k times weakly differentiable, such that $\text{TV}(f_{0j}^{(k)}) \leq c_n$, $j = 1, \dots, d$, for a sequence $c_n \geq 1$. Then for $J(g) = \text{TV}(g^{(k)})$, the following is true of the estimator defined by problem (3.24).

(a) Let $\mathcal{S}_j$ be the space of all k times weakly differentiable functions, for each $j = 1, \dots, d$. There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on b_1, b_2, σ, k, such that for

all $c \geq c_1$ and $n \geq n_0 (dc_n)^{(2k+3)/(2k+2)}$, any solution in the constrained-form additive locally adaptive spline problem (3.24), with tuning parameter $\delta = c_n$, satisfies

$$\left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_n^2 \leq cdn^{-\frac{2k+2}{2k+3}} c_n, \quad \text{and} \quad \left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_2^2 \leq cdn^{-\frac{2k+2}{2k+3}} c_n^2, \quad (3.31)$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$.

(b) Let $\mathcal{S}_j = \mathcal{G}_j$, the set of kth degree splines with knots in the set T_j in (3.17), for $j = 1, \ldots, d$. There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on b_1, b_2, σ, k, such that for $c \geq c_1$ and $n \geq (dc_n)^{(2k+3)/(2k+2)}$, any solution in the constrained-form restricted additive locally adaptive spline problem (3.24), with tuning parameter $\delta = a_k c_n$, where $a_k \geq 1$ is a constant that depends only on k, satisfies (3.31), with probability at least $1 - \exp(-c_2 c) - c_3 d/n$.

(c) Let $\mathcal{S}_j = \mathcal{H}_j$, the set of kth degree falling factorial functions defined over X_j (the jth dimension of input points), for $j = 1, \ldots, d$. Then there are constants $c_1, c_2, c_3, n_0 > 0$, depending only on b_1, b_2, σ, k, such that for all $c \geq c_1$ and $n \geq n_0 (dc_n)^{(2k+3)/(2k+2)}$, any solution in the constrained-form additive trend filtering problem (3.24), with tuning parameter $\delta = a_k c_n$, where $a_k \geq 1$ is a constant depending only on k, satisfies (3.31), with probability at least $1 - \exp(-c_2 c) - c_3 d/n$.

3.3.4 Minimax lower bounds

We consider minimax lower bounds for estimation over the class of additive functions whose components are smooth with respect to the seminorm J. We allow the dimension d to grow with n. As for the data distribution, we will use the following assumptions in place of A1, A2, A3, B1.

Assumption A4. The inputs X^i, $i = 1, \ldots, n$ are i.i.d. from the uniform distribution on $[0, 1]^d$.

Assumption B2. The responses Y^i, $i = 1, \ldots, n$ follow

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

with mean $\mu \in \mathbb{R}$, where $\int_{[0,1]^d} f_0(x)\, dx = 0$ for identifiability. The errors ϵ^i, $i = 1, \ldots, n$ are i.i.d. $N(0, \sigma^2)$, for some constant $\sigma > 0$. The errors and input points are independent.

For the regularizer J, assumed to satisfy Assumptions C1, C2, we will replace Assumption C3 by the following assumption, on the log packing and log covering (entropy) numbers.

Assumption C4. There exist constants $0 < w < 2$ and $K_1, K_2 > 0$ such that

$$\log M\big(\delta, \|\cdot\|_2, B_J(1) \cap B_\infty(1)\big) \geq K_1 \delta^{-w},$$
$$\log N\big(\delta, \|\cdot\|_2, B_J(1) \cap B_\infty(1)\big) \leq K_2 \delta^{-w}.$$

(To be clear, here $\|\cdot\|_2$ is the L_2 norm defined with respect to the uniform distribution on $[0, 1]$.)

Let us introduce the notation

$$B_J^d(\delta) = \bigg\{ \sum_{j=1}^d f_j : J(f_j) \leq \delta,\ j = 1, \ldots, d \bigg\},$$

Now we state our main minimax lower bound. The proof is given in Appendix B.13, B.14.

Theorem 3.3. Assume A4, B2 on the data distribution, and C1, C2, C4 on the seminorm J. Then there exist constants $c_0, n_0 > 0$, that depend only on $\sigma, k, L, K_1, K_2, w$, such that for all $c_n \geq 1$ and $n \geq n_0 d^{1+w/2} c_n^w$, we have

$$\inf_{\hat{f}} \sup_{f_0 \in B_J^d(c_n)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \geq c_0 d n^{-2/(2+w)} c_n^{2w/(2+w)}. \tag{3.32}$$

When we choose $J(g) = \mathrm{TV}(g^{(k)})$ as our regularizer, the additive function class $B_J^d(\delta)$ becomes

$$\mathcal{F}_k^d(\delta) = \bigg\{ \sum_{j=1}^d f_j : \mathrm{TV}(f_j^{(k)}) \leq \delta,\ j = 1, \ldots, d \bigg\},$$

and Theorem 3.3 implies the following result, whose proof is in Appendix B.15.

Corollary 3.3. Assume A4, B2 on the data distribution. Assume further that f_{0j}, $j = 1, \ldots, d$ are k times weakly differentiable. Then there are constants $c_0, n_0 > 0$, that depend only on σ, k, such that for all $c_n \geq 1$ and and $n \geq n_0 d^{(2k+3)/(2k+2)} c_n^{1/(k+1)}$,

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_k^d(c_n)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \geq c_0 d n^{-(2k+2)/(2k+3)} c_n^{2/(2k+3)}. \tag{3.33}$$

Remark 3.8 (Optimality for a fixed dimension d). For a fixed d, the estimator defined by (3.18) is minimax rate optimal over the class of additive functions f_0 such that $\sum_{j=1}^d J(f_{0j}) \leq C_n$. To see this, note that such a class of functions contains $B_J^d(C_n/d)$, therefore plugging $c_n = C_n/d$ into the right-hand side in (3.32) yields a lower bound rate of $n^{-2/(2+w)} C_n^{2w/(2+w)}$, which matches the upper bound rate in (3.20).

Furthermore, when $J(g) = \mathrm{TV}(g^{(k)})$, the lower bound rate given by plugging $c_n = C_n/d$ into the right-hand side in (3.33) is $n^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}$, matching the upper bound rate in (3.23). Hence additive locally adaptive regression splines, restricted additive locally adaptive regression splines, and additive trend filtering all achieve the minimax rate over the space of additive functions f_0 such that $\sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}) \leq C_n$.

Remark 3.9 (Optimality for a growing dimension d). For growing d, the estimator defined by (3.24) is minimax rate optimal over the class of additive functions f_0 such that $J(f_{0j}) \leq c$, $j = 1, \ldots, d$, where $c > 0$ is a constant. This is verified by noting that the lower bound rate of $dn^{-2/(2+w)}$ in (3.32) matches the upper bound rates in (3.26), (3.27).

When $J(g) = \mathrm{TV}(g^{(k)})$, and again, $c_n = c$ (a constant), the lower bound rate of $dn^{-(2k+2)/(2k+3)}$ in (3.33) matches the upper bound rates in (3.31). Thus additive locally adaptive regression splines, restricted additive locally adaptive regression splines, and additive trend filtering all attain the minimax rate over the space of additive functions f_0 with $\mathrm{TV}(f_{0j}^{(k)}) \leq c$, $j = 1, \ldots, d$.

For growing c_n, we note that the upper bounds in (3.28) and (3.31) have an inflated dependence on c_n, compared to (3.32) and (3.33). It turns out that the latter (lower bounds) are tight, and the former (upper bounds) are loose. The upper bounds can be tightened under further boundedness assumptions (see Remark 3.5).

Remark 3.10 (**Suboptimality of additive linear smoothers).** Seminal theory from Donoho & Johnstone (1998) on minimax linear rates over Besov spaces shows that, under Assumption B2, and with the inputs X^i, $i = 1, \dots, n$ being now nonrandom and occurring over the regular d-dimensional lattice $\{1/N, 2/N, \dots, 1\}^d \subseteq [0,1]^d$ with $N = n^{1/d}$, we have

$$\inf_{\hat{f} \text{ additive linear}} \sup_{f_0 \in \mathcal{F}_k^d(c_n)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \geq c_0 d n^{-(2k+1)/(2k+2)} c_n^{2/(2k+2)}, \tag{3.34}$$

for all $n \geq n_0$, where $c_0, n_0 > 0$ are constants, depending only on σ, k. On the left-hand side in (3.34) the infimum is taken over all additive linear smoothers, i.e., estimators $\hat{f} = \sum_{j=1}^d \hat{f}_j$ such that each component $\hat{f}_j$ is a linear smoother, for $j = 1, \dots, d$. The additive linear smoother lower bound (3.34) is verified in Appendix B.16.

For a fixed d, we can see that all additive linear smoothers—e.g., additive smoothing splines, additive kernel smoothing estimators, additive RKHS estimators, etc.—are suboptimal over the class of additive functions f_0 with $\sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}) \leq C_n$, as the optimal linear rate in (3.34) (set $c_n = C_n/d$) is $n^{-(2k+1)/(2k+2)} C_n^{2/(2k+2)}$, slower than the optimal rate $n^{-(2k+2)/(2k+3)} C_n^{2/(2k+2)}$ of additive locally adaptive splines and additive trend filtering in (3.23).

For growing d, and $c_n = c$ (a constant), we also see that additive linear smoothers are suboptimal over the class of additive functions f_0 such that $\mathrm{TV}(f_{0j}^{(k)}) \leq c$, $j = 1, \dots, d$, as the optimal linear rate in (3.34) is $d n^{-(2k+1)/(2k+2)}$, slower than the optimal rate $d n^{-(2k+2)/(2k+3)}$ of additive locally adaptive regression splines and additive trend filtering in (3.31).

3.4 Backfitting and the dual

We now examine computational approaches for the additive trend filtering problem (3.4). This is a convex optimization problem, and many standard approaches can be applied. For its simplicity and its ubiquity in additive modeling, we focus on the backfitting algorithm in particular.

3.4.1 Backfitting

The backfitting approach for problem (3.4) is described in Algorithm 1. We write $\mathrm{TF}_\lambda(r, X_j)$ for the univariate trend filtering fit, with a tuning parameter $\lambda > 0$, to a response vector $r = (r^1, \dots, r^n) \in \mathbb{R}^n$ over an input vector $X_j = (X_j^1, \dots, X_j^n) \in \mathbb{R}^n$. In words, the algorithm cycles over $j = 1, \dots, d$, and at each step updates the estimate for component j by applying univariate trend filtering to the jth partial residual (i.e., the current residual excluding component j). Centering in Step 2b part (ii) is optional, because the fit $\mathrm{TF}_\lambda(r, X_j)$ will have mean zero whenever r has mean zero, but centering can still be performed for numerical stability. In general, the efficiency of backfitting hinges on the efficiency of the univariate smoother employed; to implement Algorithm 1 in practice we can use fast interior

Algorithm 1 Backfitting for additive trend filtering

Given responses $Y^i \in \mathbb{R}$ and input points $X^i \in \mathbb{R}^d$, $i = 1, \ldots, n$.

1. Set $t = 0$ and initialize $\theta_j^{(0)} = 0$, $j = 1, \ldots, d$.
2. For $t = 1, 2, 3, \ldots$ (until convergence):
   a. For $j = 1, \ldots, d$:
      (i) $\theta_j^{(t)} = \mathrm{TF}_\lambda\Big(Y - \bar{Y}\mathbb{1} - \sum_{\ell<j} \theta_j^{(t)} - \sum_{\ell>j} \theta_j^{(t-1)},\ X_j\Big)$
      (ii) (Optional) $\theta_j^{(t)} = \theta_j^{(t)} - \frac{1}{n}\mathbb{1}^T\theta_j^{(t)}$
3. Return $\hat{\theta}_j$, $j = 1, \ldots, d$ (parameters $\theta_j^{(t)}$, $j = 1, \ldots, d$ at convergence).

point methods (Kim et al. 2009) or fast operator splitting methods (Ramdas & Tibshirani 2016) for univariate trend filtering, both of which result in efficient empirical performance.

Algorithm 1 is equivalent to block coordinate descent (BCD), also called exact blockwise minimization, applied to problem (3.4) over the coordinate blocks θ_j, $j = 1, \ldots, d$. A general treatment of BCD is given in Tseng (2001), who shows that for a convex criterion that decomposes into smooth plus separable terms, as does that in (3.4), all limit points of the sequence of iterates produced by BCD are optimal solutions. We are primarily interested in developing a connection between BCD for problem (3.4) and alternating projections in its dual problem (3.14), which is the topic of the next subsection.

3.4.2 Dual alternating projections

Using the additive trend filtering problem (3.4) and its dual (3.14), related by the transformation (3.15), we see that for any dimension $j = 1, \ldots, d$, the univariate trend filtering fit with response vector $r = (r^1, \ldots, r^n)$ and input vector $X_j = (X_j^1, \ldots, X_j^n)$ becomes

$$\mathrm{TF}_\lambda(r, X_j) = (\mathrm{Id} - \Pi_{U_j})(r), \tag{3.35}$$

where $U_j = \{S_j D_j^T v_j : \|u\|_\infty \leq \lambda\}$, and recall, we abbreviate $D_j = D^{(X_j, k+1)}$. Reparametrizing in terms of the primal-dual relationship $u = Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j$ (and ignoring the optional centering step), the backfitting approach in Algorithm 1 can thus be viewed as performing the updates, for $t = 1, 2, 3, \ldots$,

$$\begin{aligned} u_0^{(t)} &= Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j^{(t-1)}, \\ u_j^{(t)} &= \Pi_{U_j}\big(u_{j-1}^{(t)} + \theta_j^{(t-1)}\big), \quad j = 1, \ldots, d, \\ \theta_j^{(t)} &= \theta_j^{(t-1)} + u_{j-1}^{(t)} - u_j^{(t)}, \quad j = 1, \ldots, d. \end{aligned} \tag{3.36}$$

Thus the backfitting algorithm for (3.4), as expressed above in (3.36), is seen to be a particular type of *alternating projections* method applied to the dual problem (3.14), cycling

through projections onto U_j, $j = 1, \dots, d$. Interestingly, as opposed to the classical alternating projections approach, which would repeatedly project the current iterate $u_{j-1}^{(t)}$ onto U_j, $j = 1, \dots, d$, the steps in (3.36) repeatedly project an "offset" version $u_{j-1}^{(t)} + \theta_j^{(t-1)}$ of the current iterate, for $j = 1, \dots, d$.

3.4.3 Parallelized backfitting

We have seen that backfitting is a special type of alternating projections algorithm, applied to the dual problem (3.14). For set intersection problems (where we seek a point in the intersection of given closed, convex sets), the optimization literature offers a variety of *parallel projections* methods (in contrast to alternating projections methods) that are provably convergent. One such method can be derived using ADMM (e.g., see Section 5.1 of Boyd et al. (2011)), and a similar construction can be used for the dual problem (3.14). We first rewrite this problem as

$$\begin{aligned} \min_{u_0, u_1, \dots, u_d \in \mathbb{R}^n} \quad & \frac{1}{2}\|Y - \bar{Y}\mathbb{1} - u_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) \\ \text{subject to} \quad & u_0 = u_1, \; u_0 = u_2, \; \dots \; u_0 = u_d, \end{aligned} \tag{3.37}$$

where we write I_S for the indicator function of a set S (equal to 0 on S, and ∞ otherwise). Then we define the augmented Lagrangian, for an arbitrary $\rho > 0$, as

$$\begin{aligned} &L_\rho(u_0, u_1, \dots, u_d, \gamma_1, \dots, \gamma_d) = \\ &\qquad \frac{1}{2}\|Y - \bar{Y}\mathbb{1} - u_0\|_2^2 + \sum_{j=1}^d \bigg(I_{U_j}(u_j) + \frac{\rho}{2}\|u_0 - u_j + \gamma_j\|_2^2 - \frac{\rho}{2}\|\gamma_j\|_2^2 \bigg). \end{aligned}$$

The ADMM steps for (3.37) are now given by repeating, for $t = 1, 2, 3, \dots$,

$$\begin{aligned} u_0^{(t)} &= \frac{1}{\rho d + 1}\bigg(Y - \bar{Y}\mathbb{1} + \rho \sum_{j=1}^d (u_j^{(t-1)} - \gamma_j^{(t-1)}) \bigg) \\ u_j^{(t)} &= \Pi_{U_j}\big(u_0^{(t)} + \gamma_j^{(t-1)} \big), \quad j = 1, \dots, d \\ \gamma_j^{(t)} &= \gamma_j^{(t-1)} + u_0^{(t)} - u_j^{(t)}, \quad j = 1, \dots, d. \end{aligned} \tag{3.38}$$

Now compare (3.38) to (3.36)—the key difference is that in (3.38), the updates to u_j, $j = 1, \dots, d$, i.e., the projections onto U_j, $j = 1, \dots, d$, completely decouple and can hence be performed *in parallel*. Run properly, this could provide a large speedup over the sequential projections in (3.36).

Of course, for our current study, the dual problem (3.37) is really only interesting insofar as it is connected to the additive trend filtering problem (3.4). In Algorithm 2, we transcribe the iterations in (3.38) into an equivalent primal form, and we provide a convergence guarantee in the next theorem. For details, see Appendix B.17.

Theorem 3.4. Initialized arbitrarily, the ADMM steps (3.38) produce parameters $\hat{\gamma}_j$, $j = 1, \dots, d$ (i.e., the iterates $\gamma_j^{(t)}$, $j = 1, \dots, d$ at convergence) such that the scaled parameters $\rho\hat{\gamma}_j$, $j = 1, \dots, d$ solve additive trend filtering (3.4). Further, the outputs $\hat{\theta}_j$, $j = 1, \dots, d$ of Algorithm 2 solve additive trend filtering (3.4).

Algorithm 2 Parallel backfitting for additive trend filtering

Given responses $Y^i \in \mathbb{R}$, input points $X^i \in \mathbb{R}^d$, $i = 1, \dots, n$, and $\rho > 0$.

1. Initialize $u_0^{(0)} = 0$, $\theta_j^{(0)} = 0$ and $\theta_j^{(-1)} = 0$ for $j = 1, \dots, d$.
2. For $t = 1, 2, 3, \dots$ (until convergence):
   a. $u_0^{(t)} = \frac{1}{\rho d + 1}\left(Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j^{(t-1)}\right) + \frac{\rho d}{\rho d + 1}\left(u_0^{(t-1)} + \frac{1}{\rho d}\sum_{j=1}^d \left(\theta_j^{(t-2)} - \theta_j^{(t-1)}\right)\right)$
   b. For $j = 1, \dots, d$ (in parallel):
      (i) $\theta_j^{(t)} = \rho \cdot \mathrm{TF}_\lambda\big(u_0^{(t)} + \theta_j^{(t-1)}/\rho, X_j\big)$
      (ii) (Optional) $\theta_j^{(t)} = \theta_j^{(t)} - \frac{1}{n}\mathbb{1}^T\theta_j^{(t)}$
3. Return $\hat{\theta}_j$, $j = 1, \dots, d$ (parameters $\theta_j^{(t)}$, $j = 1, \dots, d$ at convergence).

Written in primal form, we see that the parallel backfitting approach in Algorithm 2 differs from what may be considered the "naive" approach to parallelizing the usual backfitting iterations in Algorithm 1. Consider $\rho = 1$. If we were to replace Step 2a in Algorithm 2 with $u_0^{(t)} = r^{(t-1)}$, the full residual

$$r^{(t-1)} = Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j^{(t-1)},$$

then the update steps for $\theta_j^{(t)}$, $j = 1, \dots, d$ that follow would be just given by applying univariate trend filtering to each partial residual (without sequentially updating the partial residuals between trend filtering runs). This naive parallel method has no convergence guarantees, and can fail even in simple practical examples to produce optimal solutions. Importantly, Algorithm 2 does not take $u_0^{(t)}$ to be the full residual, but as Step 2a shows, uses a less greedy choice: it basically takes $u_0^{(t)}$ to be a convex combination of the residual $r^{(t-1)}$ and its previous value $u_0^{(t-1)}$, with higher weight on the latter. The subsequent parallel updates for $\theta_j^{(t)}$, $j = 1, \dots, d$ are still given by univariate trend filtering fits, and though these steps do not exactly use partial residuals (since $u_0^{(t)}$ is not exactly the full residual), they are guaranteed to produce additive trend filtering solutions upon convergence (as per Theorem 3.4). An example of cyclic versus parallelized backfitting is given in Appendix B.18.

3.5 Experiments

Through empirical experiments, we examine the performance of additive trend filtering relative to additive smoothing splines. We also examine the efficacy of cross-validation for choosing the tuning parameter λ, as well as the use of multiple tuning parameters. All experiments were performed in R. For the univariate trend filtering solver, we used the

`trendfilter` function in the `glmgen` package; for the univariate smoothing spline solver, we used the `smooth.spline` function in base R.

3.5.1 Simulated heterogeneously-smooth data

We sampled $n = 2500$ input points in $d = 10$ dimensions, by assigning the inputs along each dimension $X_j = (X_j^1, \dots, X_j^n)$ to be a different permutation of the equally spaced points $(1/n, 2/n, \dots, 1)$, for $j = 1, \dots, 10$. For the componentwise trends, we examined sinusoids with Doppler-like spatially-varying frequencies:

$$g_{0j}(x_j) = \sin\left(\frac{2\pi}{(x_j + 0.1)^{j/10}}\right), \quad j = 1, \dots, 10.$$

We then defined the component functions as $f_{0j} = a_j g_{0j} - b_j$, $j = 1, \dots, d$, where a_j, b_j were chosen so that f_{0j} had empirical mean zero and empirical norm $\|f_{0j}\|_n = 1$, for $j = 1, \dots, d$. The responses were generated according to $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^d f_{0j}(X_j^i), \sigma^2)$, $i = 1, \dots, 2500$. By construction, in this setup, there is considerable heterogeneity in the levels of smoothness both within and between the component functions.

The left panel of Figure 3.4 shows a comparison of the MSE curves from additive trend filtering in (3.4) (of quadratic order, $k = 2$) and additive smoothing splines in (3.1) (of cubic order). We set σ^2 in the generation of the responses so that the signal-to-noise ratio (SNR) was $\|f_0\|_n^2/\sigma^2 = 4$, where $f_0 = \sum_{j=1}^d f_{0j}$. The two methods (additive trend filtering and additive smoothing splines) were each allowed their own sequence of tuning parameter values, and results were averaged over 10 repetitions from the simulation setup described above. As we can see, additive trend filtering achieves a better minimum MSE along its regularization path, and does so at a less complex model (lower df).

The right panel of Figure 3.4 shows the best-case MSEs for additive trend filtering and additive smoothing splines (i.e., the minimum MSE over their regularization paths) as the noise level σ^2 is varied so that the SNR ranges from 0.7 to 1.6, in equally spaced values on the log scale. The results were again averaged over 10 repetitions of data drawn from a simulation setup essentially the same as the one described above, except that we considered a smaller problem size, with $n = 1000$ and $d = 6$. The plot reveals that additive trend filtering performs increasingly well (in comparison to additive smoothing splines) as the SNR grows—not surprising, as here it is able to better capture the heterogeneity in the component functions.

Lastly, in Appendix B.19, we present results from an experimental setup mimicking that in this subsection, except with the component functions f_{0j}, $j = 1, \dots, d$ having homogeneous smoothness throughout. Here additive trend filtering and additive smoothing splines perform very similarly.

3.5.2 Cross-validation and multiple tuning parameters

Sticking to the simulation setup from the last subsection, but at the smaller problem size, $n = 1000$ and $d = 6$ (used to produce the right panel of Figure 3.4), we study in the left panel of Figure 3.5 the use of 5-fold cross-validation (CV) to select the tuning parameter λ for additive trend filtering and additive smoothing splines. Displayed are the resulting

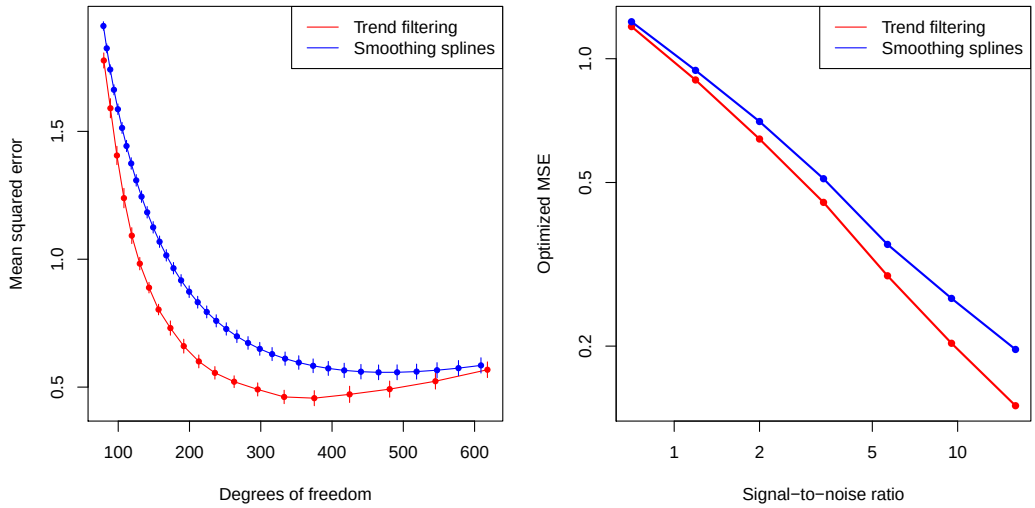

Figure 3.4: *The left panel shows the MSE curves for additive trend filtering (3.4) (of quadratic order) and additive smoothing splines (3.1) (of cubic order), computed over 10 repetitions from the heterogeneous smoothness simulation with* $n = 2500$ *and* $d = 10$*, described in Section 3.5.1, where the SNR is set to 4. Vertical segments denote* ± 1 *standard deviations. The right panel displays the best-case MSE for each method (the minimum MSE over its regularization path), in a problem setup with* $n = 1000$ *and* $d = 6$*, as the signal-to-noise ratio (SNR) varies from 0.7 to 16, in equally spaced values on the log scale.*

MSE curves as the SNR varies from 0.7 to 16. Also shown on the same plot are the oracle MSE curves (which are the same as those the right panel of Figure 3.4), in which λ has been chosen to minimize the MSE for each method. We can see that the performance of each method degrades using CV, but not by much.

In the right panel of the figure, we examine the use of multiple tuning parameters for additive smoothing splines and additive trend filtering, i.e., replacing the penalties in (3.1) and (3.4) by

$$\sum_{j=1}^{d} \lambda_j \theta_j^T Q_j \theta_j \quad \text{and} \quad \sum_{j=1}^{d} \lambda_j \big\| D^{(X_j, k+1)} S_j \theta_j \big\|_1,$$

respectively, which means we would now have d tuning parameters λ_j, $j = 1, \dots, d$. When the function we are estimating has different amounts of smoothness along different dimensions, we have argued (and seen through examples) that additive trend filtering—using only a single tuning parameter λ—can accomodate these differences, at least somewhat, thanks to its locally adaptive nature. But, when these differences in smoothness are drastic enough, it may be worthwhile to use multiple tuning parameters.

When d is moderate (even just for $d = 6$), cross-validation over a d-dimensional grid of values for λ_j, $j = 1, \dots, d$ can be prohibitive. However, as pointed out by a referee of this article, there has been a considerable amount of work dedicated to this problem by authors studying additive models built from splines (or other linear smoothers), e.g., Gu & Wahba (1991), Wood (2000), Fahrmeir & Lang (2001), Ruppert et al. (2003), Wood (2004), Kim & Gu (2004), Rue et al. (2009), Wood (2011), Wood et al. (2015, 2016). Many of these papers

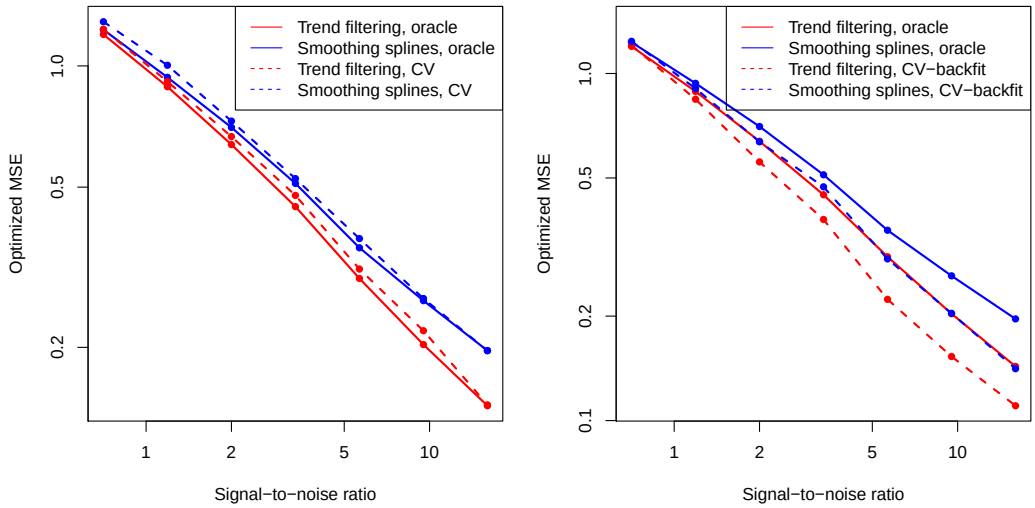

Figure 3.5: *Both panels display results from the same simulation setup as that in the right panel of Figure 3.4. The left panel shows MSE curves when the estimators are tuned by 5-fold cross-validation (CV), and also by the oracle (reflecting the minimum possible MSE). The right panel displays MSE curves when we allow each estimator to have d tuning parameters, tuned by a hybrid backfit-CV method explained in the text, versus the oracle MSE curves for a single tuning parameter.*

use an efficient computational approach based on restricted maximum likelihood (REML) for selecting λ_j, $j = 1, \ldots, d$; see also Wood (2017) for a nice introduction and description of this approach. Unfortunately, as far as we see it, REML does not easily apply to additive trend filtering.

We thus use the following simple approach for multiple tuning parameter selection: within each backfitting loop, for each component $j = 1, \ldots, d$, we use (univariate) CV to choose λ_j. While this does not solve a particular convex optimization problem, and is not guaranteed to converge in general, we have found it to work quite well in practice. The right panel of Figure 3.5 compares the performance of this so-called backfit-CV tuning to the oracle, that chooses just a single tuning parameter. Both additive trend filtering and additive smoothing splines are seen to improve with d tuning parameters, tuned by backfit-CV, in comparison to the oracle choice of tuning parameter. Interestingly, we also see that additive smoothing splines with d tuning parameters performs on par with additive trend filtering with the oracle choice of tuning parameter. (In this example, REML tuning for additive smoothing splines—as implemented by the `mgcv` R package—performed worse than backfit-CV tuning, and so we only show results from the latter.)

3.6 Discussion

We have studied additive models built around the univariate trend filtering estimator, i.e., defined by penalizing according to the sum of ℓ_1 norms of discrete derivatives of the component functions. We examined basic properties of these additive models, such as extrapolation of the fitted values to a d-dimensional surface, and uniqueness of the component fits. When the underlying regression function is additive, with components whose kth

derivatives are of bounded variation, we derived error rates for kth order additive trend filtering: $n^{-(2k+2)/(2k+3)}$ for a fixed input dimension d (under weak assumptions), and $dn^{-(2k+2)/(2k+3)}$ for a growing dimension d (under stronger assumptions). We showed these rates are sharp by establishing matching minimax lower bounds. On the computational side, we devised a provably convergent parallel backfitting algorithm for additive trend filtering. It is worth noting that our parallel backfitting method is not specific to additive trend filtering, but it can be embedded in a more general parallel coordinate descent framework (Tibshirani 2017).

A natural extension of our work is to consider the high-dimensional case, where d is comparable or possibly even much larger than n, and we fit a *sparse additive model* by employing an additional sparsity penalty in problem (3.4). Another natural extension is to consider responses $Y^i|X^i$, $i = 1, \ldots, n$ from an exponential family distribution, and we fit a *generalized additive model* by changing the loss in (3.4). After we completed an initial version of this paper, both extensions have been pursued: Tan & Zhang (2017) develop a suite of error bounds for sparse additive models, with various form of penalties (which include total variation on derivatives of components); and Haris et al. (2018) give comprehensive theory for sparse generalized additive models, with various types of penalties (which again include total variation on derivatives of components).

3.7 Extensions to exponential family losses

As is well-known (see Hastie & Tibshirani (1990)), additive models can be extended to exponential family distributions by simple link functions. Additive trend filtering can be extended to fit exponential family distributions as follows. Consider the following problem where $g : \mathbb{R} \to [0, \infty)$ is a link function:

$$\begin{aligned} \min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \quad & -(Y - \bar{Y}\mathbb{1})^T \sum_{j=1}^d \theta_j + \sum_{i=1}^n g\Big(\sum_{j=1}^d \theta_j^i\Big) + \lambda \sum_{j=1}^d \big\|D^{(X_j,k+1)} S_j \theta_j\big\|_1 \\ \text{subject to} \quad & \mathbb{1}^T \theta_j = 0, \quad j = 1, \ldots, d. \end{aligned} \tag{3.39}$$

Setting $g(x) = x^2/2, g(x) = \log(1 + e^x)$ and $g(x) = e^x$ in the above problem gives penalized maximum likelihood estimators for Gaussian, logistic and Poisson loss models respectively. Note that the Gaussian loss problem is equivalent to the original formulation (3.4).

Backfitting may again be used to solve (3.39) with a proximal Newton method to solve the inner component-wise problems. We are interested in deriving the minimax optimal rates over the additive function space $B_J^d(c_n), c_n \geq 1$ where J is the semi-norm in Theorem 3.3 and show that the estimator in (3.39) attains the rate up to logarithmic factors.

Chapter 4

A Higher Order Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test (Kolmogorov 1933, Smirnov 1948) is a classical and celebrated tool for nonparametric hypothesis testing. Let $x_1, \dots, x_m \sim P$ and $y_1, \dots, y_n \sim Q$ be independent samples. Let $X_{(m)}$ and $Y_{(n)}$ denote the two sets of samples, and also let $Z_{(N)} = X_{(m)} \cup Y_{(n)} = \{z_1, \dots, z_N\}$, where $N = m + n$. The two-sample KS test statistic is defined as

$$\max_{z \in Z_{(m+n)}} \left| \frac{1}{m} \sum_{i=1}^{m} 1\{x_i \le z\} - \frac{1}{n} \sum_{i=1}^{n} 1\{y_i \le z\} \right|. \tag{4.1}$$

In words, this measures the maximum absolute difference between the empirical cumulative distribution functions (CDFs) of $X_{(m)}$ and $Y_{(n)}$, across all points in the joint sample $Z_{(m+n)}$. Naturally, the two-sample KS test rejects the null hypothesis of $P = Q$ for large values of the statistic. The statistic (4.1) can also be written in the following variational form:

$$\sup_{f \,:\, \mathrm{TV}(f) \le 1} |\mathbb{P}_m f - \mathbb{Q}_n f|, \tag{4.2}$$

where $\mathrm{TV}(\cdot)$ denotes total variation, and we define the empirical expectation operators $\mathbb{P}_m, \mathbb{Q}_n$ via

$$\mathbb{P}_m f = \frac{1}{m} \sum_{i=1}^{m} f(x_i) \;\text{ and }\; \mathbb{Q}_n f = \frac{1}{n} \sum_{i=1}^{n} f(y_i).$$

Later, we will give a general representation result that implies the equivalence of (4.1) and (4.2) as a special case.

The KS test is a fast, general-purpose two-sample nonparametric test. But being a general-purpose test also means that it is systematically less sensitive to some types of differences, such as tail differences (Bryson 1974). Intuitively, this is because the empirical CDFs of $X_{(m)}$ and $Y_{(n)}$ must both tend to 0 as $z \to -\infty$ and to 1 as $z \to \infty$, so the gap in the tails will not be large.

The insensitivity of the KS test to tail differences is well-known. Several authors have proposed modifications to the KS test to improve its tail sensitivity, based on variance-reweighting (Anderson & Darling 1952), or Renyi-type statistics (Mason & Schuenemeyer

1983, Calitz 1987), to name a few ideas. In a different vein, Wang et al. (2014) recently proposed a higher-order extension of the KS two-sample test, which replaces the total variation constraint on f in (4.2) with a total variation constraint on a derivative of f. These authors show empirically that, in some cases, this modification can lead to better tail sensitivity. In the current work, we refine the proposal of Wang et al. (2014), and give theoretical backing for this new test.

4.1 A higher-order KS test

Our test statistic has the form of an integral probability metric (IPM). For a function class $\mathcal{F}$, the IPM between distributions P and Q, with respect to $\mathcal{F}$, is defined as (Muller 1997)

$$\rho(P, Q; \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{P}f - \mathbb{Q}f| \tag{4.3}$$

where we define the expectation operators $\mathbb{P}, \mathbb{Q}$ by

$$\mathbb{P}f = \mathbb{E}_{X \sim P}[f(X)] \quad \text{and} \quad \mathbb{Q}f = \mathbb{E}_{Y \sim Q}[f(Y)].$$

For a given function class $\mathcal{F}$, the IPM $\rho(\cdot, \cdot\,; \mathcal{F})$ is a pseudometric on the space of distributions. Note that the KS test in (4.2) is precisely $\rho(P_m, Q_n; \mathcal{F}_0)$, where P_m, Q_n are the empirical distributions of $X_{(m)}, Y_{(n)}$, respectively, and $\mathcal{F}_0 = \{f : \mathrm{TV}(f) \leq 1\}$.

Consider an IPM given by replacing $\mathcal{F}_0$ with $\mathcal{F}_k = \{f : \mathrm{TV}(f^{(k)}) \leq 1\}$, for an integer $k \geq 1$ (where we write $f^{(k)}$ for the kth weak derivative of f). Some motivation is as follows. In the case $k = 0$, we know that the *witness functions* in the KS test (4.2), i.e., the functions in $\mathcal{F}_0$ that achieve the supremum, are piecewise constant step functions (cf. the equivalent representation (4.1)). These functions can only have so much action in the tails. By moving to $\mathcal{F}_k$, which is essentially comprised of the kth order antiderivative of functions in $\mathcal{F}_0$, we should expect that the witness functions over $\mathcal{F}_k$ are kth order antiderivatives of piecewise constant functions, i.e., kth degree piecewise polynomial functions, which can have much more sensitivity in the tails.

But simply replacing $\mathcal{F}_0$ by $\mathcal{F}_k$ and proposing to compute $\rho(P_m, Q_n; \mathcal{F}_k)$ leads to an ill-defined test. This is due to the fact that $\mathcal{F}_k$ contains all polynomials of degree k. Hence, if the ith moments of P_m, Q_n differ, for any $i \in [k]$ (where we abbreviate $[a] = \{1, \ldots, a\}$ for an integer $a \geq 1$), then $\rho(P_m, Q_n; \mathcal{F}_k) = \infty$.

As such, we must modify $\mathcal{F}_k$ to control the growth of its elements. While there are different ways to do this, not all result in computable IPMs. The approach we take yields an exact representer theorem (generalizing the equivalence between (4.1) and (4.2)). Define

$$\begin{aligned} \mathcal{F}_k = \big\{ & f : \mathrm{TV}(f^{(k)}) \leq 1, \\ & f^{(j)}(0) = 0, \ j \in \{0\} \cup [k-1], \\ & f^{(k)}(0+) = 0 \text{ or } f^{(k)}(0-) = 0 \big\}. \end{aligned} \tag{4.4}$$

Here $f^{(k)}(0+)$ and $f^{(k)}(0-)$ denote one-sided limits at 0 from above and below, respectively. Informally, the functions in $\mathcal{F}_k$ are pinned down at 0, with all lower-order derivatives (and

the limiting kth derivative from the right or left) equal to 0, which limits their growth. Now we define the *kth-order KS test statistic* as

$$\rho(P_m, Q_n; \mathcal{F}_k) = \sup_{f \in \mathcal{F}_k} |\mathbb{P}_m f - \mathbb{Q}_n f|. \tag{4.5}$$

An important remark is that for $k = 0$, this recovers the original KS test statistic (4.2), because $\mathcal{F}_0$ contains all step functions of the form $g_t(x) = 1\{x \leq t\}$, $t \geq 0$.

Another important remark is that for any $k \geq 0$, the function class $\mathcal{F}_k$ in (4.4) is "rich enough" to make the IPM in (4.5) a metric. We state this formally next; its proof, as with all other proofs in this chapter, is in Appendix C.

Proposition 4.1. For any $k \geq 0$, and any P, Q with k moments, $\rho(P, Q; \mathcal{F}_k) = 0$ if and only if $P = Q$.

Motivating Example. Figure 4.1 shows the results of a simple simulation comparing the proposed higher-order tests (4.5), of orders $k = 1$ through 5, against the usual KS test (corresponding to $k = 0$). For the simulation setup, we used $P = N(0, 1)$ and $Q = N(0, 1.44)$. For 500 "alternative" repetitions, we drew $m = 250$ samples from P, drew $n = 250$ samples from Q, and computed test statistics; for another 500 "null" repetitions, we permuted the $m + n = 500$ samples from the corresponding alternative repetition, and again computed test statistics. For each test, we varied the rejection threshold for each test, we calculated its true positive rate using the alternative repetitions, and calculated its false positive rate using the null repetitions. The oracle ROC curve corresponds to the likelihood ratio test (which knows the exact distributions P, Q). Interestingly, we can see that power of the higher-order KS test improves as we increase the order from $k = 0$ up to $k = 2$, then stops improving by $k = 3, 4, 5$.

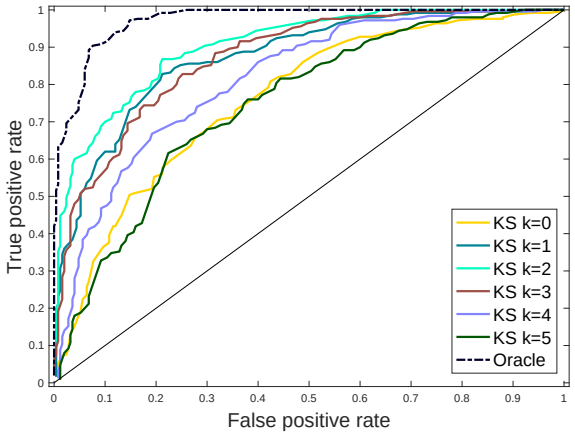

Figure 4.1: *ROC curves from an experiment comparing the proposed higher-order KS tests in* (4.5) *(for various k) to the usual KS test, when $P = N(0, 1)$ and $Q = N(0, 1.44)$.*

Figure 4.2 displays the witness function (which achieves the supremum in (4.5)) for a large-sample version of the higher-order KS test, across orders $k = 0$ through 5. We used the same distributions as in Figure 4.1, but now $n = m = 10^4$. We will prove in Section 4.2

that, for the kth order test, the witness function is always a kth degree piecewise polynomial (in fact, a rather simple one, of the form $g_t(x) = (x-t)_+^k$ or $g_t(x) = (t-x)_+^k$ for a knot t). Recall the underlying distributions P, Q here have different variances, and we can see from their witness functions that all higher-order KS tests choose to put weight on tail differences. Of course, the power of any test of is determined by the size of the statistic under the alternative, relative to typical fluctuations under the null. As we place more weight on tails, in this particular setting, we see diminishing returns at $k = 3$, meaning the null fluctuations must be too great.

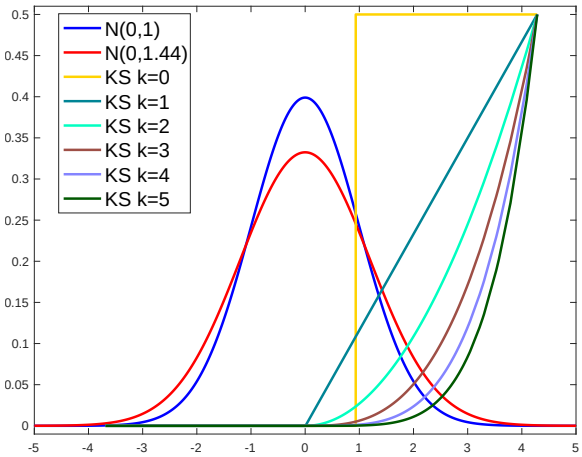

Figure 4.2: *Witness functions (normalized for plotting purposes) for the higher-order KS tests, when $P = N(0,1)$ and $Q = N(0, 1.44)$. They are always of piecewise polynomial form; and here they all place weight on tail differences.*

Summary of Contributions. Our contributions in this work are as follows.

- We develop an exact representer theorem for the higher-order KS test statistic (4.5). This enables us to compute the test statistic in linear-time, for all $k \leq 5$. For $k \geq 6$, we develop a nearly linear-time approximation to the test statistic.
- We derive the asymptotic null distribution of the our higher-order KS test statistic, based on empirical process theory. For $k \geq 6$, our approximation to the test statistic has the same asymptotic null.
- We provide concentration tail bounds for the test statistic. Combined with the metric property from Proposition 4.1, this shows that our higher-order KS test is asymptotically powerful against any pair of fixed, distinct distributions P, Q.
- We perform extensive numerical studies to compare the newly proposed tests with several others.

Other Related Work. Recently, IPMs have been gaining in popularity due in large part to energy distance tests (Szekely & Rizzo 2004, Baringhaus & Franz 2004) and kernel

maximum mean discrepancy (MMD) tests (Gretton et al. 2012), and in fact, there is an equivalence between the two classes (Sejdinovic et al. 2013). An IPM with a judicious choice of $\mathcal{F}$ gives rise to a number of common distances between distributions, such as Wasserstein distance or total variation (TV) distance. While IPMs look at differences $dP - dQ$, tests based on ϕ-divergences (such as Kullback-Leibler, or Hellinger) look at ratios dP/dQ, but can be hard to efficiently estimate in practice (Sriperumbudur et al. 2009). The TV distance is the only IPM that is also a ϕ-divergence, but it is impossible to estimate.

There is also a rich class of nonparametric tests based on graphs. Using minimum spanning trees, Friedman & Rafsky (1979) generalized both the Wald-Wolfowitz runs test and the KS test. Other tests are based on k-nearest neighbors graphs (Schilling 1986, Henze 1988) or matchings (Rosenbaum 2005). The Mann-Whitney-Wilcoxon test has a multivariate generalization using the concept of data depth (Liu & Singh 1993). Bhattacharya (2016) established that many computationally efficient graph-based tests have suboptimal statistical power, but some inefficient ones have optimal scalings.

Different computational-statistical tradeoffs were also discovered for IPMs (Ramdas, Reddi, Poczos, Singh & Wasserman 2015). Further, as noted by Janssen (2000) (in the context of one-sample testing), every nonparametric test is essentially powerless in an infinity of directions, and has nontrivial power only against a finite subspace of alternatives. In particular, this implies that no single nonparametric test can uniformly dominate all others; improved power in some directions generally implies weaker power in others. This problem only gets worse in high-dimensional settings (Ramdas, Reddi, Pczos, Singh & Wasserman 2015, Arias-Castro et al. 2018). Therefore, the question of which test to use for a given problem must be guided by a combination of simulations, computational considerations, a theoretical understanding of the pros/cons of each test, and a practical understanding of the data at hand.

Outline. In Section 4.2, we give computational details for the higher-order KS test statistic (4.5). We derive its asymptotic null in Section 4.3, and give concentration bounds (for the statistic around the population-level IPM) in Section 4.4. We give numerical experiments in Section 4.5, and conclude in Section 4.6 with a discussion.

4.2 Computation

Write $T = \rho(P_m, Q_n; \mathcal{F}_k)$ for the test statistic in (4.5). In this section, we derive a representer theorem for T, develop a linear-time algorithm for $k \leq 5$, and a nearly linear-time approximation for $k \geq 6$.

4.2.1 Representer theorem

The higher-order KS test statistic in (4.5) is defined by an infinite-dimensional maximization over $\mathcal{F}_k$ in (4.4). Fortunately, we can restrict our attention to a simpler function class, as we show next.

Theorem 4.1. Fix $k \geq 0$. Let $g_t^+(x) = (x-t)_+^k/k!$ and $g_t^-(x) = (t-x)_+^k/k!$ for $t \in \mathbb{R}$, where we write $(a)_+ = \max\{a, 0\}$. For the statistic T defined by (4.5),

$$T = \max\Big\{ \sup_{t \geq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+|, \quad \sup_{t \leq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^-| \Big\}. \tag{4.6}$$

The proof of this theorem uses a key result from Mammen (1991), where it is shown that we can construct a spline interpolant to a given function at given points, such that its higher-order total variation is no larger than that of the original function.

Remark 4.1. When $k = 0$, note that for $t \geq 0$,

$$\begin{aligned} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+| &= \bigg| \frac{1}{m} \sum_{i=1}^m 1\{x_i > t\} - \frac{1}{n} \sum_{i=1}^n 1\{y_i > t\} \bigg| \\ &= \bigg| \frac{1}{m} \sum_{i=1}^m 1\{x_i \leq t\} - \frac{1}{n} \sum_{i=1}^n 1\{y_i \leq t\} \bigg| \end{aligned}$$

and similarly for $t \leq 0$, $|(\mathbb{P}_m - \mathbb{Q}_n) g_t^-|$ reduces to the same expression in the second line above. As we vary t from $-\infty$ to ∞, this only changes at values $t \in Z_{(N)}$, which shows (4.6) and (4.1) are the same, i.e., Theorem 4.1 recovers the equivalence between (4.2) and (4.1).

Remark 4.2. For general $k \geq 0$, we can interpret (4.6) as a comparison between truncated kth order moments, between the empirical distributions P_m and Q_n. The test statistic T the maximum over all possible truncation locations t. The critical aspect here is *truncation*, which makes the higher-order KS test statistic a metric (recall Proposition 4.1). A comparison of moments, alone, would not be enough to ensure such a property.

Theorem 4.1 itself does not immediately lead to an algorithm for computing T, as the range of t considered in the suprema is infinite. However, through a bit more work, detailed in the next two subsections, we can obtain an exact linear-time algorithm for all $k \leq 5$, and a linear-time approximation for $k \geq 6$.

4.2.2 Linear-time algorithm for $k \leq 5$

The key fact that we will exploit is that the criterion in (4.6), as a function of t, is a piecewise polynomial of order k with knots in $Z_{(N)}$. Assume without a loss of generality that $z_1 < \cdots < z_N$. Also assume without a loss of generality that $z_1 \geq 0$ (this simplifies notation, and the general case follows by the repeating the same arguments separately for the points in $Z_{(N)}$ on either side of 0). Define $c_i = \mathbb{1}\{z_i \in X_{(m)}\}/m - \mathbb{1}\{z_i \in Y_{(n)}\}/n$, $i \in [N]$, and

$$\phi_i(t) = \frac{1}{k!} \sum_{j=i}^N c_j (z_j - t)^k, \quad i \in [N]. \tag{4.7}$$

Then the statistic in (4.6) can be succinctly written as

$$T = \max_{i \in [N]} \sup_{t \in [z_{i-1}, z_i]} \phi_i(t), \tag{4.8}$$

where we let $z_0 = 0$ for convenience. Note each $\phi_i(t)$, $i \in [N]$ is a kth degree polynomial. We can compute a representation for these polynomials efficiently.

Lemma 4.1. Fix $k \geq 0$. The polynomials in (4.7) satisfy the recurrence relations

$$\phi_i(t) = \frac{1}{k!} c_i (z_i - t)^k + \phi_{i+1}(t), \quad i \in [N]$$

(where $\phi_{N+1} = 0$). Given the monomial expansion

$$\phi_{i+1}(t) = \sum_{\ell=0}^{k} a_{i+1,\ell} t^\ell,$$

we can compute an expansion for ϕ_i, with coefficients $a_{i\ell}$, $\ell \in \{0\} \cup [k]$, in $O(1)$ time. So we can compute all coefficients $a_{i,\ell}$, $i \in [N]$, $\ell \in \{0\} \cup [k]$ in $O(N)$ time.

To compute T in (4.8), we must maximize each polynomial ϕ_i over its domain $[z_{i-1}, z_i]$, for $i \in [N]$, and then compare maxima. Once we have computed a representation for these polynomials, as Lemma 4.1 ensures we can do in $O(N)$ time, we can use this to analytically maximize each polynomial over its domain, provided the order k is small enough. Of course, maximizing a polynomial over an interval can be reduced to computing the roots of its derivative, which is an analytic computation for any $k \leq 5$ (since the roots of any quartic have a closed-form, see, e.g., Rosen 1995). The next result summarizes.

Proposition 4.2. For any $0 \leq k \leq 5$, the test statistic in (4.8) can be computed in $O(N)$ time.

Maximizing a polynomial of degree $k \geq 6$ is not generally possible in closed-form. However, developments in semidefinite optimization allow us to approximate its maximum efficiently, investigated next.

4.2.3 Linear-time approximation for $k \geq 6$

Seminal work of Shor (1998), Nesterov (2000) shows that the problem of maximizing a polynomial over an interval can be cast as a semidefinite program (SDP). The number of variables in this SDP depends only on the polynomial order k, and all constraint functions are self-concordant. Using say an interior point method to solve this SDP, therefore, leads to the following result.

Proposition 4.3. Fix $k \geq 6$ and $\epsilon > 0$. For each polynomial in (4.7), we can compute an ϵ-approximation to its maximum in $c_k \log(1/\epsilon)$ time, for a constant $c_k > 0$ depending only on k. As we can compute a representation for all these polynomials in $O(N)$ time (Lemma 4.1), this means we can compute an ϵ-approximation to the statistic in (4.6) in $O(N \log(1/\epsilon))$ time.

Remark 4.3. Let T_ϵ denote the ϵ-approximation from Proposition 4.3. Under the null $P = Q$, we would need to have $\epsilon = o(1/\sqrt{N})$ in order for the approximation T_ϵ to share the asymptotic null distribution of T, as we will see in Section 4.3.3. Taking say, $\epsilon = 1/N$, the statistic $T_{1/N}$ requires $O(N \log N)$ computational time, and this is why in various places we make reference to a *nearly* linear-time approximation when $k \geq 6$.

4.2.4 Simple linear-time approximation

We conclude this section by noting a simple approximation to (4.6) given by

$$T^* = \max\Big\{ \max_{t \in Z^0_{(N)}, t \geq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+|, \quad \max_{t \in Z^0_{(N)}, t \leq 0} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^-| \Big\}, \tag{4.9}$$

where $Z^0_{(N)} = \{0\} \cup Z_{(N)}$. Clearly, for $k = 0$ or 1, the maximizing t in (4.6) must be one of the sample points $Z_{(N)}$, so $T^* = T$ and there is no approximation error in (4.9). For $k \geq 2$, we can control the error as follows.

Lemma 4.2. For $k \geq 2$, the statistics in (4.6), (4.9) satisfy

$$T - T^* = \frac{\delta_N}{(k-1)!} \bigg(\frac{1}{m} \sum_{i=1}^m |x_i|^{k-1} + \frac{1}{n} \sum_{i=1}^n |y_i|^{k-1} \bigg),$$

where δ_N is the maximum gap between sorted points in $Z^0_{(N)}$.

Remark 4.4. We would need to have $\delta_N = o_P(1/\sqrt{N})$ in order for T^* to share the asymptotic null of T, see again Section 4.3.3 (this is assuming that P has $k-1$ moments, so the sample moments concentrate for large enough N). This will not be true of δ_N, the maximum gap, in general. But it does hold when P is continuous, having compact support, and a density bounded from below on its support; here, in fact, $\delta_N = o_P(\log N/N)$ (see, e.g., Wang et al. 2014).

Although it does not have the strong guarantees of the approximation from Proposition 4.3, the statistic in (4.9) is simple and efficient—we must emphasize that it can be computed in $O(N)$ linear time, as a consequence of Lemma 4.1 (the evaluations of $\phi_i(t)$ at the sample points $t \in Z_{(N)}$ are the constant terms a_{i0}, $i \in [N]$ in their monomial expansions)—and is likely a good choice for most practical purposes.

4.3 Asymptotic null

To study the asymptotic null distribution of the proposed higher-order KS test, we will appeal to uniform central limit theorems (CLTs) from the empirical process theory literature, reviewed here for completeness. For functions f, g in a class $\mathcal{F}$, let $\mathbb{G}_{P,\mathcal{F}}$ denote a Gaussian process indexed by $\mathcal{F}$ with mean and covariance

$$\begin{gathered} \mathbb{E}(\mathbb{G}_{P,\mathcal{F}} f) = 0, \quad f \in \mathcal{F}, \\ \mathrm{Cov}(\mathbb{G}_{P,\mathcal{F}} f, \mathbb{G}_{P,\mathcal{F}} g) = \mathrm{Cov}_{X \sim P}(f(X), g(X)), \quad f, g \in \mathcal{F}. \end{gathered}$$

For functions l, u, let $[l, u]$ denote the set of functions $\{f : l(x) \leq f(x) \leq u(x), \text{ for all } x\}$. Call $[l, u]$ a *bracket* of size $\|u - l\|_2$, where $\|\cdot\|_2$ denotes the $L_2(P)$ norm, defined as

$$\|f\|_2^2 = \int f(x)^2 \, dP(x).$$

Finally, let $N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{F})$ be the smallest number of ϵ-sized brackets that are required to cover $\mathcal{F}$. Define the bracketing integral of $\mathcal{F}$ as

$$J_{[]}(\|\cdot\|_2, \mathcal{F}) = \int_0^1 \sqrt{\log N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{F})}\, d\epsilon.$$

Note that this is finite when $\log N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{F})$ grows slower than $1/\epsilon^2$. We now state an important uniform CLT from empirical process theory.

Theorem 4.2 (Theorem 11.1.1 in Dudley 1999)**.** If $\mathcal{F}$ is a class of functions with finite bracketing integral, then when $P = Q$ and $m, n \to \infty$, the process

$$\sqrt{\frac{mn}{m+n}}\{\mathbb{P}_m f - \mathbb{Q}_n f\}_{f \in \mathcal{F}}$$

converges weakly to the Gaussian process $\mathbb{G}_{P,\mathcal{F}}$. Hence,

$$\sqrt{\frac{mn}{m+n}} \sup_{f \in \mathcal{F}} |\mathbb{P}_m f - \mathbb{Q}_n f| \xrightarrow{d} \sup_{f \in \mathcal{F}} |\mathbb{G}_{P,\mathcal{F}} f|.$$

4.3.1 Bracketing integral calculation

To derive the asymptotic null of the higher-order KS test, based on its formulation in (4.5), and Theorem 4.2, we would need to bound the bracketing integral of $\mathcal{F}_k$. While there are well-known entropy (log covering) number bounds for related function classes (e.g., Birman & Solomyak 1967, Babenko 1979), and the conversion from covering to bracketing numbers is standard, these results unfortunately require the function class to be uniformly bounded in the sup norm, which is certainly not true of $\mathcal{F}_k$.

Note that the representer result in (4.6) can be written as $T = \rho(P_m, Q_n; \mathcal{G}_k)$, where

$$\mathcal{G}_k = \{g_t^+ : t \geq 0\} \cup \{g_t^- : t \leq 0\}. \tag{4.10}$$

We can hence instead apply Theorem 4.2 to $\mathcal{G}_k$, whose bracketing number can be bounded by direct calculation, assuming enough moments on P.

Lemma 4.3. Fix $k \geq 0$. Assume $\mathbb{E}_{X \sim P}|X|^{2k+\delta} \leq M < \infty$, for some $\delta > 0$. For the class $\mathcal{G}_k$ in (4.10), there is a constant $C > 0$ depending only on k, δ such that

$$\log N_{[]}(\epsilon, \|\cdot\|_2, \mathcal{G}_k) \leq C \log \frac{M^{1+\frac{\delta(k-1)}{2k+\delta}}}{\epsilon^{2+\delta}}.$$

4.3.2 Asymptotic null for higher-order KS

Applying Theorem 4.2 and Lemma 4.3 to the higher-order KS test statistic (4.6) leads to the following result.

Theorem 4.3. Fix $k \geq 0$. Assume $\mathbb{E}_{X \sim P}|X|^{2k+\delta} < \infty$, for some $\delta > 0$. When $P = Q$, the test statistic in (4.6) satisfies, as $m, n \to \infty$,

$$\sqrt{\frac{mn}{m+n}} T \xrightarrow{d} \sup_{g \in \mathcal{G}_k} |\mathbb{G}_{P,k} g|,$$

where $\mathbb{G}_{P,k}$ is an abbreviation for the Gaussian process indexed by the function class $\mathcal{G}_k$ in (4.10).

Remark 4.5. When $k = 0$, note that for $t \geq s \geq 0$, the covariance function is

$$\mathrm{Cov}_{X\sim P}(1\{X > s\}, 1\{X > t\}) = F_P(s)(1 - F_P(t)),$$

where F_P denotes the CDF of P. For $s \leq t \leq 0$, the covariance function is again equal to $F_P(s)(1 - F_P(t))$. The supremum of this Gaussian process over $t \in \mathbb{R}$ is that of a Brownian bridge, so Theorem 4.3 recovers the well-known asymptotic null distribution of the KS test, which (remarkably) does not depend on P.

Remark 4.6. When $k \geq 1$, it is not clear how strongly the supremum of the Gaussian process from Theorem 4.3 depends on P; it appears it must depend on the first k moments of P, but is not clear whether it *only* depends on these moments. Section 3.5 investigates empirically. Currently, we do not have a precise understanding of whether the asymptotic null is useable in practice, and we suggest using a permutation null instead.

4.3.3 Asymptotic null under approximation

The approximation from Proposition 4.3 shares the same asymptotic null, provided $\epsilon > 0$ is small enough.

Corollary 4.1. Fix $k \geq 0$. Assume $\mathbb{E}_{X\sim P}|X|^{2k+\delta} < \infty$, for some $\delta > 0$. When $P = Q$, as $m, n \to \infty$ such that m/n converges to a positive constant, the test statistic T_ϵ from Proposition 4.3 converges at a $\sqrt{N}$-rate to the supremum of the same Gaussian process in Theorem 4.3, provided $\epsilon = o(1/\sqrt{N})$.

The approximation in (4.9) shares the same asymptotic null, provided P is continuous with compact support.

Corollary 4.2. Fix $k \geq 0$. Assume that P is continuous, compactly supported, with density bounded from below on its support. When $P = Q$, as $m, n \to \infty$ such that m/n converges to a positive constant, the test statistic T^* in (4.9) converges at a $\sqrt{N}$-rate to the supremum of the same Gaussian process in Theorem 4.3.

4.4 Tail concentration

We examine the convergence of our test statistics to their population analogs. In general, if the population-level IPM $\rho(P, Q; \mathcal{F}_k)$ is large, then the concentration bounds below will imply that the empirical statistic $\rho(P_m, Q_n; \mathcal{F}_k)$ will be large for m, n sufficiently large, and the test will have power.

We first review the necessary machinery, again from empirical process theory. For $p \geq 1$, and a function f of a random variable $X \sim P$, recall the $L_p(P)$ norm is defined as $\|f\|_p = [\mathbb{E}(f(X)^p)]^{1/p}$. For $p > 0$, recall the *exponential Orlicz norm* of order p is defined as

$$\|f\|_{\Psi_p} = \inf\left\{t > 0 : \mathbb{E}[\exp(|X|^p/t^p)] - 1 \leq 1\right\}.$$

(These norms depend on the measure P, since they are defined in terms of expectations with respect to $X \sim P$, though this is not explicit in our notation.)

We now state an important concentration result.

Theorem 4.4 (Theorems 2.14.2 and 2.14.5 in van der Vaart & Wellner 1996)**.** Let $\mathcal{F}$ be a class functions with an envelope function F, i.e., $f \leq F$ for all $f \in \mathcal{F}$. Define

$$W = \sqrt{n} \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f|,$$

and abbreviate $J = J_{[]}(\|\cdot\|, \mathcal{F})$. For $p \geq 2$, if $\|F\|_p < \infty$, then for a constant $c_1 > 0$,

$$[\mathbb{E}(W^p)]^{1/p} \leq c_1 \Big(\|F\|_2 J + n^{-1/2+1/p} \|F\|_p \Big),$$

and for $0 < p \leq 1$, if $\|F\|_{\Psi_p} < \infty$, then for a constant $c_2 > 0$,

$$\|W\|_{\Psi_p} \leq c_2 \Big(\|F\|_2 J + n^{-1/2} (1 + \log n)^{1/p} \|F\|_{\Psi_p} \Big).$$

The two-sample test statistic $T = \rho(P_m, Q_n; \mathcal{G}_k)$ satisfies (following by a simple argument using convexity)

$$|T - \rho(P, Q; \mathcal{F}_k)| \leq \rho(P, P_m; \mathcal{F}_k) + \rho(Q, Q_n; \mathcal{F}_k).$$

The terms on the right hand side can each be bounded by Theorem 4.4, where we can use the envelope function $F(x) = |x|^k / k!$ for $\mathcal{G}_k$. Using Markov's inequality, we can then get a tail bound on the statistic.

Theorem 4.5. Fix $k \geq 0$. Assume that P, Q both have p moments, where $p \geq 2$ and $p > 2k$. For the statistic in (4.6), for any $\alpha > 0$, with probability $1 - \alpha$,

$$|T - \rho(P, Q; \mathcal{G}_k)| \leq c(\alpha) \bigg(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \bigg),$$

where $c(\alpha) = c_0 \alpha^{-1/p}$, and $c_0 > 0$ is a constant. If P, Q both have finite exponential Orlicz norms of order $0 < p \leq 1$, then the above holds for $c(\alpha) = c_0 (\log(1/\alpha))^{1/p}$.

When we assume k moments, the population IPM for $\mathcal{F}_k$ also has a representer in $\mathcal{G}_k$; by Proposition 4.1, this implies $\rho(\cdot, \cdot\,; \mathcal{G}_k)$ is also a metric.

Corollary 4.3. Fix $k \geq 0$. Assuming P, Q both have k moments, $\rho(P, Q; \mathcal{F}_k) = \rho(P, Q; \mathcal{G}_k)$. Therefore, by Proposition 4.1, $\rho(\cdot, \cdot\,; \mathcal{G}_k)$ is a metric (over the space of distributions P, Q with k moments).

Putting this metric property together with Theorem 4.5 gives the following.

Corollary 4.4. Fix $k \geq 0$. For $\alpha_N = o(1)$ and $1/\alpha_N = o(N^{p/2})$, reject when the higher-order KS test statistic (4.6) satisfies $T > c(\alpha_N)(1/\sqrt{m} + 1/\sqrt{n})$, where $c(\cdot)$ is as in Theorem 4.5. For any P, Q that meet the moment conditions of Theorem 4.5, as $m, n \to \infty$ in such a way that m/n approaches a positive constant, we have type I error tending to 0, and power tending to 1, i.e., the higher-order KS test is *asymptotically powerful*.

4.5 Numerical experiments

We present numerical experiments that examine the convergence of our test statistic to its asymptotic null, its power relative to other general purpose nonparametric tests, and its power when P, Q have densities with local differences. Experiments comparing to the MMD test with a polynomial kernel are deferred to Appendix C.

Convergence to Asymptotic Null. In Figure 4.3, we plot histograms of finite-sample higher-order KS test statistics and their asymptotic null distributions, when $k = 1, 2$. We considered both $P = N(0, 1)$ and $P = \text{Unif}(-\sqrt{3}, \sqrt{3})$ (the uniform distribution standardized to have mean 0 and variance 1). For a total of 1000 repetitions, we drew two sets of samples from P, each of size $m = n = 2000$, then computed the test statistics. For a total of 1000 times, we also approximated the supremum of the Gaussian process from Theorem 4.3 via discretization. We see that the finite-sample statistics adhere closely to their asymptotic distributions. Interestingly, we also see that the distributions look roughly similar across all four cases considered. Future work will examine more thoroughly.

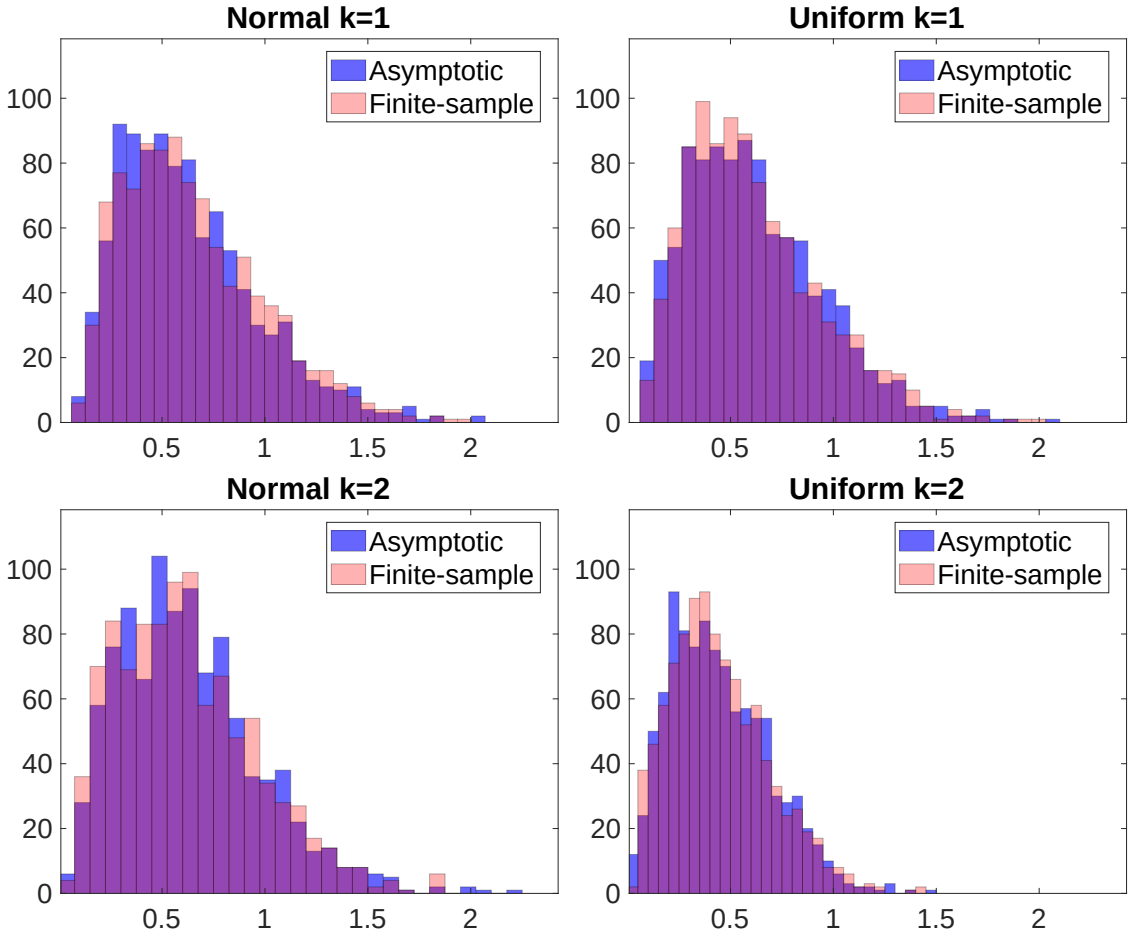

Figure 4.3: *Histograms comparing finite-sample test statistics to their asymptotic null distribution.*

Comparison to General-Purpose Tests. In Figures 4.4 and 4.5, we compare the higher-order KS tests to the KS test, and other widely-used nonparametric tests from the literature: the kernel maximum mean discrepancy (MMD) test (Gretton et al. 2012) with a Gaussian kernel, the energy distance test (Szekely & Rizzo 2004), and the Anderson-Darling test (Anderson & Darling 1954). The simulation setup is the same as that in the introduction, where we considered P, Q with different variances, except here we study different means: $P = N(0, 1)$, $Q = N(0.2, 1)$, and different third moments: $P = N(0, 1)$, $Q = t(3)$, where $t(3)$ denotes Student's t-distribution with 3 degrees of freedom. The higher-order KS tests generally perform favorably, and in each setting there is a choice of k that yields better power than KS. In the mean difference setting, this is $k = 1$, and the power degrades for $k = 3, 5$, likely because these tests are "smoothing out" the mean difference too much; see Proposition 4.4.

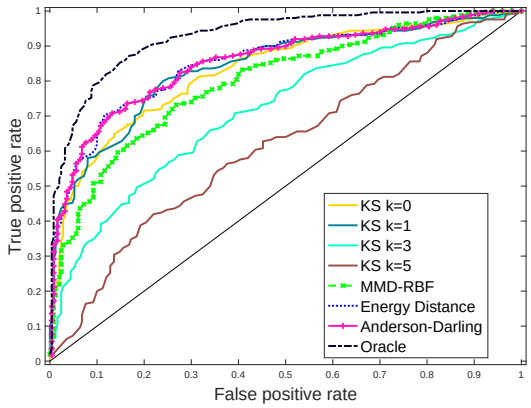

Figure 4.4: *ROC curves for* $P = N(0, 1)$, $Q = N(0.2, 1)$.

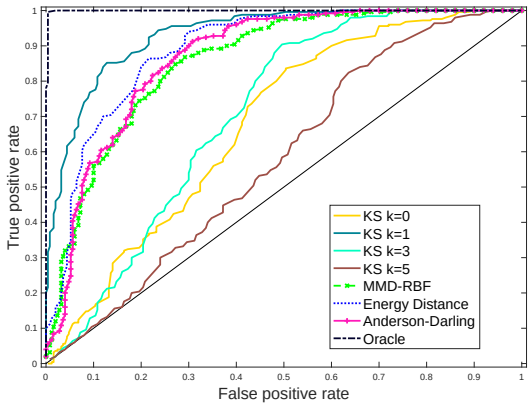

Figure 4.5: *ROC curves for* $P = N(0, 1)$, $Q = t(3)$.

Local Density Differences. In Figures 4.6 and 4.7, we examine the higher-order KS tests and the KS test, in cases where P, Q have densities p, q such that $p - q$ has sharp local changes. Figure 4.6 shows a case where $p - q$ is piecewise constant with a few short departures from 0 (see Appendix C for a plot) and $m = n = 500$. The KS test is very powerful, and the higher-order KS tests all perform poorly; in fact, the KS test here has

better power than all commonly-used nonparametric tests we tried (results not shown). Figure 4.7 displays a case where $p - q$ changes sharply in the right tail (see Appendix C for a plot) and $m = n = 2000$. The power of the higher-order KS test appears to increase with k, likely because the witness functions are able to better concentrate on sharp departures for large k.

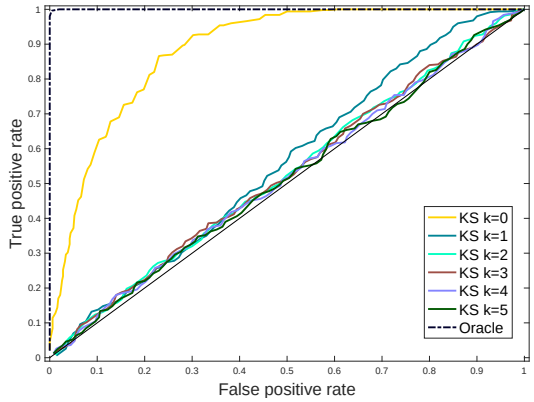

Figure 4.6: *ROC curves for piecewise constant $p - q$.*

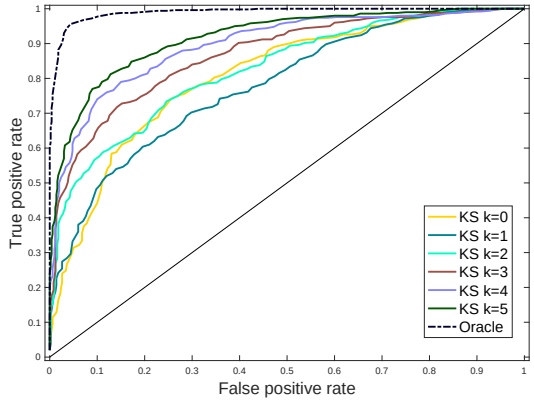

Figure 4.7: *ROC curves for tail departure in $p - q$.*

4.6 Discussion

This paper began by noting the variational characterization of the classical KS test as an IPM with respect to functions of bounded total variation, and then proposed a generalization to higher-order total variation classes. This generalization was nontrivial, with subtleties arising in defining the right class of functions so that the statistic was finite and amenable for simplification via a representer result, challenges in computing the statistic efficiently, and challenges in studying asymptotic convergence and concentration due to the fact that the function class is not uniformly sup norm bounded. The resulting class of linear-time higher-order KS tests was shown empirically to be more sensitive to tail differences than the usual KS test, and to have competitive power relative to several other popular tests.

In future work, we intend to more formally study the power properties of our new higher-order tests relative to the KS test. The following is a lead in that direction. For $k \geq 1$, define I^k to be the kth order integral operator, acting on a function f, via

$$(I^k f)(x) = \int_0^x \int_0^{t_k} \cdots \int_0^{t_2} f(t_1)\, dt_1 \cdots dt_k.$$

Denote by F_P, F_Q the CDFs of the distributions P, Q. Notice that the population-level KS test statistic can be written as $\rho(P, Q; \mathcal{F}_0) = \|F_P - F_Q\|_\infty$, where $\|\cdot\|_\infty$ is the sup norm. Interestingly, a similar representation holds for the higher-order KS tests.

Proposition 4.4. Assuming P, Q have k moments,

$$\rho(P, Q; \mathcal{F}_k) = \|(I^k)^*(F_P - F_Q)\|_\infty,$$

where $(I^k)^*$ is the adjoint of the bounded linear operator I^k, with respect to the usual L_2 inner product. Further, if P, Q are supported on $[0, \infty)$, or their first k moments match, then we have the more explicit representation

$$\rho(P, Q; \mathcal{F}_k) = \sup_{x \in \mathbb{R}} \left| \int_x^\infty \int_{t_k}^\infty \cdots \int_{t_2}^\infty (F_P - F_Q)(t_1)\, dt_1 \cdots dt_k \right|.$$

The representation in Proposition 4.4 could provide one avenue for power analysis. When P, Q are supported on $[0, \infty)$, or have k matching moments, the representation is particularly simple in form. This form confirms the intuition that detecting higher-order moment differences is hard: as k increases, the k-times integrated CDF difference $F_P - F_Q$ becomes smoother, and hence the differences are less accentuated.

In future work, we also intend to further examine the asymptotic null of the higher-order KS test (the Gaussian process from Theorem 4.3), and determine to what extent it depends on the underlying distribution P (beyond say, its first k moments). Lastly, some ideas in this paper seem extendable to the multivariate and graph settings, another direction for future work.

Chapter 5

Discussion and Conclusion

We studied extensions of trend filtering in regular lattice designs and additive models. We showed that our trend filtering estimators in these restricted multivariate settings achieve optimal rates in a minimax sense. We further showed that linear smoothers—this class includes many interesting and popular methods—cannot achieve these optimal rates on these higher order TV classes, thus extending classical results for 1d from Donoho & Johnstone (1998). The proof techniques used in additive trend filtering and lattices are different in nature. For upper bounds in additive models, we used results from empirical process theory and generalization bounds based on Rademacher complexity of the class of regression functions. In lattice design, we relied heavily on the spectrum of the Laplacian of the lattice. For lower bounds in additive models, we used Yang and Barron's method which in turn uses covering and packing numbers of the true function class and Fano's inequality. On the other hand, for lattice design the minimax rate lower bound was obtained by embedding ℓ_1 balls and Holder balls of appropriate sizes and using the known lower bounds on these inscribed balls.

The higher-order KS test statistic is an integral probability metric that can be computed quickly and is sensitive to some tail differences. We gave its asymptotic null distribution and showed that it concentrates to its population version. Coupled with the metric property, the test is asymptotically powerful.

Before concluding, we discuss a few interesting aspects of the problems that we worked on so far.

Weak sparsity and strong sparsity. Our estimation error bounds in lattice design and additive models are over a class of *weakly sparse* signals in the form

$$\mathcal{W}_1(C_n) = \{\theta : \|\Delta\theta\|_1 \leq C_n\}$$

where Δ is a penalty operator. It is interesting to study error bounds over *strongly sparse* signals of the form

$$\mathcal{W}_0(C_n) = \{\theta : \|\Delta\theta\|_0 \leq C_n\}$$

where $\|x\|_0$ is the number of nonzeros in a Euclidean vector x. Hutter & Rigollet (2016) give such bounds in lattice design setting and Tan & Zhang (2017) give such bounds for additive models with kth order total variation components. Chatterjee & Goswami (2019)

study a different notion of sparsity—they consider adaptivity of total variation denoising to signals with a few axis-parallel piecewise constant regions.

Difficulty with TV regularization in high dimensions. In univariate setting, locally adaptive regression splines Mammen & van de Geer (1997) defined by

$$\hat{f} = \underset{f}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \mathrm{TV}(f^{(k)}) \tag{5.1}$$

achieves minimax optimal rate over kth order TV classes in (1.3) for an appropriate choice of λ. However, it is non-trivial to extend this formulation to more than one dimensions. For continuous settings in two (or more) dimensions and $k = 0$, the optimizer in (5.1) satisfies $\hat{f}(x_i) = y_i$ for all $i \in [n]$ and $\mathrm{TV}(\hat{f}) = 0$. In other words, the estimator interpolates in a trivial way and its MSE will be σ^2. Discrete formulations such as our Kronecker trend filtering, graph trend filtering from Wang et al. (2016) and the Hardy-Krause variation based (discrete) formulation from Fang et al. (2019) do not suffer from this problem.

Difficulty of fitting splines in multivariate setting. It is nontrivial to fit splines in multiple dimensions with automatically selected knots (recall, this is the aim of KTF and GTF). In fact, fitting multivariate splines (say using multivariate adaptive regression splines (MARS) (Friedman 1991)) is similar in difficulty level to fitting neural networks with rectified linear units (ReLUs). Note that both the methods aim to find piecewise linear functions that minimize the prediction error. See Zhang & Goh (2016) where MARS and a three-layer feedforward neural network are compared in a civil engineering application. See also Balestriero & Baraniuk (2018) for a connection between neural networks and splines.

Non-lattice designs. Consider an arbitrary design where the d-dimensional input points do not lie on a lattice. One way of denoising in this setting is to build a k-Nearest Neighbor (kNN) graph or an ϵ-neighborhood graph of the input points based on some metric (say Euclidean distance between input points). Padilla et al. (2018) studies the denoising problem in this setting and shows that the graph TV denoising estimator (GTF with $k = 0$) achieves the minimax optimal rate on a class of true functions which are piecewise Lipschitz. One may apply a kth order GTF estimator on the kNN or ϵ-neighborhood graph and hope to recover smoother functions with better error bounds; but it is an open problem as of now. Note that there are nonparametric methods such as MARS, CART and neural networks which are applicable to non-lattice designs without modifications.

Isotropic and anisotropic TV denoising. The TV denoising method that we studied in Chapter 2 uses anisotropic TV penalty ($\mathrm{TV}_{\mathrm{aniso}}$), as opposed to the isotropic TV penalty ($\mathrm{TV}_{\mathrm{iso}}$) proposed by Rudin et al. (1992). Notice that $\mathrm{TV}_{\mathrm{iso}}(\theta) \leq \mathrm{TV}_{\mathrm{aniso}}(\theta)$ and $\mathrm{TV}_{\mathrm{aniso}}(\theta) \leq \sqrt{d}\mathrm{TV}_{\mathrm{aniso}}(\theta)$ for all $\theta \in \mathbb{R}^n$ on a d-dimensional grid. Therefore, the minimax optimal rates (and minimax linear rates) are same for the isotropic and anisotropic TV balls of same radius, up to factors of d.

Other notions of variation. KTF penalizes variation of only axis-parallel derivatives; it does not penalize variation of cross-derivatives. Hardy-Krause variation as defined in the recent work by Fang et al. (2019) considers cross-derivatives. However, we believe that a discrete roughness penalty that mimics

$$\sum_{\alpha \geq 0, \|\alpha\|_1 = k+1} \int \left| \frac{\partial^\alpha f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \right| dx$$

will result in an estimate with a desirable piecewise kth degree polynomial structure.

Conclusion

Linear smoothers are typically cheap to compute but they do not have good worst-case performance on TV and higher-order TV classes. Nonlinear smoothers such as trend filtering methods (KTF, GTF, additive trend filtering) are somewhat more expensive computationally but they have optimal worst-case performance on these signals. Further, the piecewise polynomial structure of estimates is interpretable and often desirable. KTF and GTF can be extended to exponential family losses and can be used for density estimation or classification. Also, the higher-order KS test seems to be sensitive to tail differences. In summary, we believe our theoretically sound methods with nice properties will be useful to practitioners.

Bibliography

Acar, R. & Vogel, C. R. (1994), 'Analysis of bounded variation penalty methods for ill-posed problems', *Inverse Problems* **10**, 1217–1229.

Anderson, T. W. & Darling, D. A. (1952), 'Asymptotic theory of certain goodness of fit criteria based on stochastic processes', *Annals of Mathematical Statistics* **23**(2), 193–212.

Anderson, T. W. & Darling, D. A. (1954), 'A test of goodness of fit', *Journal of the American Statistical Association* **49**(268), 765–769.

Ando, R. & Zhang, T. (2006), 'Learning on graph with Laplacian regularization', *Advances in Neural Information Processing Systems* **9**.

Aptekarev, A. I., Denisov, S. & Tulyakov, D. N. (2016), 'On a problem by Steklov', *Journal of American Mathematical Society* **29**(4), 1117–1165.

Arias-Castro, E., Pelletier, B. & Saligrama, V. (2018), 'Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension', *Journal of Nonparametric Statistics* **30**(2), 448–471.

Babenko, K. (1979), *Theoretical Foundations and Construction of Numerical Algorithms for the Problems of Mathematical Physics*. In Russian.

Balestriero, R. & Baraniuk, R. (2018), A spline theory of deep learning, *in* 'Proceedings of the 35th International Conference on Machine Learning', Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, Stockholmsmässan, Stockholm Sweden, pp. 374–383.

Barbero, A. & Sra, S. (2018), 'Modular proximal optimization for multidimensional total-variation regularization', *J. Mach. Learn. Res.* **19**(1), 2232–2313.

Baringhaus, L. & Franz, C. (2004), 'On a new multivariate two-sample test', *Journal of Multivariate Analysis* **88**(1), 190–206.

Bartlett, P., Bousquet, O. & Mendelson, S. (2005), 'Local Rademacher complexities', *Annals of Statistics* **33**(4), 1497–1537.

Belkin, M. & Niyogi, P. (2002), 'Using manifold structure for partially labelled classification', *Advances in Neural Information Processing Systems* **15**.

Belkin, M. & Niyogi, P. (2003), 'Laplacian eigenmaps for dimensionality reduction and data representation', *Neural Computation* **15**(6), 1373–1396.

Belkin, M. & Niyogi, P. (2004), 'Semi-supervised learning on Riemannian manifolds', *Machine Learning* **56**(1–3), 209–239.

Belkin, M. & Niyogi, P. (2005), 'Towards a theoretical foundation for Laplacian-based manifold methods', *Conference on Learning Theory* **18**.

Belkin, M., Niyogi, P. & Sindhwani, V. (2005), 'On manifold regularization', *International Conference on Artificial Intelligence and Statistics* **8**.

Bhattacharya, B. B. (2016), Power of graph-based two-sample tests, PhD thesis, Stanford University.

Birge, L. & Massart, P. (2001), 'Gaussian model selection', *Journal of the European Mathematical Society* **3**(3), 203–268.

Birman, M. & Solomyak, M. (1967), 'Piecewise-polynomial approximations of functions of the classes W_p^α', *Mathematics of the USSR-Sbornik* **73**(115), 331–335. In Russian.

Bogoya, J. M., Bottcher, A., Grudsky, S. M. & Maximenko, E. A. (2016), 'Eigenvectors of Hermitian Toeplitz matrices with smooth simple-loop symbols', *Linear Algebra and its Applications* **493**, 606–637.

Bottcher, A. & Grudsky, S. M. (2005), *Spectral Properties of Banded Toeplitz Matrices*, Society for Industrial and Applied Mathematics.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternative direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–122.

Bredies, K., Kunisch, K. & Pock, T. (2010), 'Total generalized variation', *SIAM Journal on Imaging Sciences* **3**(3), 492–526.

Breiman, L. & Friedman, J. (1985), 'Estimating optimal transformations for multiple regression and correlation', *Journal of the American Statistical Association* **80**(391), 614–619.

Bryson, M. C. (1974), 'Heavy-tailed distributions: Properties and tests', *Technometrics* **16**(1), 61–68.

Buja, A., Hastie, T. & Tibshirani, R. (1989), 'Linear smoothers and additive models', *Annals of Statistics* **17**(2), 453–510.

Calitz, F. (1987), 'An alternative to the Kolmogorov-Smirnov test for goodness of fit', *Communications in Statistics: Theory and Methods* **16**(12), 3519–3534.

Chambolle, A. & Darbon, J. (2009), 'On total variation minimization and surface evolution using parametric maximum flows', *International Journal of Computer Vision* **84**, 288–307.

Chambolle, A. & Lions, P.-L. (1997), 'Image recovery via total variation minimization and related problems', *Numerische Mathematik* **76**(2), 167–188.

Chambolle, A. & Pock, T. (2011), 'A first-order primal-dual algorithm for convex problems with applications to imaging', *Journal of Mathematical Imaging and Vision* **40**, 120–145.

Chatterjee, S. & Goswami, S. (2019), 'New risk bounds for 2d total variation denoising'.

Condat, L. (2012), A direct algorithm for 1d total variation denoising. HAL: 00675043.

Conte, S. & de Boor, C. (1980), *Elementary Numerical Analysis: An Algorithmic Approach*, McGraw-Hill, New York. International Series in Pure and Applied Mathematics.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer.

Dobson, D. & Santosa, F. (1996), 'Recovery of blocky images from noisy and blurred data', *SIAM Journal on Applied Mathematics* **56**(4), 1181–1198.

Donoho, D. & Johnstone, I. (1994*a*), 'Minimax risk over ℓ_p-balls for ℓ_q-error', *Probability Theory and Related Fields* **99**(2), 277–303.

Donoho, D. L. & Johnstone, I. M. (1994*b*), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**(3), 425–455.

Donoho, D. L. & Johnstone, I. M. (1998), 'Minimax estimation via wavelet shrinkage', *Annals of Statistics* **26**(8), 879–921.

Donoho, D., Liu, R. & MacGibbon, B. (1990), 'Minimax risk over hyperrectangles, and implications', *Annals of Statistics* **18**(3), 1416–1437.

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.

Dudley, R. M. (1999), *Uniform Central Limit Theorems*, Cambridge University Press.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association* **81**(394), 461–470.

Fahrmeir, L. & Lang, S. (2001), 'Bayesian inference for generalized additive mixed models based on Markov random field priors', *Journal of the Royal Statistical Society: Series C* **50**(2), 201–220.

Fang, B., Guntuboyina, A. & Sen, B. (2019), 'Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy-krause variation'.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines', *Annals of Statistics* **19**(1), 1–67.

Friedman, J. H. & Rafsky, L. C. (1979), 'Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests', *Annals of Statistics* **7**(4), 697–717.

Friedman, J. & Stuetzle, W. (1981), 'Projection pursuit regression', *Journal of the American Statistical Association* **76**(376), 817–823.

Godunov, S. & Ryabenkii, V. (1987), *Difference Schemes: An Introduction to the Underlying Theory*, Elsevier, Amsterdam. Number 19 in Studies in Mathematics and Its Applications.

Green, P. & Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall/CRC Press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schelkopf, B. & Smola, A. (2012), 'A kernel two-sample test', *Journal of Machine Learning Research* **13**, 723–773.

Gu, C. & Wahba, G. (1991), 'Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method', *SIAM Journal of Scientific and Statistical Computing* **12**(2), 383–398.

Gyorfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression*, Springer.

Harchaoui, Z. & Levy-Leduc, C. (2010), 'Multiple change-point estimation with a total variation penalty', *Journal of the American Statistical Association* **105**(492), 1480–1493.

Haris, A., Simon, N. & Shojaie, A. (2018), Generalized sparse additive models.

Hastie, T. (1983), Non-parametric logistic regression. Technical Report, Stanford University.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman and Hall.

Henze, N. (1988), 'A multivariate two-sample test based on the number of nearest neighbor type coincidences', *Annals of Statistics* **16**(2), 772–783.

Hoefling, H. (2010), 'A path algorithm for the fused lasso signal approximator', *Journal of Computational and Graphical Statistics* **19**(4), 984–1006.

Hsu, D., Kakde, S. & Zhang, T. (2012), 'A tail inequality for quadratic forms of subgaussian random vectors', *Electronic Communications in Probability* **17**(52), 1–6.

Hutter, J.-C. & Rigollet, P. (2016), 'Optimal rates for total variation denoising', *Annual Conference on Learning Theory* **29**, 1115–1146.

Janssen, A. (2000), 'Global power functions of goodness of fit tests', *Annals of Statistics* **28**(1), 239–253.

Johnson, N. (2013), 'A dynamic programming algorithm for the fused lasso and l_0-segmentation', *Journal of Computational and Graphical Statistics* **22**(2), 246–260.

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

Kim, Y.-J. & Gu, C. (2004), 'Smoothing spline Gaussian regression: more scalable computation via efficient approximation', *Journal of the Royal Statistical Society: Series B* **66**(2), 337–356.

Kolmogorov, A. (1933), 'Sulla determinazione empirica di una legge di distribuzione', *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91.

Kolmogorov, A. N. & Tikhomirov, V. M. (1959), 'ϵ-entropy and ϵ-capacity of sets in function spaces', *Uspekhi Matematicheskikh Nauk* **14**(2), 3–86.

Korostelev, A. P. & Tsybakov, A. B. (2003), *Minimax Theory of Image Reconstructions*, Springer.

Kovac, A. & Smith, A. (2011), 'Nonparametric regression on a graph', *Journal of Computational and Graphical Statistics* **20**(2), 432–447.

Kunsch, H. (1994), 'Robust priors for smoothing and image restoration', *Annals of the Institute of Statistical Mathematics* **46**(1), 1–19.

Lin, Y. & Zhang, H. H. (2006), 'Component selection and smoothing in multivariate nonparametric regression', *Annals of Statistics* **34**(5), 2272–2297.

Liu, R. Y. & Singh, K. (1993), 'A quality index based on data depth and multivariate rank tests', *Journal of the American Statistical Association* **88**(421), 252–260.

Lou, Y., Bien, J., Caruana, R. & Gehrke, J. (2016), 'Sparse partially linear additive models', *Journal of Computational and Graphical Statistics* **25**(4), 1126–1140.

Mammen, E. (1991), 'Nonparametric regression under qualitative smoothness assumptions', *Annals of Statistics* **19**(2), 741–759.

Mammen, E. & van de Geer, S. (1997), 'Locally apadtive regression splines', *Annals of Statistics* **25**(1), 387–413.

Mason, D. M. & Schuenemeyer, J. H. (1983), 'A modified Kolmogorov-Smirnov test sensitive to tail alternatives', *Annals of Statistics* **11**(3), 933–946.

Meier, L., van de Geer, S. & Buhlmann, P. (2009), 'High-dimensional additive modeling', *Annals of Statistics* **37**(6), 3779–3821.

Muller, A. (1997), 'Integral probability metrics and their generating classes of functions', *Advances in Applied Probability* **29**(2), 429–443.

Nesterov, Y. (2000), *Squared Functional Systems and Optimization Problems*, Springer, pp. 405–440.

Neuman, C. P. & Schonbach, D. I. (1974), 'Discrete (Legendre) orthogonal polynomials—a survey', *International Journal for Numerical Methods in Engineering* **8**(4), 743–770.

Ng, M., Chan, R. & Tang, W.-C. (1999), 'A fast algorithm for deblurring models with Neumann boundary conditions', *SIAM Journal on Scientific Computing* **21**(3), 851–866.

Okamoto, M. (1973), 'Distinctness of the eigenvalues of a quadratic form in a multivariate sample', *Annals of Statistics* **1**(4), 763–765.

Padilla, O. H. M., Sharpnack, J., Chen, Y. & Witten, D. M. (2018), 'Adaptive nonparametric regression with the k-nn fused lasso'.

Padilla, O. H. M., Sharpnack, J., Scott, J., & Tibshirani, R. J. (2016), The DFS fused lasso: Linear-time denoising over general graphs. arXiv: 1608.03384.

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

Poschl, C. & Scherzer, O. (2008), Characterization of minimizers of convex regularization functionals, *in* 'Frames and Operator Theory in Analysis and Signal Processing', Vol. 451, AMS eBook Collections, pp. 219–248.

Ramdas, A., Reddi, S., Pczos, B., Singh, A. & Wasserman, L. (2015), 'On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions.', *Twenty-Ninth Conference on Artificial Intelligence* pp. 3571–3577.

Ramdas, A., Reddi, S., Poczos, B., Singh, A. & Wasserman, L. (2015), 'Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing', *arXiv preprint arXiv:1508.00655* .

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

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

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

Rinaldo, A. (2009), 'Properties and refinements of the fused lasso', *Annals of Statistics* **37**(5), 2922–2952.

Rosen, M. I. (1995), 'Niels Hendrik Abel and equations of the fifth degree', *The American Mathematical Monthly* **102**(6), 495–505.

Rosenbaum, P. R. (2005), 'An exact distribution-free test comparing two multivariate distributions based on adjacency', *Journal of the Royal Statistical Society: Series B* **67**(4), 515–530.

Rudin, L. I., Osher, S. & Faterni, E. (1992), 'Nonlinear total variation based noise removal algorithms', *Physica D: Nonlinear Phenomena* **60**(1), 259–268.

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations', *Journal of the Royal Statistical Society: Series B* **71**(2), 319–392.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Sadhanala, V. & Tibshirani, R. J. (2017), Additive models with trend filtering. arXiv: 1702.05037.

Sadhanala, V., Wang, Y.-X., Ramdas, A. & Tibshirani, R. J. (2019), A higher-order kolmogorov-smirnov test, *in* 'Proceedings of Machine Learning Research', Vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 2621–2630.

Sadhanala, V., Wang, Y.-X., Sharpnack, J. & Tibshirani, R. J. (2017), 'Higher-total variation classes on grids: Minimax theory and trend filtering methods', *Advances in Neural Information Processing Systems* **30**.

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

Sardy, S. & Tseng, P. (2004), 'AMlet, RAMlet and GAMlet: Automatic nonlinear fitting of additive model, robust and generalized with wavelets', *Journal of Computational and Graphical Statistics* **13**(2), 283–309.

Schilling, M. F. (1986), 'Multivariate two-sample tests based on nearest neighbors', *Journal of the American Statistical Association* **81**(395), 799–806.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. (2013), 'Equivalence of distance-based and RKHS-based statistics in hypothesis testing', *Annals of Statistics* **41**(5), 2263–2291.

Sharpnack, J., Rinaldo, A. & Singh, A. (2012), 'Sparsistency via the edge lasso', *International Conference on Artificial Intelligence and Statistics* **15**.

Sharpnack, J. & Singh, A. (2010), 'Identifying graph-structured activation patterns in networks', *Advances in Neural Information Processing Systems* **23**.

Shor, N. Z. (1998), *Nondifferentiable Optimization and Polynomial Problems*, Nonconvex Optimization and Its Applications, Springer.

Smirnov, N. (1948), 'Table for estimating the goodness of fit of empirical distributions', *Annals of Mathematical Statistics* **19**(2), 279–281.

Smola, A. & Kondor, R. (2003), 'Kernels and regularization on graphs', *Proceedings of the Annual Conference on Learning Theory* **16**.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Scholkopf, B. & Lanckriet, G. R. G. (2009), 'On integral probability metrics, ϕ-divergences and binary classification', *arXiv preprint arXiv:0901.2698* .

Steidl, G., Didas, S. & Neumann, J. (2006), 'Splines in higher order TV regularization', *International Journal of Computer Vision* **70**(3), 214–255.

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

Strikwerda, J. C. (2004), *Finite difference schemes and partial differential equations*, Society for Industrial and Applied Mathematics.

Szekely, G. J. & Rizzo, M. L. (2004), 'Testing for equal distributions in high dimension', *InterStat* **5**(16.10), 1249–1272.

Tan, Z. & Zhang, C.-H. (2017), Penalized estimation in additive regression with high-dimensional data. arXiv: 1704.07229.

Tansey, W. & Scott, J. (2015), A fast and flexible algorithm for the graph-fused lasso. arXiv: 1505.06475.

Tibshirani, R. (1983), Non-parametric estimation of relative risk. Technical Report, Stanford University.

Tibshirani, R. J. (2013), 'The lasso problem and uniqueness', *Electronic Journal of Statistics* **7**, 1456–1490.

Tibshirani, R. J. (2014), 'Adaptive piecewise polynomial estimation via trend filtering', *Annals of Statistics* **42**(1), 285–323.

Tibshirani, R. J. (2015), 'Degrees of freedom and model search', *Statistica Sinica* **25**(3), 1265–1296.

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. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Annals of Statistics* **40**(2), 1198–1232.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society: Series B* **67**(1), 91–108.

Tropp, J. A. (2012), 'User-friendly tail bounds for sums of random matrices', *Foundations of Computational Mathematics* **12**(4), 389–434.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for nondifferentiable minimization', *Journal of Optimization Theory and Applications* **109**(3), 475–494.

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, Springer.

van de Geer, S. (1990), 'Estimating a regression function', *Annals of Statistics* **18**(2), 907–924.

van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambdrige University Press.

van de Geer, S. (2014), 'On the uniform convergence of empirical norms and inner products, with application to causal inference', *Electronic Journal of Statistics* **8**, 543–574.

van der Burg, E. & de Leeuw, J. (1983), 'Non-linear canonical correlation', *British Journal of Mathematical and Statistical Psychological* **36**(1), 54–80.

van der Vaart, A. & Wellner, J. (1996), *Weak Convergence*, Springer.

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wang, Y.-X., Sharpnack, J., Smola, A. & Tibshirani, R. J. (2016), 'Trend filtering on graphs', *Journal of Machine Learning Research* **17**(105), 1–41.

Wang, Y.-X., Smola, A. & Tibshirani, R. J. (2014), 'The falling factorial basis and its statistical applications', *International Conference on Machine Learning* **31**.

Wang, Y., Yang, J., Yin, W. & Zhang, Y. (2008), 'A new alternating minimization algorithm for total variation image reconstruction', *SIAM Journal on Imaging Sciences* **1**(3), 248–272.

Wood, S. N. (2000), 'Modelling and smoothing parameter estimation with multiple quadratic penalties', *Journal of the Royal Statistical Society: Series B* **62**(2), 413–428.

Wood, S. N. (2004), 'Stable and efficient multiple smoothing parameter estimation for generalized additive models', *Journal of the American Statistical Association* **99**(467), 673–686.

Wood, S. N. (2011), 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society: Series B* **73**(1), 3–36.

Wood, S. N. (2017), *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC Press.

Wood, S. N., Goude, Y. & Shaw, S. (2015), 'Generalized additive models for large data sets', *Journal of the Royal Statistical Society: Series C* **64**(1), 139–155.

Wood, S. N., Pya, N. & Safken, B. (2016), 'Smoothing parameter and model selection for general smooth models', *Journal of the American Statistical Association* **111**(516), 1548–1575.

Yang, Y. & Barron, A. (1999), 'Information-theoretic determination of minimax rates of convergence', *Annals of Statistics* **27**(5), 1564–1599.

Zhang, S. & Wong, M.-Y. (2003), 'Wavelet threshold estimation for additive regression models', *Annals of Statistics* **31**(1), 152–173.

Zhang, W. & Goh, A. T. (2016), 'Multivariate adaptive regression splines and neural network models for prediction of pile drivability', *Geoscience Frontiers* **7**(1), 45 – 52. Special Issue: Progress of Machine Learning in Geosciences.

Zhou, D., Huang, J. & Scholkopf, B. (2005), 'Learning from labeled and unlabeled data on a directed graph', *Proceedings of the International Conference on Machine Learning* **22**.

Zhu, X., Ghahramani, Z. & Lafferty, J. (2003), 'Semi-supervised learning using Gaussian fields and harmonic functions', *Proceedings of the International Conference on Machine Learning* **20**.

Appendix A

Appendix for Trend Filtering on Grids

A.1 Certain properties of GTF/KTF operators

A.1.1 Proof of Lemma 2.1

The nullity of $\widetilde{\Delta}^{(k+1)}$ is the number of nonzero singular values of $\widetilde{\Delta}^{(k+1)}$, or equivalently, the number of nonzero eigenvalues of $(\widetilde{\Delta}^{(k+1)})^T \widetilde{\Delta}^{(k+1)}$. Following from (2.11), and abbreviating $D = D^{(k+1)}_{1\mathrm{d}}$,

$$(\widetilde{\Delta}^{(k+1)})^T \widetilde{\Delta}^{(k+1)} = D^T D \otimes I \otimes \cdots \otimes I \; + \; I \otimes D^T D \otimes \cdots \otimes I \; + \; \ldots \\ + \; I \otimes I \otimes \cdots \otimes D^T D,$$

the Kronecker sum of $D^T D$ with itself, a total of d times. Using a standard fact about Kronecker sums, if ρ_i, $i = 1, \ldots, N$ denote the eigenvalues of $D^T D$ then

$$\rho_{i_1} + \rho_{i_2} + \cdots + \rho_{i_d}, \quad i_1, \ldots, i_d \in \{1, \ldots, N\}$$

are the eigenvalues of $(\widetilde{\Delta}^{(k+1)})^T \widetilde{\Delta}^{(k+1)}$. By counting the multiplicity of the zero eigenvalue, we arrive at a nullity for $\widetilde{\Delta}^{(k+1)}$ of $(k+1)^d$. It is straightforward to check that the vectors specified in the lemma, given by evaluations of polynomials, are in the null space, and that these are linearly independent, which completes the proof. □

A.1.2 Proof of Lemma 2.2

Let us define

$$\widetilde{D} = \begin{bmatrix} C^{(k+1)}_{1\mathrm{d}} \\ D^{(k+1)}_{1\mathrm{d}} \end{bmatrix} \in \mathbb{R}^{N \times N},$$

where the first $k+1$ rows are given by a matrix $C^{(k+1)} \in \mathbb{R}^{(k+1) \times N}$ that completes the row space, as in Lemma 2 of Wang et al. (2014). And now, again by Lemma 2 of Wang et al. (2014),

$$\big(H^{(k)}_{1\mathrm{d}}\big)^{-1} = \frac{1}{k!} \widetilde{D}, \tag{A.1}$$

where $H_{1\mathrm{d}}^{(k)} \in \mathbb{R}^{N \times N}$ is the falling factorial basis matrix of order k, which has elements

$$\big(H_{1\mathrm{d}}^{(k)}\big)_{ij} = h_j(i/N), \quad i, j = 1, \ldots, N,$$

with h_i, $i = 1, \ldots, N$ denoting the falling factorial basis functions in (1.7).

Let us write the KTF problem in (2.7), (2.11) explicitly as

$$\min_{\theta \in \mathbb{R}^n} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \left\| \begin{bmatrix} D_{1\mathrm{d}}^{(k+1)} \otimes I \otimes \cdots \otimes I \\ I \otimes D_{1\mathrm{d}}^{(k+1)} \otimes \cdots \otimes I \\ \vdots \\ I \otimes I \otimes \cdots \otimes D_{1\mathrm{d}}^{(k+1)} \end{bmatrix} \theta \right\|_1. \tag{A.2}$$

We now transform variables in this problem by defining $\theta = (H_{1\mathrm{d}}^{(k)} \otimes \cdots \otimes H_{1\mathrm{d}}^{(k)})\alpha$ and using (A.1), which turns (A.2) into an equivalent basis form,

$$\min_{\alpha \in \mathbb{R}^n} \ \frac{1}{2}\Big\|y - \Big(H_{1\mathrm{d}}^{(k)} \otimes \cdots \otimes H_{1\mathrm{d}}^{(k)}\Big)\alpha\Big\|_2^2 + \lambda k! \left\| \begin{bmatrix} I^0 \otimes H_{1\mathrm{d}}^{(k)} \otimes \cdots \otimes H_{1\mathrm{d}}^{(k)} \\ H_{1\mathrm{d}}^{(k)} \otimes I^0 \otimes \cdots \otimes H_{1\mathrm{d}}^{(k)} \\ \vdots \\ H_{1\mathrm{d}}^{(k)} \otimes H_{1\mathrm{d}}^{(k)} \otimes \cdots \otimes I^0 \end{bmatrix} \alpha \right\|, \tag{A.3}$$

where $I^0 = [\, 0_{(N-k-1)\times(k+1)} \ I_{(N-k-1)} \,]$.

Interestingly, the penalty in (A.3) is not a pure sparsity penalty on the coefficients α (as it is in basis form in 1d) but a sparsity penalty on aggregated (sums of) coefficients. This makes the penalty a little hard to interpret, but to glean intuition, we can rewrite the problem once more via the transformation

$$f = \sum_{i_1, \ldots, i_d = 1}^{N} \alpha_{i_1, \ldots, i_d} (h_{i_1} \otimes h_{i_2} \otimes \cdots \otimes h_{i_d}), \tag{A.4}$$

where recall we are indexing the components of α by $\alpha_{i_1, \ldots, i_d}$, for $i_1, \ldots, i_d = 1, \ldots, N$ (and the summands above use tensor products of univariate functions). To be concrete, note that the function f defined in (A.4) evaluates to

$$f(x) = \sum_{i_1, \ldots, i_d = 1}^{N} \alpha_{i_1, \ldots, i_d} h_{i_1}(x) h_{i_2}(x_2) \cdots h_{i_d}(x_d), \quad x \in [0,1]^d.$$

Thus we can equivalently write the basis form in (A.3) in functional form

$$\min_{f \in H_d} \ \frac{1}{2} \sum_{x \in Z_d} \big(y(x) - f(x)\big)^2 + \lambda \sum_{j=1}^{d} \sum_{x_{-j} \in Z_{d-1}} \mathrm{TV}\bigg(\frac{\partial^k f(\cdot, x_{-j})}{\partial x_j^k}\bigg), \tag{A.5}$$

where recall $f(\cdot, x_{-j})$ denotes f as function of the jth dimension with all other dimensions fixed at x_{-j}, $\partial^k / \partial x_j^k(\cdot)$ denotes the kth partial weak derivative operator with respect to

x_j, for $j = 1, \dots, d$, and $\mathrm{TV}(\cdot)$ denotes the total variation operator. To see the equivalence between the penalty terms in (A.3) and (A.5), it can be directly checked that

$$k! \Big(I^0 \otimes H_{1\mathrm{d}}^{(k)} \otimes \cdots \otimes H_{1\mathrm{d}}^{(k)} \Big) \alpha$$

contains the differences of the function $\partial^k f / \partial x_1^k$ over all pairs of grid positions that are adjacent in the x_1 direction, where f is as in (A.4). This, combined with the fact that $\partial^k f / \partial x_1^k$ is constant in between lattice positions, means that

$$k! \Big\| \Big(I^0 \otimes H_{1\mathrm{d}}^{(k)} \otimes \cdots \otimes H_{1\mathrm{d}}^{(k)} \Big) \alpha \Big\|_1 = \sum_{x_{-1} \in Z_{d-1}} \mathrm{TV} \bigg(\frac{\partial^k f(\cdot, x_{-1})}{\partial x_1^k} \bigg),$$

the total variation of $\partial^k f / \partial x_1^k$ added up over all slices of the lattice Z_d in the x_1 direction. Similar arguments apply to the penalty terms corresponding to dimensions $j = 2, \dots, d$, and this completes the proof. □

A.2 Canonical scaling

A.2.1 Proof of Lemma A.5

Suppose that $\theta \in \mathcal{H}_d(1)$ that is a discretization of a 1-Lipschitz function f, i.e., $\theta_i = f(i_1/N \dots, i_d/N)$, $i = 1, \dots, n$. We first we compute and bound its squared Sobolev norm

$$\begin{aligned} \|D\theta\|_2^2 = \sum_{(i,j) \in E} (\theta_i - \theta_j)^2 &= \sum_{(i,j) \in E} \big(f(i_1/N, \dots, i_d/N) - f(j_1/N, \dots, j_d/N) \big)^2 \\ &\leq \sum_{(i,j) \in E} \big\| (i_1/N, \dots, i_d/N) - (j_1/N, \dots, j_d/N) \big\|_\infty^2 \\ &= m/N^2, \end{aligned}$$

where, recall, we denote by $m = |E|$ the number of edges in the grid. In the second line we used the 1-Lipschitz property of f, and in the third we used that multi-indices corresponding to adjacent locations on the grid are exactly 1 apart, in ℓ_∞ distance. Thus we see that setting $C_n' = \sqrt{m}/N$ gives the desired containment $\mathcal{S}_d(C_n') \supseteq \mathcal{H}_d(1)$. It is always true that $m \asymp n$ for a d-dimensional grid (though the constant may depend on d), so that $\sqrt{m}/N \asymp n^{1/2-1/d}$. This completes the proof for the Sobolev class scaling. □

A.3 Proofs of upper bounds for GTF/KTF

A.3.1 Proof of Theorem 2.2

For $d = 2$, it is shown in the proof of Corollary 8 in Wang et al. (2016) that the GTF operator $\Delta^{(k+1)}$ satisfies the incoherence property, as defined in Theorem 2.1, for any choice of cutoff $i_0 \geq 1$, and with a constant $\mu = 4$ when k is even and $\mu = 2$ when k is odd. Here we establish the incoherence property for $d \geq 2$. We treat the cases where k is odd and even separately.

If k is odd we can extend the argument from Corollary 8 in Wang et al. (2016) in a straightforward manner. The GTF operator is $\Delta^{(k+1)} = L^{(k+1)/2}$ where L is the Laplacian of the d-dimensional grid graph. Denoting the Laplacian of the chain graph of length N by $L_{1\mathrm{d}}$, L is given by

$$L = L_{1\mathrm{d}} \otimes I \otimes I + I \otimes L_{1\mathrm{d}} \otimes I + I \otimes I \otimes L_{1\mathrm{d}}$$

for $d = 3$ and

$$L = L_{1\mathrm{d}} \otimes I \cdots \otimes I + I \otimes L_{1\mathrm{d}} \cdots \otimes I + \cdots + I \otimes \ldots I \otimes L_{1\mathrm{d}}$$

for general d where each term in the summation is a Kronecker product of d matrices. Let $\lambda_i, u_i, i \in [N]$ are the eigenvalues and eigenvectors of $L_{1\mathrm{d}}$. As shown in Wang et al. (2016), in 1d, we have the incoherence property $\|u_i\|_\infty \leq \sqrt{2/N}$ for all $i \in [N]$. The eigenvalues of L are $\sum_{j=1}^d \lambda_{i_j}$ and the corresponding eigenvectors are $u_{i_1} \otimes \cdots \otimes u_{i_d}$ for $i_1, \ldots, i_d \in [N]$. Clearly, incoherence holds for the eigenvectors of L with constant $\mu = 2^{d/2}$.

If k is even, then the left the singular vectors of $\Delta^{(k+1)}$ are the same as those of $\Delta^{(1)}$. We know that both the left and right singular vectors of $D^{(1)}_{1\mathrm{d}}$ satisfy the incoherence property with constant $\mu = \sqrt{2}$ (see Corollary 7 in Wang et al. (2016)). Setting $D = D^{(1)}_{1\mathrm{d}}$ in Lemma A.1, we see that the left singular vectors of $\Delta^{(1)}$ and hence those of $\Delta^{(k+1)}$ satisfy incoherence property with constant $2^{d/2}$. Therefore, for all integers $k \geq 0$, the left singular vectors of $\Delta^{(k+1)}$ are incoherent with constant $2^{d/2}$.

From the incoherence property and Theorem 2.1, the GTF estimator $\hat{\theta}$ in (2.7), (2.8) satisfies

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{i_0}{n} + \frac{\mu}{n}\sqrt{\frac{\log r}{n}\sum_{i_0+1}^{q}\frac{1}{\xi_i^2}} \cdot \|\Delta\theta_0\|_1\right), \tag{A.6}$$

where $\xi_i, i \in [n-1]$ are the nonzero singular values of $\Delta^{(k+1)}$, $q = n-1$ and $\mu = 2^{d/2}$. It suffices to upper bound the partial sum term $\sum_{i=i_0+1}^{n-1} \xi_i^{-2}$.

Set $\beta = i_0^{1/d}$ and consider $\sum_{\|i\| \geq \beta} \frac{1}{\xi_i^2}$. The number of $i \in [N]^d$ satisfying $\|i\| \leq \beta$ is $\Theta(\beta^d) = \Theta(i_0)$. Lemma A.10 gives the key calculation where it is shown that for large enough n and each $i_0 \geq 1$,

$$\sum_{\|i\| \geq \beta} \frac{1}{\xi_i^2} = \sum_{\|i\| \geq \beta} \frac{1}{\rho_i^{k+1}} \leq c \begin{cases} n & 2(k+1) < d \\ n \log(n/i_0) & 2(k+1) = d \\ n(n/i_0)^{(2k+2-d)/d} & 2(k+1) > d \end{cases}$$

where $\rho_i, i \in [n-1]$ are nonzero eigenvalues of the Laplacian L and $c > 0$ is a constant that depends only on k . For $k < d/2 - 1$, to minimize to the upper bound given in (A.6), we want to balance

$$\frac{i_0}{n} \quad \text{with} \quad \frac{\mu C_n}{n}\sqrt{\log n}.$$

This leads us to choose $i_0 \asymp C_n\sqrt{\log n}$, and plugging this in gives the result for $k < d/2 - 1$.

If $k = d/2 - 1$ is an integer, then we want to balance

$$\frac{i_0}{n} \quad \text{with} \quad \frac{\mu C_n}{n}\sqrt{\log n \log(n/i_0)}.$$

This leads us to choose $i_0 \asymp C_n \log n$, and plugging this in gives the result for $k = d/2 - 1$.

For $k > d/2 - 1$, we want to balance

$$\frac{i_0}{n} \quad \text{with} \quad \frac{\mu C_n}{n\sqrt{n}} \sqrt{n(n/i_0)^{(2k+2-d)/d} \log n}$$

This leads us to take

$$i_0 \asymp (C_n \sqrt{\log r})^{\frac{2d}{2k+2+d}} n^{\frac{2k+2-d}{2k+2+d}}$$

and plugging this in completes the proof for $k > d/2 - 1$. □

A.3.2 Proof of Theorem 2.3

The KTF operator (2.11), is the Δ in (A.7) with $D = D_{1d}^{(k+1)}$. Abbreviate $N' = N - k - 1$. Let β_i, u_i, v_i be a triplet of nonzero singular value, left singular vector, and right singular vector of $D_{1d}^{(k+1)}$, for $i \in [N']$ and let p_j, $j \in [k+1]$ form an orthogonal basis for the null space of $D_{1d}^{(k+1)}$. From Lemma A.1 it suffices to show incoherence of u_i, v_i, $i \in [N']$, and p_i, $i \in [k+1]$. Incoherence of u_i, $i \in [N']$ is established in Lemma A.12 and of v_i, $i \in [N']$ in Lemma A.13, using specialized approximations for eigenvectors of Toeplitz matrices from Bogoya et al. (2016). Incoherence of p_i, $i \in [k+1]$ may be seen by choosing, e.g., these vectors to be the discrete Legendre orthogonal polynomials as in Neuman & Schonbach (1974). We have thus shown that $\widetilde{\Delta}^{(k+1)}$ satisfies the incoherence property, as defined in Theorem 2.1, for any choice of $i_0 \geq 1$.

Now we address the partial sum term $\sum_{i=i_0+1}^{n-1} \xi_i^{-2}$. Lemma A.11 shows that for large enough n and a constant $c > 0$ depending only on k,

$$\sum_{i=i_0+1}^{n-(k+1)^d} \frac{1}{\xi_i^2} \leq c \begin{cases} n & 2(k+1) < d \\ n \log(n/i_0) & 2(k+1) = d \\ n(n/i_0)^{(2k+2-d)/d} & 2(k+1) > d \end{cases}$$

just as was the case for GTF. (In fact, this result is proved by tying the singular values of the KTF operator to those of the GTF operator.) Repeating the same arguments as in the proof of Theorem 2.2 gives the desired result. □

A.4 Incoherence of GTF/KTF penalty operators for d-dimensional grids

Let

$$\Delta = \begin{bmatrix} D \otimes I \otimes \cdots \otimes I \\ I \otimes D \otimes \cdots \otimes I \\ \vdots \\ I \otimes I \otimes \cdots \otimes D \end{bmatrix} \tag{A.7}$$

where each Kronecker product has d terms. With $D = D_{1d}^{(k+1)}$ where $D_{1d}^{(k+1)} \in \mathbb{R}^{N-k-1 \times N}$, we get the KTF penalty operator $\Delta = \widetilde{\Delta}^{(k+1)}$. With $D = D_{1d}^{(1)}$ where $D_{1d}^{(1)} \in \mathbb{R}^{N-1 \times N}$, we get the GTF penalty operator $\Delta = \Delta^{(1)}$, of order $k = 0$.

Lemma A.1. Let Δ be as defined in (A.7) for a matrix $D \in \mathbb{R}^{N' \times N}$ with $N' \leq N$. Let $\gamma_i, u_i, v_i, i \in [N]$ denote the singular values of D^T, its right and left singular vectors. Note that $\gamma_i = 0, u_i = 0, v_i \in \text{null}(D)$ for $i \in [p]$ where $p = \text{nullity}(D)$. If these singular vectors are incoherent, that is $\|v_i\|_\infty \leq \mu/\sqrt{N}, \|u_i\|_\infty \leq \mu/\sqrt{N'}$ for a constant $\mu \geq 1$, then the left singular vectors ν of Δ are incoherent with a constant μ^d, that is, $\|\nu\|_\infty \leq \mu^d/\sqrt{N^{d-1}N'}$.

Note that $p = 1$ when Δ is the GTF penalty operator with $D = D^{(1)}_{1\text{d}}$ and $p = k + 1$ when Δ is the KTF penalty operator with $D = D^{(k+1)}_{1\text{d}}$.

Proof of Lemma A.1. Abbreviate $\lambda_i = \gamma_i^2$ for $i \in [N]$. We are looking for a total of $N^d - p^d$ eigenvectors for $\Delta\Delta^T$. Assume for exposition that $d = 3$. For any $(i, j, k) \in [N]^d \setminus [p]^d$ (where $\setminus$ is the set difference operator), the vectors

$$z_{i,j,k} := \frac{1}{\sqrt{\lambda_i + \lambda_j + \lambda_k}} \begin{bmatrix} \gamma_i \cdot u_i \otimes v_j \otimes v_k \\ \gamma_j \cdot v_i \otimes u_j \otimes v_k \\ \gamma_k \cdot v_i \otimes v_j \otimes u_k \end{bmatrix} \tag{A.8}$$

are eigenvectors of $\Delta\Delta^T$ as verified below.

$$\begin{aligned} \Delta\Delta^T \begin{bmatrix} \gamma_i \cdot u_i \otimes v_j \otimes v_k \\ \gamma_j \cdot v_i \otimes u_j \otimes v_k \\ \gamma_k \cdot v_i \otimes v_j \otimes u_k \end{bmatrix} &= \Delta \ \left(\gamma_i^2 + \gamma_j^2 + \gamma_k^2\right) v_i \otimes v_j \otimes v_k \\ &= (\lambda_i + \lambda_j + \lambda_k) \begin{bmatrix} \gamma_i \cdot u_i \otimes v_j \otimes v_k \\ \gamma_j \cdot v_i \otimes u_j \otimes v_k \\ \gamma_k \cdot v_i \otimes v_j \otimes u_k \end{bmatrix} \end{aligned} \tag{A.9}$$

We see all $N^d - p^d$ eigenvectors of $\Delta\Delta^T$ here. Notice that $\|z_{i,j,k}\|_2 = 1$ and the incoherence is readily available provided the left and right singular vectors of D are incoherent.

For general d, these $N^d - p^d$ eigenvectors are given by

$$z_{i_1,i_2,.,i_d} = \frac{1}{\sqrt{\sum_{j=1}^d \lambda_{i_j}}} \begin{bmatrix} \gamma_{i_1} \cdot u_{i_1} \otimes v_{i_2} \otimes \ldots v_{i_d} \\ \gamma_{i_2} \cdot v_{i_1} \otimes u_{i_2} \otimes \ldots v_{i_d} \\ \vdots \\ \gamma_{i_d} \cdot v_{i_1} \otimes v_{i_2} \otimes \ldots u_{i_d} \end{bmatrix} \tag{A.10}$$

with eigenvalues $\sum_{j=1}^d \lambda_{i_j}$ and are easily seen to be incoherent. □

A.5 Proofs of lower bounds for GTF and KTF classes

Here and henceforth, we use the notation $B_p(r) = \{x : \|x\|_p \leq r\}$ for the ℓ_p ball of radius r, where $p, r > 0$ (and the ambient dimension will be determined based on the context).

We begin with a very simple lemma, that will help us embed ℓ_1 balls inside the GTF and KTF classes.

Lemma A.2. Let $\mathcal{T}(r) = \{\theta \in \mathbb{R}^n : \|\Delta\theta\|_1 \leq r\}$ for a matrix Δ and $r > 0$. Recall that $\|\Delta\|_{1,\infty} = \max_{i\in[n]} \|\Delta_i\|_1$ where Δ_i is the ith column of Δ. Then for any $r > 0$, it holds that $B_1(r/\|\Delta\|_{1,\infty}) \subseteq \mathcal{T}(r)$.

Proof. The proof follows from the observation that, for any θ,

$$\|\Delta\theta\|_1 = \Big\| \sum_{i=1}^n \Delta_i \theta_i \Big\|_1 \leq \sum_{i=1}^n \|\Delta_i\|_1 |\theta_i| \leq \Big(\max_{i=1,\ldots,n} \|\Delta_i\|_1 \Big) \|\theta\|_1 = d_{\max} \|\theta\|_1.$$

□

Corollary A.1. For any $r > 0$, and integers $d \geq 1, k \geq 0$,

$$B_1(r/(2^{k+1}d)) \subseteq \mathcal{T}_d^k(r), \quad B_1(r/(2^{k+1}d)) \subseteq \widetilde{\mathcal{T}}_d^k(r).$$

Proof. These containments follow from Lemma A.2 and the facts $\|\Delta^{(k+1)}\|_{1,\infty} = 2^{k+1}d$, $\|\widetilde{\Delta}^{(k+1)}\|_{1,\infty} = 2^{k+1}d$. □

To prove Theorems 2.4,2.5 we will rely on a result from Birge & Massart (2001), which gives a lower bound for the risk in a normal means problem, over ℓ_p balls. Another related, earlier result is that of Donoho & Johnstone (1994*a*); however, the Birge and Massart result places no restrictions on the radius of the ball in question, whereas the Donoho and Johnstone result does. Translated into our notation, the Birge and Massart result is as follows.

Lemma A.3 (Proposition 5 of Birge & Massart (2001))**.** Assume i.i.d. observations $y_i \sim N(\theta_{0,i}, \sigma^2)$, $i = 1, \ldots, n$, and $n \geq 2$. Then the minimax risk over the ℓ_p ball $B_p(r_n)$, where $0 < p < 2$, satisfies

$$n \cdot R(B_p(r_n)) \geq c \cdot \begin{cases} \sigma^{2-p} r_n^p \Big[1 + \log\Big(\frac{\sigma^p n}{r_n^p} \Big) \Big]^{1-p/2} & \text{if } \sigma\sqrt{\log n} \leq r_n \leq \sigma n^{1/p}/\sqrt{\rho_p} \\ r_n^2 & \text{if } r_n < \sigma\sqrt{\log n} \\ \sigma^2 n/\rho_p & \text{if } r_n > \sigma n^{1/p}/\sqrt{\rho} \end{cases}.$$

Here $c > 0$ is a universal constant, and $\rho_p > 1.76$ is the unique solution of $\rho_p \log \rho_p = 2/p$.

A.5.1 Proof of Theorem 2.5

It sufficies to show that the minimax optimal risk $R\big(\widetilde{\mathcal{T}}_d^k(C_n)\big)$ is lower bounded by the three terms present in the statement's lower bound separately:

$$\begin{aligned} R\big(\widetilde{\mathcal{T}}_d^k(C_n)\big) &= \Omega\Big(\frac{\kappa\sigma^2}{n} \Big), \\ R\big(\widetilde{\mathcal{T}}_d^k(C_n)\big) &= \Omega\Big(\Big(\frac{C_n}{n} \Big)^{\frac{2d}{2k+2+d}} \Big), \\ R\big(\widetilde{\mathcal{T}}_d^k(C_n)\big) &= \Omega\Big(\frac{C_n}{n} \Big) \end{aligned} \tag{A.11}$$

where $\kappa = \text{nullity}\big(\widetilde{\Delta}^{(k+1)}\big) = (k+1)^d$. First, as the null space of $\widetilde{\Delta}^{(k+1)}$ has dimension κ, we get the first lower bound:

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \widetilde{\mathcal{T}}_d^k(C_n)} \frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta_0\|_2^2 \;\geq\; \inf_{\hat{\theta}} \sup_{\theta_0 \in \text{null}\left(\widetilde{\Delta}^{(k+1)}\right)} \frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta_0\|_2^2 \;\geq\; \frac{\kappa\sigma^2}{n}.$$

We get the second lower bound in (A.11) by using the ℓ_1-ball embedding

$$B_1\left(C_n/d_{\max}\right) \subset \widetilde{\mathcal{T}}_d^k(C_n)$$

from Corollary A.1 and then using A.3.

Finally, we will show that

$$R\big(\mathcal{H}_d^{k+1}(L_n)\big) = \Omega(n^{-\frac{2k+2}{2k+2+d}} L_n^{\frac{2d}{2k+2+d}}). \tag{A.12}$$

Taking $L_n = C_n/n^{1-(k+1)/d}$ and applying Lemma 2.3 would then give the third lower bound in (A.11). This result is "nearly" a textbook result on Holder classes in nonparametric regression. A standard result (e.g., see Chapter 2.8 of Korostelev & Tsybakov (2003)) is that, in a model

$$y_i = f_0(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n$$

where the design points $x_i \in [0,1]^d$, $i = 1, \ldots, n$ are fixed and arbitrary, we have

$$\inf_{\hat{f}} \sup_{f_0 \in H(k+1, L_n; [0,1]^d)} \mathbb{E}\|\hat{f} - f_0\|_2^2 = \Omega(n^{-\frac{2k+2}{2k+2+d}} L_n^{\frac{2d}{2k+2+d}}), \tag{A.13}$$

where $\|\cdot\|_2$ denotes the L_2 norm on functions, defined as

$$\|f\|_2^2 = \int_{[0,1]^d} f(x)^2\, dx.$$

Note that we can rewrite the desired result (A.12) as

$$\inf_{\hat{f}} \sup_{f_0 \in H(k+1, L_n; [0,1]^d)} \mathbb{E}\|\hat{f} - f_0\|_n^2 = \Omega(n^{-\frac{2k+2}{2k+2+d}} L_n^{\frac{2d}{2k+2+d}}), \tag{A.14}$$

where the design points are $\{x_1, \ldots, x_n\} = Z_d$, the regular lattice on $[0,1]^d$, and where $\|\cdot\|_n$ denotes the empirical norm on functions, defined as

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2.$$

The proof of (A.13) reduces the estimation problem to a multiple hypothesis testing problem, and then constructs a sufficiently hard set of hypothesis by taking linear combinations of kernel "bump" functions and applying the Varshamov–Gilbert lemma (e.g., see Sections 2.7, 2.8 of Korostelev & Tsybakov (2003), or Section 2.6 of Tsybakov (2009)). But in the standard construction, the bump functions are not only orthogonal with respect to the L_2 inner product, but also with respect to the empirical inner product, since their supports

are nonoverlapping. Thus the exact same sequence of arguments leads to (A.14), i.e., leads to (A.12), provided the empirical norm a bump function is at least of the same order as its L_2 norm, as verified below.

Consider a partition of $[0,1]^d$ into $m \asymp n^{d/(2k+2+d)}$ hypercubes, each hypercube having side length $h = 1/m^{1/d} \asymp n^{-1/(2k+2+d)}$. Denote by z_i, $i = 1, \ldots, m$ the hypercube centers and consider bump functions $\varphi_i(x) = \varphi(x - z_i)$, $i = 1, \ldots, n$, where

$$\varphi(x) = h^{k+1} K\left(\frac{2\|x\|_2}{h}\right), \quad \text{where } K(u) = \exp\left(\frac{-1}{1-u^2}\right) 1\{|u| < 1\}.$$

In the L_2 norm, it holds that $\|\varphi_i\|_2^2 \asymp h^{2k+2+d}$, $i = 1, \ldots, n$. We want to show the empirical norms are lower bounded at the same rate. By symmetry, it suffices to study one bump function, say, φ_1. Denote by U_1 the set of grid points lying in a sphere of radius $h/(2\sqrt{2})$ around z_1. As $K(u) \geq 1/e^2$ for $|u| \leq 1/\sqrt{2}$, we have $\varphi_1(x) \geq h^{k+1}/e^2$ for $x \in U_1$. But the number of elements in U_1 is on the order of nh^d, and therefore $\|\varphi_1\|_n^2 = \Omega(h^d h^{2k+2}) = \Omega(h^{2k+2+d})$, as desired. □

A.5.2 Proof of Theorem 2.4

As in the proof of Theorem 2.5, it is sufficient to show three similar lower bounds. We get the first two lower bounds just as in the proof of Theorem 2.5 by using the fact that $\text{nullity}\big(\Delta^{(k+1)}\big) = 1$ and the ℓ_1-ball embedding

$$B_1(C_n/(2^{k+1}d)) \subset \mathcal{T}_d^k(C_n)$$

from Corollary A.1. The third lower bound is obtained in a different route as follows. Define a class

$$\mathcal{S}_d^{k+1} = \{\theta \in \mathbb{R}^n : \|\Delta^{(k+1)}\theta\|_2 \leq B_n\} = \{\theta \in \mathbb{R}^n : \theta^T L^{k+1} \theta \leq B_n^2\}.$$

Notice that $\mathcal{S}_d^{k+1}(B_n) \subseteq \mathcal{T}_d^{k+1}(C_n)$ provided $B_n = C_n/\sqrt{r}$, where $r \asymp n$ is the number of rows of of $\Delta^{(k+1)}$, owing to the simple inequality $\|x\|_1 \leq \sqrt{r}\|x\|_2$ for $x \in \mathbb{R}^n$. We will show that

$$R\big(\mathcal{S}_d^{k+1}(B_n)\big) = \Omega(n^{-\frac{d}{2k+2+d}} B_n^{\frac{2d}{2k+2+d}}). \tag{A.15}$$

Taking $B_n \asymp C_n/\sqrt{n}$ would then give the result.

Letting $L = U\Lambda U^T$ be an eigendecomposition, and note that for any estimator $\hat{\theta}$ of θ_0,

$$\|\hat{\theta} - \theta_0\|_2 = \|U^T\hat{\theta} - U^T\theta_0\|_2,$$

which means that we may rotate the parameter space and equivalently consider the minimax error over the rotated class

$$\widetilde{\mathcal{S}}_d^{k+1} = \bigg\{\gamma \in \mathbb{R}^n : \sum_{i=1}^n \lambda_i^{k+1}\gamma_i^2 \leq B_n^2\bigg\},$$

where we have denoted the eigenvalues (diagonal elements of Λ) as λ_i, $i \in [n]$. We will now seek to embed a hyperrectangle in the above class and make use of results of Donoho et al. (1990).

Write $\gamma = (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^{n-1}$, and order $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, so the above class becomes

$$\widetilde{\mathcal{S}}_d^{k+1} = \left\{ (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^{n-1} : \sum_{i=2}^n \lambda_i^{k+1} \beta_i^2 \leq B_n^2 \right\} := \mathbb{R} \times \mathcal{E}(B_n),$$

where we have used the fact that $\lambda_1 = 0$. (Here and henceforth, although unconventional, we will index β according to components $i = 2, \ldots, n$, rather than $i = 1, \ldots, n-1$, because it simplifies notation later.) The minimax risk (writing $\gamma_0 = U^T \theta_0$, and $\gamma_0 = (\alpha_0, \beta_0)$) satisfies

$$\inf_{\hat{\gamma}} \sup_{\gamma_0 \in \widetilde{\mathcal{S}}_d^{k+1}} \frac{1}{n} \mathbb{E} \|\hat{\gamma} - \gamma_0\|_2^2 = \frac{\sigma^2}{n} + \inf_{\hat{\beta}} \sup_{\beta_0 \in \mathcal{E}(B_n)} \frac{1}{n} \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2.$$

We focus on the second term. The ellipsoid $\mathcal{E}(B_n)$ is compact, convex, orthosymmetric and quadratically convex, the latter property as defined in Donoho et al. (1990). We can therefore use Lemma 6 and Theorem 7 in their work to conclude that the minimax risk over $\mathcal{E}(B_n)$ is at least four-fifths of the minimax linear risk of its hardest hyperrectangle,

$$\inf_{\hat{\beta}} \sup_{\beta_0 \in \mathcal{E}(B_n)} \frac{1}{n} \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2 \geq \frac{4}{5} \sup_{H \subseteq \mathcal{E}(B_n)} \inf_{\hat{\beta}\,\text{linear}} \sup_{\beta_0 \in H} \frac{1}{n} \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2, \tag{A.16}$$

where the outer sup on the right-hand side is over hyperrectangles H contained in $\mathcal{E}(B_n)$. Consider hyperrectangles parametrized by a threshold τ,

$$H(\tau) = \{\beta \in \mathbb{R}^{n-1} : |\beta_i| \leq t_i(\tau),\ i = 2, \ldots, n\},$$

where for all $i = 2, \ldots, n$, using multi-index notation $i = (i_1, \ldots, i_d)$, we let

$$t_{i+1}(\tau) = \begin{cases} B_n / (\sum_{i_1, \ldots, i_d \leq \tau} \lambda_i^{k+1})^{1/2} & \text{if } i_1, \ldots, i_d \leq \tau \\ 0 & \text{else.} \end{cases}$$

It is not hard to check that $H(\tau) \subseteq \mathcal{E}(B_n)$. The minimax linear risk over $H(\tau)$ decomposes, and can be evaluated exactly, as in Donoho et al. (1990),

$$\inf_{\hat{\beta}\,\text{linear}} \sup_{\beta_0 \in H(\tau)} \frac{1}{n} \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2 = \frac{1}{n} \sum_{i=2}^n \frac{t_i(\tau)^2 \sigma^2}{t_i(\tau)^2 + \sigma^2} = \frac{1}{n} \frac{(\tau^d - 1)\sigma^2 B_n^2}{B_n^2 + \sum_{i_1, \ldots, i_d \leq \tau} \lambda_i^{k+1}}.$$

Lemma A.7 provides an upper bound on the sum in the denominator above, and plugging this in, we get

$$\inf_{\hat{\beta}\,\text{linear}} \sup_{\beta_0 \in H(\tau)} \frac{1}{n} \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2 \geq \frac{1}{n} \frac{(\tau^d - 1)\sigma^2 B_n^2}{B_n^2 + c \frac{\tau^{2k+2+d}}{N^{2k+2}}},$$

for a constant $c > 0$. This lower bound is maximized at $\tau \asymp (B_n^2 N^{2k+2})^{\frac{1}{2k+2+d}}$, in which case, we see

$$\inf_{\hat{\beta}\,\text{linear}} \sup_{\beta_0 \in H(\tau)} \frac{1}{n} \mathbb{E} \|\hat{\beta} - \beta_0\|_2^2 = \Omega(n^{-\frac{d}{2k+2+d}} B_n^{\frac{2d}{2k+2+d}}).$$

Recalling (A.16), we have hence shown (A.15), and this completes the proof.

A.6 Proof of minimax linear rate for GTF and KTF classes

A.6.1 Proof of Theorem 2.6 (minimax linear rates over TV classes)

First we recall a few definitions, from Donoho et al. (1990). Given a set $A \subseteq \mathbb{R}^k$, its *quadratically convex hull* $\mathrm{qconv}(A)$ is defined as

$$\mathrm{qconv}(A) = \big\{(x_1, \ldots, x_k) : (x_1^2, \ldots, x_k^2) \in \mathrm{conv}(A_+^2)\big\}, \quad \text{where}$$
$$A_+^2 = \big\{(a_1^2, \ldots, a_k^2) : a \in A,\ a_i \geq 0,\ i = 1, \ldots, k\big\}.$$

(Here $\mathrm{conv}(B)$ denotes the convex hull of a set B.) Furthermore, the set A is called *quadratically convex* provided that $\mathrm{qconv}(A) = A$. Also, A is called *orthosymmetric* provided that $(a_1, \ldots, a_k) \in A$ implies $(\sigma_1 a_1, \ldots, \sigma_k a_k) \in A$, for any choice of signs $\sigma_1, \ldots, \sigma_k \in \{-1, 1\}$.

Now we proceed with the proof. Following from equation (7.2) of Donoho et al. (1990),

$$\mathrm{qconv}\big(B_1(C_n/d_{\max})\big) = B_2(C_n/d_{\max}).$$

Theorem 11 of Donoho et al. (1990) states that, for orthosymmetric, compact sets, such as $B_1(C_n/d_{\max})$, the minimax linear risk equals that of its quadratically convex hull. Moreover, Theorem 7 of Donoho et al. (1990) tells us that for sets that are orthosymmetric, compact, convex, and quadratically convex, such as $B_2(C_n/d_{\max})$, the minimax linear risk is the same as the minimax linear risk over the worst rectangular subproblem. We consider $B_\infty(C_n/(d_{\max}\sqrt{n}))$, and abbreviate $r_n = C_n/(d_{\max}\sqrt{n})$. It is fruitful to study rectangles because the problem separates across dimensions, as in

$$\begin{aligned}\inf_{\hat\theta \text{ linear}} \sup_{\theta_0 \in B_\infty(r_n)} \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n (\hat\theta_i - \theta_{0,i})^2\bigg] &= \frac{1}{n}\sum_{i=1}^n \bigg[\inf_{\hat\theta_i \text{ linear}} \sup_{|\theta_{0,i}| \leq r_n} \mathbb{E}(\hat\theta_i - \theta_{0,i})^2\bigg] \\ &= \inf_{\hat\theta_1 \text{ linear}} \sup_{|\theta_{0,1}| \leq r_n} \mathbb{E}(\hat\theta_1 - \theta_{0,1})^2.\end{aligned}$$

Thus it suffices to compute the minimax linear risk over the 1d class $\{\theta_{0,1} : |\theta_{0,1}| \leq r_n\}$. It is easily shown (e.g., see Section 2 of Donoho et al. (1990)) that this is $r_n^2\sigma^2/(r_n^2 + \sigma_2^2)$, and so this is precisely the minimax linear risk for $B_2(C_n/d_{\max})$, and for $B_1(C_n/d_{\max})$.

To get the first lower bound as stated in the theorem, we simply take a maximum of $r_n^2\sigma^2/(r_n^2 + \sigma_2^2)$ and σ^2/n, as the latter is the minimax risk for estimating a 1-dimensional mean parameter given n observations in a normal model with variance σ^2. To get the second, we use the fact that $2ab/(a+b) \geq \min\{a, b\}$. This completes the proof. □

A.6.2 Alternative proof of Theorem 2.6

Here, we reprove Theorem 2.6 using elementary arguments. We write $y = \theta_0 + \epsilon$, for $\epsilon \sim N(0, \sigma^2 I)$. Given an arbitary linear estimator, $\hat\theta = Sy$ for a matrix $S \in \mathbb{R}^{n\times n}$, observe that

$$\begin{aligned}\mathbb{E}\big[\mathrm{MSE}(\hat\theta, \theta_0)\big] = \frac{1}{n}\mathbb{E}\|\hat\theta - \theta_0\|_2^2 &= \frac{1}{n}\mathbb{E}\|S(\theta_0 + \epsilon) - \theta_0\|_2^2 \\ &= \frac{1}{n}\mathbb{E}\|S\epsilon\|_2^2 + \frac{1}{n}\|(S - I)\theta_0\|_2^2 \\ &= \frac{\sigma^2}{n}\|S\|_F^2 + \frac{1}{n}\|(S - I)\theta_0\|_2^2, \qquad \text{(A.17)}\end{aligned}$$

which we may view as the variance and (squared) bias terms, respectively. Now denote by e_i the ith standard basis vector, and consider

$$\begin{aligned}
\frac{\sigma^2}{n}\|S\|_F^2 + \left(\sup_{\theta_0:\|D\theta_0\|_1 \le C_n} \frac{1}{n}\|(S-I)\theta_0\|_2^2\right) &\ge \frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{d_{\max}^2 n}\left(\max_{i=1,\ldots,n} \|(I-S)e_i\|_2^2\right) \\
&\ge \frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{d_{\max}^2 n^2}\sum_{i=1}^n \|(I-S)e_i\|_2^2 \\
&= \frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{d_{\max}^2 n^2}\|(I-S)\|_F^2 \\
&\ge \frac{\sigma^2}{n}\sum_{i=1}^n S_{ii}^2 + \frac{C_n^2}{d_{\max}^2 n^2}\sum_{i=1}^n (1-S_{ii})^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(\sigma^2 S_{ii}^2 + \frac{C_n^2}{d_{\max}^2 n}(1-S_{ii})^2\right).
\end{aligned}$$

Here S_{ii}, $i = 1, \ldots, n$ denote the diagonal entries of S. To bound each term in the sum, we apply the simple inequality $ax^2 + b(1-x)^2 \ge ab/(a+b)$ for all x (since a short calculation shows that the quadratic in x here is minimized at $x = b/(a+b)$). We may continue on lower bounding the last displayed expression, giving

$$\frac{\sigma^2}{n}\|S\|_F^2 + \left(\sup_{\theta_0:\|D\theta_0\|_1 \le C_n} \frac{1}{n}\|(S-I)\theta_0\|_2^2\right) \ge \frac{\sigma^2 C_n^2}{C_n^2 + \sigma^2 d_{\max}^2 n}.$$

Lastly, we may take the maximum of this with σ^2/n in order to derive a final lower bound, as argued in the proof of Theorem 2.6. □

A.6.3 Proof of Lemma 2.5 (mean estimator over TV classes)

For this estimator, the smoother matrix is $S = \mathbb{1}\mathbb{1}^T/n$ and so $\|S\|_F^2 = 1$. From (A.17), we have

$$\mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{mean}}, \theta_0)\big] = \frac{\sigma^2}{n} + \frac{1}{n}\|\theta_0 - \bar{\theta}_0\mathbb{1}\|_2^2,$$

where $\bar{\theta}_0 = (1/n)\sum_{i=1}^n \theta_{0,i}$. Now

$$\begin{aligned}
\sup_{\theta_0:\|D\theta_0\|_1 \le C_n} \frac{1}{n}\|\theta_0 - \bar{\theta}_0\mathbb{1}\|_2^2 &= \sup_{x \in \mathrm{row}(D):\|Dx\|_1 \le C_n} \frac{1}{n}\|x\|_2^2 \\
&= \sup_{z \in \mathrm{col}(D):\|z\|_1 \le C_n} \frac{1}{n}\|D^\dagger z\|_2^2 \\
&\le \sup_{z:\|z\|_1 \le C_n} \frac{1}{n}\|D^\dagger z\|_2^2 \\
&= \frac{C_n^2}{n}\max_{i=1,\ldots,n} \|D_i^\dagger\|_2^2 \\
&\le \frac{C_n^2 M_n^2}{n},
\end{aligned}$$

which establishes the desired bound. □

A.6.4 General proof

Proof of Theorem 2.7. The proofs of Theorems 2.7 and 2.8 follow the same line of reasoning. We give the proof only for Theorem 2.8 because it is slightly more involved to control the summation of inverse of eigenvalues of $\Delta^T\Delta$ where $\Delta = \widetilde{\Delta}^{(k+1)}$. □

Proof of Theorem 2.8. For brevity, denote $\Delta = \widetilde{\Delta}^{(k+1)}$ in the context of this proof. The minimax linear risk for the class $\widetilde{\mathcal{T}}_d^k(C_n)$ is

$$\begin{aligned} R_L(\widetilde{\mathcal{T}}_d^k(C_n)) &= \inf_{S\in\mathbb{R}^{n\times n}} \sup_{\theta_0\in\widetilde{\mathcal{T}}_d^k(C_n)} \frac{1}{n}\mathbb{E}\|Sy-\theta_0\|_2^2 \\ &= \inf_{S} \sup_{\theta_0\in\widetilde{\mathcal{T}}_d^k(C_n)} \frac{1}{n}\mathbb{E}\|S(\theta_0+\epsilon)-\theta_0\|_2^2 \\ &= \frac{1}{n}\inf_{S} \sup_{\theta_0\in\widetilde{\mathcal{T}}_d^k(C_n)} \sigma^2\|S\|_F^2 + \|(S-I)\theta_0\|_2^2 \end{aligned}$$

where in the last line we used the assumption that $\epsilon_i, i \in [n]$ are i.i.d. with mean zero and variance σ^2 and used the notation $\|A\|_F$ for the Frobenius norm of a matrix A. The infimum can be restricted to the set of linear smoothers

$$\mathcal{S} = \{S : \text{null}(S-I) \supseteq \text{null}(\Delta)\}$$

because if for a linear smoother S, if there exists $\eta \in \text{null}(\Delta)$ suc that $(S-I)\eta \neq 0$, then the inner supremum above will be ∞, that is, its risk will be ∞. If the outer infimum is over $\mathcal{S}$, then the supremum can be restricted to $\{\theta_0 \in \text{row}(\Delta) : \theta \in \widetilde{\mathcal{T}}_d^k(C_n)\}$. We continue to lower bound minimax linear risk as follows:

$$\begin{aligned} R_L(\widetilde{\mathcal{T}}_d^k(C_n)) &= \frac{1}{n}\inf_{S\in\mathcal{S}} \sigma^2\|S\|_F^2 + \sup_{\theta_0\in\text{row}(\Delta):\|\Delta\theta_0\|_1\le C_n} \|(S-I)\theta_0\|_2^2 \\ &= \frac{1}{n}\inf_{S\in\mathcal{S}} \sigma^2\|S\|_F^2 + \sup_{z:\|z\|_1\le C_n} \|(S-I)\Delta^+ z\|_2^2 \\ &= \frac{1}{n}\inf_{S\in\mathcal{S}} \sigma^2\|S\|_F^2 + C_n^2 \max_{i\in[m]} \left\|\left((S-I)\Delta^+\right)_i\right\|_2^2 \qquad \text{(A.18)} \\ &\ge \frac{1}{n}\inf_{S\in\mathcal{S}} \sigma^2\|S\|_F^2 + \frac{C_n^2}{m}\sum_{i=1}^{m} \left\|\left((S-I)\Delta^+\right)_i\right\|_2^2 \\ &\ge \inf_{S\in\mathcal{S}} \underbrace{\frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{mn}\left\|(S-I)\Delta^+\right\|_F^2}_{=:r(S)} \qquad \text{(A.19)} \end{aligned}$$

In the third line, $(A)_i$ denotes the ith column of matrix A and m denotes the number of rows in Δ. In the fourth line, we used the fact that the maximum of a set is at least as much as their average. In the last line, we use the fact that $m \ge dn$ and also – within the context of this proof – define the quantity $r(S)$ which is a lower bound on the risk of the linear smoother $S \in \mathcal{S}$.

Notice that $r(\cdot)$ is a quadratic in the entries of S and the constraint $S \in \mathcal{S}$ translates to linear constraints on the entries of S. Writing the KKT conditions, after some work, we see that $r(\cdot)$ is minimized at

$$S_0 = a_n \left(\sigma^2 L^{(k+1)} + a_n I\right)^{-1} \tag{A.20}$$

where we denote $a_n = \frac{C_n^2}{m}$ and $L^{(k+1)} = \Delta^T \Delta$. Further, $S_0 \in \mathcal{S}$. Therefore,

$$R_L(\widetilde{\mathcal{T}}_d^k(C_n)) \geq r(S_0). \tag{A.21}$$

We simplify the expression for $r(S_0)$ now. Let $\lambda_i, i \in [n]$ be the eigenvalues of $L^{(k+1)}$. Then the eigenvalues of S_0 are

$$\frac{a_n}{\sigma^2 \lambda_i + a_n}, i \in [n]$$

and the non-zero squared singular values of $(S - I)\Delta^+$ are given by

$$\frac{\sigma^4 \lambda_i}{(\sigma^2 \lambda_i + a_n)^2}, \quad \kappa < i \leq n.$$

Using the fact that the squared Frobenius norm of a matrix is the sum of squares of its singular values, substituting the above eigenvalues and singular values in (A.19), we have

$$\begin{aligned} r(S_0) &= \frac{\sigma^2}{n} \sum_{i=1}^{n} \left(\frac{a_n}{\sigma^2 \lambda_i + a_n}\right)^2 + \frac{a_n}{n} \sum_{i=1}^{n} \frac{\sigma^4 \lambda_i}{(\sigma^2 \lambda_i + a_n)^2} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2 \lambda_i + a_n}. \end{aligned} \tag{A.22}$$

Now we upper bound the risk $R(S_0)$ of the linear smoother defined by S_0. From (A.18), we can write

$$R(S_0) = \frac{\sigma^2}{n} \|S_0\|_F^2 + \frac{C_n^2}{n} \max_{i \in [m]} \left\| \left((S_0 - I)\Delta^+\right)_i \right\|_2^2.$$

Let $\Delta = U \Sigma V^T$ be the singular value decomposition of Δ. Also let the eigen-decomposition of $S_0 - I = V \Lambda V^T$. Then using incoherence of columns of U, that is, the fact that there exists a constant $c > 1$ that depends only on k, d such that $U_{ij}^2 \leq \frac{c}{m}$ for all $i \in [m], j \in [n]$, we can write

$$\begin{aligned} \max_{i \in [m]} \left\| \left((S_0 - I)\Delta^+\right)_i \right\|_2^2 &= \max_{i \in [m]} \left\| V \Lambda V^T V \Sigma^+ (U^T)_i \right\|_2^2 \\ &= \max_{i \in [m]} (U^T)_i^T (\Lambda \Sigma^+)^2 (U^T)_i \\ &\leq \frac{c}{m} \mathrm{tr}\left((\Lambda \Sigma^+)^2\right) \\ &= \frac{c}{m} \sum_{i=1}^{n} \frac{\sigma^4 \lambda_i}{(\sigma^2 \lambda_i + a_n)^2}. \end{aligned}$$

Plugging this back in the previous display and also using the fact that the squared Frobenius norm of a matrix is equal to the sum of the squares of its eigenvalues,

$$\begin{aligned} R(S_0) &= \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{a_n}{\sigma^2\lambda_i + a_n}\right)^2 + \frac{c\cdot a_n}{n}\sum_{i=1}^{n}\frac{\sigma^4\lambda_i}{(\sigma^2\lambda_i + a_n)^2} \\ &\le c\cdot r(S_0) \end{aligned}$$

Combining this with the lower bound in (A.21), we have

$$r(S_0) \le R_L(\widetilde{\mathcal{T}}_d^k(C_n)) \le \min\left\{\sigma^2, R(S_0)\right\} \le \min\left\{\sigma^2,\ \ c\cdot r(S_0)\right\}. \tag{A.23}$$

In other words, the minimax linear rate is essentially $r(S_0)$ up to a constant factor. Further, one of the estimators $\hat{y} = S_0 y$, $\hat{y} = y$ achieves the minimax linear rate up to a constant factor.

Now we bound $r(S_0)$. Let $\kappa = (k+1)^d$ denote the nullity of Δ. Recall from (A.22)

$$r(S_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2 a_n}{\sigma^2\lambda_i + a_n} = \frac{\kappa\sigma^2}{n} + \frac{1}{n}\sum_{i=\kappa+1}^{n}\frac{\sigma^2 a_n}{\sigma^2\lambda_i + a_n}. \tag{A.24}$$

Lower bounding $r(S_0)$. First, we give two lower bounds on $r(S_0)$. By using the fact that arithmetic mean of positive numbers is at least as large as their harmonic mean, we have

$$\begin{aligned} r(S_0) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2 a_n}{\sigma^2\lambda_i + a_n} \\ &\ge \frac{n\sigma^2 a_n}{\sum_{i=1}^{n}(\sigma^2\lambda_i + a_n)} \\ &= \frac{n\sigma^2 a_n}{na_n + \sigma^2\|\Delta\|_F^2} \\ &= \frac{n\sigma^2 a_n}{na_n + \sigma^2 dn^{1-1/d}\|D_{1\mathrm{d}}^{(k+1)}\|_F^2} \\ &= \frac{\sigma^2 a_n}{a_n + \sigma^2 dn^{-1/d}(n^{1/d} - k - 1)\binom{2k+2}{k+1}} \\ &\ge \frac{\sigma^2 a_n}{a_n + \sigma^2 d 4^{k+1}} \end{aligned} \tag{A.25}$$

Now we bound in a different way. Let n_1 be the cardinality of $\{i \in [n] : \sigma^2\lambda_i \le a_n\}$. Then

$$r(S_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2 a_n}{\sigma^2\lambda_i + a_n} \ge \frac{1}{n}\sum_{i=1}^{n_1}\frac{\sigma^2 a_n}{a_n + a_n} = \frac{n_1\sigma^2}{2n}.$$

Note that $n_1 = \lfloor nF(a_n/\sigma^2)\rfloor$ where F is the spectral distribution of $(\widetilde{\Delta}^{(k+1)})^T\widetilde{\Delta}^{(k+1)}$ defined

in Lemma A.4. Applying Lemma A.4, we get

$$\begin{aligned} r(S_0) &\geq \frac{\sigma^2}{2}\left(F\left(\frac{a_n}{\sigma^2}\right) - \frac{1}{n}\right) \\ &\geq c_1 \frac{\sigma^2}{2}\left(\frac{a_n}{\sigma^2}\right)^{\frac{d}{2k+2}} \\ &= c\sigma^{2-\frac{d}{k+1}} a_n^{\frac{d}{2k+2}} \end{aligned} \tag{A.26}$$

From (A.24),(A.25) and (A.26) we have the lower bound

$$r(S_0) \geq \max\left\{\frac{\kappa\sigma^2}{n}, \frac{\sigma^2 a_n}{a_n + \sigma^2 d 2^{2k+2}}, c\sigma^{2-\frac{d}{k+1}} a_n^{\frac{d}{2k+2}}\right\}. \tag{A.27}$$

Upper bounding $r(S_0)$. If $2k+2 < d$, then

$$\begin{aligned} r(S_0) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2\lambda_i + a_n} \\ &\leq \frac{\kappa\sigma^2}{n} + \frac{1}{n}\sum_{i=\kappa+1}^{n} \frac{\sigma^2 a_n}{\sigma^2\lambda_i} \\ &= \frac{\kappa\sigma^2}{n} + \frac{a_n}{n}\sum_{i=1}^{\kappa+1} \frac{1}{\lambda_i} \\ &\leq \frac{\kappa\sigma^2}{n} + \frac{a_n}{n}(c_3 n) \\ &= \frac{\kappa\sigma^2}{n} + c_3 a_n \end{aligned} \tag{A.28}$$

We used Lemma A.10 to control the second term in the third line. By the same reasoning, if $2k+2 = d$, $r(S_0) \leq \frac{\kappa\sigma^2}{n} + c_3 a_n \log n$. For the case $2k+2 > d$, we can write

$$\begin{aligned} r(S_0) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2\lambda_i + a_n} \\ &\leq \frac{1}{n}\sum_{i=1}^{n_1} \frac{\sigma^2 a_n}{a_n} + \frac{1}{n}\sum_{i=n_1+1}^{n} \frac{\sigma^2 a_n}{2\sigma^2\lambda_i} \\ &= \frac{n_1\sigma^2}{n} + \frac{a_n}{2n}\sum_{i=n_1+1}^{n} \frac{1}{\lambda_i} \\ &\leq \frac{\kappa\sigma^2}{n} + c_2\sigma^2\left(\frac{a_n}{\sigma^2}\right)^{\frac{d}{2k+2}} + \frac{a_n}{2n} n^{\frac{2k+2}{d}} \left(n(a_n/\sigma^2)^{\frac{d}{2k+2}}\right)^{1-(2k+2)/d} \\ &\leq \frac{\kappa\sigma^2}{n} + c\sigma^{2-\frac{d}{k+1}} a_n^{\frac{d}{2k+2}} \end{aligned} \tag{A.29}$$

In the fourth line, we used Lemma A.4 to bound n_1 and Lemma A.10 to bound the summation. From the upper bounds in (A.28),(A.29) we conclude that the lower bound in (A.27) is tight up to a constant factor (or a $\log n$ factor in the case $2k+2 = d$).

Risk of $\hat{\theta}^{\text{poly}}$. For brevity, denote $\Pi = P_{\text{null}(\Delta)}$. Note that $(I - \Pi)\Delta^+ = \Delta^+$. From (A.38),

$$\begin{aligned}\sup_{\theta_0 \in \widetilde{\mathcal{T}}_d^k(C_n)} \mathbb{E}\big[\text{MSE}(\hat{\theta}^{\text{poly}}, \theta_0)\big] &= \frac{\sigma^2}{n}\|\Pi\|_F^2 + \max_{i \in [m]} \|\big((\Pi - I)\Delta^+\big)_i\|_2^2 \\ &= \frac{\kappa\sigma^2}{n} + \max_{i \in [m]} \|\Delta_i^+\|_2^2\end{aligned}$$

Then using incoherence of columns of U, that is, the fact that there exists a constant $c > 1$ that depends only on k, d such that $U_{ij}^2 \leq \frac{c}{m}$ for all $i \in [m], j \in [n]$, we can write

$$\begin{aligned}\max_{i \in [m]} \|\Delta_i^+\|_2^2 &= \max_{i \in [m]} \big\|V\Sigma^+(U^T)_i\big\|_2^2 \\ &= \max_{i \in [m]} (U^T)_i^T (\Sigma^+)^2 (U^T)_i \\ &\leq \frac{c}{m} \text{tr}\left((\Sigma^+)^2\right) \\ &= \frac{c}{m} \sum_{i=\kappa+1}^{n} \frac{1}{\lambda_i}\end{aligned}$$

Plugging this back in the above display and using the bound on $\sum_{i=\kappa+1}^n \frac{1}{\lambda_i}$ from Lemma A.11, we get the desired result. □

Remark A.1. In the case $2k+2 \leq d$, the desired lower bound may also be obtained by embedding the ℓ_1-ball $B_1(C_n/(2^{k+1}d))$ in $\mathcal{T}_d^k(C_n)$ in Theorem 2.7 (or $\widetilde{\mathcal{T}}_d^k(C_n)$ in Theorem 2.8). We apply results from Donoho et al. (1990) just as in the proof of Theorem 2.6 to get the lower bound.

Lemma A.4. Let $\widetilde{\Delta}^{(k+1)}$ be the $(k+1)$th order KTF operator on a d-dimensional regular grid with n vertices. Let $\lambda_i, i \in [n]$ be the eigenvalues of $(\widetilde{\Delta}^{(k+1)})^T \widetilde{\Delta}^{(k+1)}$ with the ordering $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Define

$$F(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\lambda_i \leq t\}, \qquad \text{for } t \in [0, \lambda_n].$$

Then there exists a constants $c_1, c_2 > 0$ independent of n such that

$$c_1 t^{\frac{d}{2k+2}} \leq F(t) - F(0) \leq c_2 t^{\frac{d}{2k+2}}$$

for all $t \in [0, \lambda_n]$.

Proof of Lemma A.4. Let $N = n^{1/d}, N' = N - k - 1$. Let $D = D_{\text{1d}}^{(k+1)} \in \mathbb{R}^{N' \times N}$ and let G be the $(k+1)$th order GTF operator on a chain of length N. As in Lemma A.2 Sadhanala et al. (2017), we tie together the eigenvalues of DD^T and GG^T using Cauchy interlacing theorem.

Let $N'' = N - \mathbb{1}\{k \text{ is even}\}$ for brevity. Let $\alpha_i, i \in [N'']$ be the eigenvalues of GG^T and $\beta_i, i \in [N']$ the eigenvalues of DD^T. Cauchy interlacing theorem (as applied in Lemma A.2 Sadhanala et al. (2017)) tells us that

$$\alpha_i \leq \beta_i \leq \alpha_{i+N''-N'}, \qquad \text{for } i \in [N'].$$

From the Kronecker product structure of $(\widetilde{\Delta}^{(k+1)})^T\widetilde{\Delta}^{(k+1)}$, we can index its eigenvalues $\lambda_i, i \in [n]$ alternatively using the grid positions, as in

$$\lambda_{i_1,\ldots,i_d} = \rho_{i_1} + \cdots + \rho_{i_d}, \qquad \text{for } (i_1,\ldots,i_d) \in [N]^d$$

where $\rho_1 = \cdots = \rho_{k+1} = 0$ and $\rho_{i+k+1} = \beta_i, i \in [N']$. From the interlacing result displayed above, and the fact that the eigenvalues of the Laplacian are given by $4\sin^2\frac{\pi(i-1)}{2N}$ for $i \in [N]$, we have

$$\left(4\sin^2\frac{\pi(i-k-1)_+}{2N}\right)^{k+1} \le \rho_i \le \left(4\sin^2\frac{\pi(i-1)}{2N}\right)^{k+1}, \qquad \text{for } i \in [N] \tag{A.30}$$

where $(x)_+ = \max\{x, 0\}$ for $x \in \mathbb{R}$. The lower bound can be derived as follows. Certainly, $F(t) \ge F(0) = \kappa/n$ for $t \ge 0$. We can write

$$\begin{aligned}
nF(t) &= \sum_{i\in[N]^d} \mathbb{1}\{\lambda_{i_1,\ldots,i_d} \le t\} \\
&= \sum_{i\in[N]^d} \mathbb{1}\{\sum_{j=1}^d \rho_{i_j} \le t\} \\
&= \sum_{i\in[N]^d} \mathbb{1}\{\sum_{j=1}^d 4^{k+1}\sin^{2k+2}\frac{\pi(i_j-1)}{2N} \le t\} \\
&\ge \sum_{i\in[N]^d} \mathbb{1}\{\sum_{j=1}^d \pi^{2k+2}(i_j-1)^{2k+2} \le tN^{2k+2}\} \\
&\ge c_1 n t^{\frac{d}{2k+2}}
\end{aligned}$$

In the third line, we used (A.30) and in the fourth line, we used the fact that $\sin x \le x$ for $x \ge 0$. In the last line, we used the fact that the the number of (integer) lattice points in the ℓ_{2k+2} body

$$x_1^{2k+2} + \cdots + x_d^{2k+2} \le r^{2k+2}$$

is close to its volume, which is given by $c_1' r^d$ for a constant c_1' that depends only on d, k.

The upper bound can be argued in a similar manner.

$$\begin{aligned}
nF(t) &= \sum_{i\in[N]^d} \mathbb{1}\{\lambda_{i_1,\ldots,i_d} \le t\} \\
&= \sum_{i\in[N]^d} \mathbb{1}\{\sum_{j=1}^d \rho_{i_j} \le t\} \\
&\le \sum_{i\in[N]^d} \mathbb{1}\{\sum_{j=1}^d 4^{k+1}\sin^{2k+2}\frac{\pi(i_j-k-1)_+}{2N} \le t\} \\
&\le \sum_{i\in[N]^d} \mathbb{1}\{\sum_{j=1}^d \left(\frac{\pi}{2}\right)^{2k+2}(i_j-k-1)_+^{2k+2} \le tN^{2k+2}\} \\
&\le \kappa + c_2 n t^{\frac{d}{2k+2}}
\end{aligned}$$

In the third line, we again use (A.30) and in the fourth line, we use the fact that $\sin x \geq x/2$ for $x \in [0, \pi/2]$. In the last line, we argue about the the number of lattice points in the ℓ_{2k+2} body just as in the lower bound result in the preceding display. □

A.7 Analysis over Sobolev classes

Define the Sobolev ball $\mathcal{S}_d(C_n')$, of radius C_n' is defined as

$$\mathcal{S}_d(C_n') = \big\{\theta : \|\Delta^{(1)}\theta\|_2 \leq C_n'\big\}. \tag{A.31}$$

The following lemma shows that the Sobolev class contains a Holder ball of certain radius.

Lemma A.5. For any integers $k \geq 0$, $d \geq 1$, the (discretized) Holder and Sobolev classes defined in (2.15), (A.31) satisfy $\mathcal{H}_d^1(L) \subseteq \mathcal{S}_d(cLn^{1/2-1/d})$, where $c > 0$ is a constant depending only on k.

Our first result here is a lower bound on the minimax risk of the Sobolev class $\mathcal{S}_d(C_n')$ in (A.31).

Theorem A.1. For a universal constant $c > 0$,

$$R(\mathcal{S}_d(C_n')) \geq \frac{c}{n}\Big((n\sigma^2)^{\frac{2}{d+2}}(C_n')^{\frac{2d}{d+2}} \wedge n\sigma^2 \wedge n^{2/d}(C_n')^2\Big) + \frac{\sigma^2}{n}.$$

Elegant tools for minimax analysis from Donoho et al. (1990), which leverage the fact that the ellipsoid $\mathcal{S}_d(C_n')$ is orthosymmetric and quadratically convex (after a rotation), are used to prove the result.

The next theorem gives upper bounds, certifying that the above lower bound is tight, and showing that Laplacian eigenmaps and Laplacian smoothing, both linear smoothers, are optimal over $\mathcal{S}_d(C_n')$ for all d and for $d = 1, 2$, or 3 respectively.

Theorem A.2. For Laplacian eigenmaps, $\hat{\theta}^{\mathrm{LE}}$ in (2.5), with $k \asymp ((n(C_n')^d)^{2/(d+2)} \vee 1) \wedge n$, we have

$$\sup_{\theta_0 \in \mathcal{S}_d(C_n')} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LE}}, \theta_0)\big] \leq \frac{c}{n}\Big((n\sigma^2)^{\frac{2}{d+2}}(C_n')^{\frac{2d}{d+2}} \wedge n\sigma^2 \wedge n^{2/d}(C_n')^2\Big) + \frac{c\sigma^2}{n},$$

for a universal constant $c > 0$, and n large enough. When $d = 1, 2$, or 3, the same bound holds for Laplacian smoothing $\hat{\theta}^{\mathrm{LS}}$ in (2.5), with $\lambda \asymp (n/(C_n')^2)^{2/(d+2)}$ (and a possibly different constant c).

Remark A.2. As shown in the proof, Laplacian smoothing is nearly minimax rate optimal over $\mathcal{S}_d(C_n')$ when $d = 4$, just incurring an extra log factor. It is unclear to us whether this method is still (nearly) optimal when $d \geq 5$; based on insights from our proof technique, we conjecture that it is not.

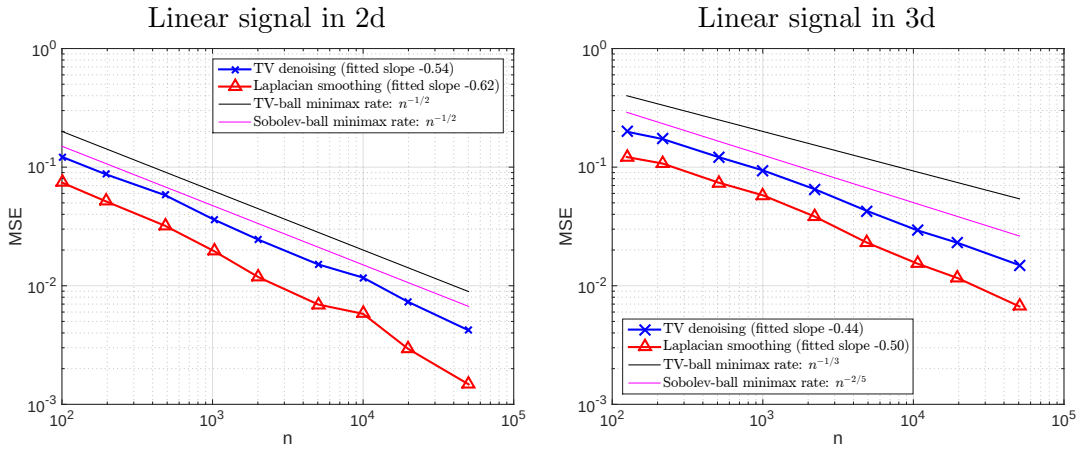

Figure A.1: *MSE curves for estimating a "linear" signal, a very smooth signal, over 2d and 3d grids. For each n, the results were averaged over 5 repetitions, and Laplacian smoothing and TV denoising were tuned for best average MSE performance. The signal was set to satisfy $\|D\theta_0\|_2 \asymp n^{1/2-1/d}$, matching the canonical scaling.*

A.8 A phase transition, and adaptivity

The TV class $\mathcal{T}_d^0(C_n)$ in (2.13) (with $k = 0$) and the Sobolev class in (A.31), display a curious relationship. We reflect on the minimax optimal rates from Theorems 2.4 and A.1, using, for concreteness, the canonical scalings $C_n \asymp n^{1-1/d}$ and $C'_n \asymp n^{1/2-1/d}$, that, recall, guarantee $\mathcal{S}_d(C'_n) \subseteq \mathcal{T}_d(C_n)$. (Similar statements could also be made outside of this case, subject to an appropriate relationship with $C_n/C'_n \asymp \sqrt{n}$.) When $d = 1$, both the TV and Sobolev classes have a minimax rate of $n^{-2/3}$ (this TV result is actually due to Donoho & Johnstone (1998), as stated in (1.12), not Theorem 2.4). When $d = 2$, both the TV and Sobolev classes again have the same minimax rate of $n^{-1/2}$, the caveat being that the rate for TV class has an extra $\sqrt{\log n}$ factor. But for all $d \geq 3$, the rates for the canonical TV and Sobolev classes differ, and the smaller Sobolev spaces have faster rates than their inscribing TV spaces. This may be viewed as a phase transition at $d = 3$; see Table A.1.

Function class	Dimension 1	Dimension 2	Dimension $d \geq 3$
TV ball $\mathcal{T}_d(n^{1-1/d})$	$n^{-2/3}$	$n^{-1/2}\sqrt{\log n}$	$n^{-1/d}\sqrt{\log n}$
Sobolev ball $\mathcal{S}_d(n^{1/2-1/d})$	$n^{-2/3}$	$n^{-1/2}$	$n^{-\frac{2}{2+d}}$

Table A.1: *Summary of rates for canonically-scaled TV and Sobolev spaces.*

We may paraphrase to say that 2d is just like 1d, in that expanding the Sobolev ball into a larger TV ball does not hurt the minimax rate, and methods like TV denoising are automatically *adaptive*, i.e., optimal over both the bigger and smaller classes. However, as soon as we enter the 3d world, it is no longer clear whether TV denoising can adapt to the smaller, inscribed Sobolev ball, whose minimax rate is faster, $n^{-2/5}$ versus $n^{-1/3}$ (ignoring log factors). Theoretically, this is an interesting open problem that we do not approach here and leave to future work.

We do, however, investigate the matter empirically: see Figure A.1, where we run Laplacian smoothing and TV denoising on a highly smooth "linear" signal θ_0. This is constructed so that each component θ_i is proportional to $i_1 + i_2 + \ldots + i_d$ (using the multi-index notation $(i_1, \ldots, i_d)$ of (2.15) for grid location i), and the Sobolev norm is $\|D\theta_0\|_2 \asymp n^{1/2-1/d}$. Arguably, these are among the "hardest" types of functions for TV denoising to handle. The left panel, in 2d, is a case in which we know that TV denoising attains the minimax rate; the right panel, in 3d, is a case in which we do not, though empirically, TV denoising surely seems to be doing better than the slower minimax rate of $n^{-1/3}$ (ignoring log terms) that is associated with the larger TV ball.

Even if TV denoising is shown to be minimax optimal over the inscribed Sobolev balls when $d \geq 3$, note that this does not necessarily mean that we should scrap Laplacian smoothing in favor of TV denoising, in all problems. Laplacian smoothing is the unique Bayes estimator in a normal means model under a certain Markov random field prior (e.g., Sharpnack & Singh (2010)); statistical decision theory therefore tells that it is *admissible*, i.e., no other estimator—TV denoising included—can uniformly dominate it.

A.9 Proof of minimax rates over Sobolev classes

A.9.1 Proof of Theorem A.1 (minimax rates over Sobolev classes)

Recall that we denote by $L = V\Sigma V^T$ the eigendecomposition of the graph Laplacian $L = D^T D$, where $\Sigma = \mathrm{diag}(\rho_1, \ldots, \rho_n)$ with $0 = \rho_1 < \rho_2 \leq \ldots \leq \rho_n$, and where $V \in \mathbb{R}^{n \times n}$ has orthonormal columns. Also denote by $D = U\Sigma^{1/2}V^T$ the singular value decomposition of the edge incidence matrix D, where $U \in \mathbb{R}^{m \times n}$ has orthonormal columns.[1] First notice that

$$\|D\theta_0\|_2 = \|U\Sigma^{1/2}V^T\theta_0\|_2 = \|\Sigma^{1/2}V^T\theta_0\|_2.$$

This suggests that a rotation by V^T will further simplify the minimax risk over $\mathcal{S}_d(C_n')$, i.e.,

$$\begin{aligned} \inf_{\hat{\theta}} \sup_{\theta_0 : \|\Sigma^{1/2}V^T\theta_0\|_2 \leq C_n'} \frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta_0\|_2^2 &= \inf_{\hat{\theta}} \sup_{\theta_0 : \|\Sigma^{1/2}V^T\theta_0\|_2 \leq C_n'} \frac{1}{n}\mathbb{E}\|V^T\hat{\theta} - V^T\theta_0\|_2^2 \\ &= \inf_{\hat{\gamma}} \sup_{\gamma_0 : \|\Sigma^{1/2}\gamma_0\|_2 \leq C_n'} \frac{1}{n}\mathbb{E}\|\hat{\gamma} - \gamma_0\|_2^2, \end{aligned} \tag{A.32}$$

where we have rotated and now consider the new parameter $\gamma_0 = V^T\theta_0$, constrained to lie in

$$\mathcal{E}_d(C_n') = \bigg\{ \gamma : \sum_{i=2}^n \rho_i \gamma_i^2 \leq (C_n')^2 \bigg\}.$$

To be clear, in the rotated setting (A.32) we observe a vector $y' = V^T y \sim N(\gamma_0, \sigma^2 I)$, and the goal is to estimate the mean parameter γ_0. Since there are no constraints along the first dimension, we can separate out the MSE in (A.32) into that incurred on the first

[1] When $d = 1$, we have $m = n - 1$ edges, and so it is not be possible for U to have orthonormal columns; however, we can just take its first column to be all 0s, and take the rest as the eigenbasis for $\mathbb{R}^{n-1}$, and all the arguments given here will go through.

component, and all other components. Decomposing $\gamma_0 = (\alpha_0, \beta_0) \in \mathbb{R}^{1\times(n-1)}$, with similar notation for an estimator $\hat{\gamma}$,

$$\begin{aligned}\inf_{\hat{\gamma}} \sup_{\gamma_0 \in \mathcal{E}_d(C_n')} \frac{1}{n}\mathbb{E}\|\hat{\gamma} - \gamma_0\|_2^2 &= \inf_{\hat{\alpha}} \sup_{\alpha_0} \frac{1}{n}\mathbb{E}(\hat{\alpha} - \alpha_0)^2 + \inf_{\hat{\beta}} \sup_{\beta_0 \in P_{-1}(\mathcal{E}_d(C_n'))} \frac{1}{n}\mathbb{E}\|\hat{\beta} - \beta_0\|_2^2 \\ &= \frac{\sigma^2}{n} + \inf_{\hat{\beta}} \sup_{\beta_0 \in P_{-1}(\mathcal{E}_d(C_n'))} \frac{1}{n}\mathbb{E}\|\hat{\beta} - \beta_0\|_2^2, \end{aligned} \tag{A.33}$$

where P_{-1} projects onto all coordinate axes but the 1st, i.e., $P_{-1}(x) = (0, x_2, \ldots, x_n)$, and in the second line we have used the fact that the minimax risk for estimating a 1-dimensional parameter α_0 given an observation $z \sim N(\alpha_0, \sigma^2)$ is simply σ^2.

Let us lower bound the second term in (A.33), i.e., $R(P_{-1}(\mathcal{E}_d(C_n')))$. The ellipsoid $P_{-1}(\mathcal{E}_d(C_n'))$ is orthosymmetric, compact, convex, and quadratically convex, hence Theorem 7 in Donoho et al. (1990) tells us that its minimax linear risk is the minimax linear risk of its hardest rectangular subproblem. Further, Lemma 6 in Donoho et al. (1990) then tells us the minimax linear risk of its hardest rectangular subproblem is, up to a constant factor, the same as the minimax (nonlinear) risk of the full problem. More precisely, Lemma 6 and Theorem 7 from Donoho et al. (1990) imply

$$\frac{5}{4} R(P_{-1}(\mathcal{E}_d(C_n'))) \geq R_L(P_{-1}(\mathcal{E}_d(C_n'))) = \sup_{H \subseteq P_{-1}(\mathcal{E}_d(C_n'))} R_L(H), \tag{A.34}$$

where the supremum above is taken over all rectangular subproblems, i.e., all rectangles H contained in $P_{-1}(\mathcal{E}_d(C_n'))$.

To study rectangular subproblems, it helps to reintroduce the multi-index notation for a location i on the d-dimensional grid, writing this as $(i_1, \ldots, i_d) \in \{1, \ldots, N\}^d$, where $N = n^{1/d}$. For a parameter $2 \leq \tau \leq N$, we consider rectangular subsets of the form[2]

$$\begin{gathered} H(\tau) = \big\{\beta \in \mathbb{R}^{n-1} : |\beta_i| \leq t_i(\tau),\ i = 2, \ldots, n\big\}, \quad \text{where} \\ t_i(\tau) = \begin{cases} C_n' / (\sum_{j_1,\ldots,j_d \leq \tau} \rho_{j_1,\ldots,j_d})^{1/2} & \text{if } i_1, \ldots, i_d \leq \tau \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i = 2, \ldots, n. \end{gathered}$$

It is not hard to check that $H(\tau) \subseteq \{\beta \in \mathbb{R}^{n-1} : \sum_{i=2}^n \rho_i \beta_i^2 \leq (C_n')^2\} = P_{-1}(\mathcal{E}_d(C_n'))$. Then, from (A.34),

$$\begin{aligned} \frac{5}{4} R(P_{-1}(\mathcal{E}_d(C_n'))) \geq \sup_{\tau} R_L(H(\tau)) &= \sup_{\tau} \frac{1}{n} \sum_{i=1}^n \frac{t_i(\tau)^2 \sigma^2}{t_i(\tau)^2 + \sigma^2} \\ &= \sup_{\tau} \frac{1}{n} \frac{(\tau^d - 1)\sigma^2 (C_n')^2}{(C_n')^2 + \sum_{j_1,\ldots,j_d \leq \tau} \rho_{j_1,\ldots,j_d}}. \end{aligned}$$

The first equality is due to the fact that the minimax risk for rectangles decouples across dimensions, and the 1d minimax linear risk is straightforward to compute for an interval, as

[2]Here, albeit unconventional, it helps to index $\beta \in H(\tau) \subseteq \mathbb{R}^{n-1}$ according to components $i = 2, \ldots, n$, rather than $i = 1, \ldots, n-1$. This is so that we may keep the index variable i to be in correspondence with positions on the grid.

argued in the proof Theorem 2.6; the second equality simply comes from a short calculation following the definition of $t_i(\tau)$, $i = 2, \dots, n$. Applying Lemma A.8, on the eigenvalues of the graph Laplacian matrix L for a d-dimensional grid, we have that for a constant $c > 0$,

$$\frac{(\tau^d - 1)\sigma^2 (C_n')^2}{(C_n')^2 + \sum_{j_1,\dots,j_d \le \tau} \rho_{j_1,\dots,j_d}} \ge \frac{(\tau^d - 1)\sigma^2 (C_n')^2}{(C_n')^2 + c\sigma^2 \tau^{d+2}/N^2} \ge \frac{1}{2} \frac{\sigma^2 (C_n')^2}{(C_n')^2 \tau^{-d} + c\sigma^2 \tau^2 / N^2}.$$

We can choose τ to maximize the expression on the right above, given by

$$\tau^* = \left(\frac{N^2 (C_n')^2}{c\sigma^2} \right)^{\frac{1}{d+2}}.$$

When $2 \le \tau^* \le N$, this provides us with the lower bound on the minimax risk

$$\frac{5}{4} R(P_{-1}(\mathcal{E}_d(C_n'))) \ge R_L(H(\tau^*)) \ge \frac{1}{2n} \frac{\tau^{d} \sigma^2 (C_n')^2}{2(c\sigma^2)^{\frac{d}{d+2}} (C_n')^{\frac{4}{d+2}} N^{-\frac{2d}{d+2}}} = \frac{c_1}{n} (n\sigma^2)^{\frac{2}{d+2}} (C_n')^{\frac{2d}{d+2}}, \tag{A.35}$$

for a constant $c_1 > 0$. When $\tau^* < 2$, we can use $\tau = 2$ as lower bound on the minimax risk,

$$\frac{5}{4} R(P_{-1}(\mathcal{E}_d(C_n'))) \ge R_L(H(2)) \ge \frac{1}{2n} \frac{\sigma^2 N^2 (C_n')^2}{N^2 (C_n')^2 2^{-d} + c\sigma^2 2^2} \ge \frac{c_2}{n} N^2 (C_n')^2, \tag{A.36}$$

for a constant $c_2 > 0$, where in the last inequality, we used the fact that $N^2 (C_n')^2 \le c\sigma^2 2^{d+2}$ (just a constant) since we are in the case $\tau^* < 2$. Finally, when $\tau^* > N$, we can use $\tau = N$ as a lower bound on the minimax risk,

$$\frac{5}{4} R(P_{-1}(\mathcal{E}_d(C_n'))) \ge R_L(H(N)) \ge \frac{1}{2n} \frac{\sigma^2 (C_n')^2}{N^{-d} (C_n')^2 + c\sigma^2} \ge c_3 \sigma^2, \tag{A.37}$$

for a constant $c_3 > 0$, where in the last inequality, we used that $c\sigma^2 \le N^{-d} (C_n')^2$ as we are in the case $\tau^* > N$. Taking a minimum of the lower bounds in (A.35), (A.36), (A.37), as a way to navigate the cases, gives us a final lower bound on $R(P_{-1}(\mathcal{E}_d(C_n')))$, and completes the proof.

A.9.2 Proof of Theorem A.2 (Laplacian eigenmaps and Laplacian smoothing over Sobolev classes)

We will prove the results for Laplacian eigenmaps and Laplacian separately.

Given an arbitary linear estimator, $\hat{\theta} = Sy$ for a matrix $S \in \mathbb{R}^{n \times n}$, observe that

$$\begin{aligned} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big] = \frac{1}{n} \mathbb{E} \|\hat{\theta} - \theta_0\|_2^2 &= \frac{1}{n} \mathbb{E} \|S(\theta_0 + \epsilon) - \theta_0\|_2^2 \\ &= \frac{1}{n} \mathbb{E} \|S\epsilon\|_2^2 + \frac{1}{n} \|(S - I)\theta_0\|_2^2 \\ &= \frac{\sigma^2}{n} \|S\|_F^2 + \frac{1}{n} \|(S - I)\theta_0\|_2^2, \end{aligned} \tag{A.38}$$

which we may view as the variance and (squared) bias terms, respectively.

Laplacian eigenmaps. The smoother matrix for this estimator is $S_k = V_{[k]}V_{[k]}^T$, for a tuning parameter $k = 1, \dots, n$. From (A.38),

$$\mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LE}}, \theta_0)\big] = \frac{\sigma^2}{n}k + \frac{1}{n}\|(I - S_k)\theta_0\|_2^2.$$

Now we write $k = \tau^d$, and analyze the max risk of the second term,

$$\begin{aligned}
\sup_{\theta_0 : \|D\theta_0\|_2 \le C_n'} \frac{1}{n}\|(I - S_k)\theta_0\|_2^2 &= \sup_{z : \|z\|_2 \le C_n'} \frac{1}{n}\|(I - S_k)D^\dagger z\|_2^2 \\
&= \frac{(C_n')^2}{n}\sigma_{\max}^2\big((I - S_k)D^\dagger\big) \\
&\le \frac{(C_n')^2}{n}\frac{1}{4\sin^2(\pi\tau/(2N))} \\
&\le \frac{(C_n')^2}{n}\frac{4N^2}{\pi^2\tau^2}.
\end{aligned}$$

Here we denote by $\sigma_{\max}(A)$ the maximum singular value of a matrix A. The last inequality above used the simple lower bound $\sin(x) \ge x/2$ for $x \in [0, \pi/2]$. The earlier inequality used that

$$(I - S_k)D^\dagger = (I - V_{[k]}V_{[k]}^T)V^T(\Sigma^\dagger)^{1/2}U^T = \big[0, \dots, 0, V_{k+1}, \dots, V_n\big](\Sigma^\dagger)^{1/2}U^T,$$

where we have kept the same notation for the singular value decomposition of D as in the proof of Theorem A.1. Therefore $\sigma_{\max}^2((I - S_k)D^\dagger)$ is the reciprocal of the $(k+1)$st smallest eigenvalue ρ_{k+1} of the graph Laplacian L. For any subset A of the set of eigenvalues $\lambda(L) = \{\rho_1, \dots, \rho_n\}$ of the Laplacian, with $|A| = k$, note that $\rho_{k+1} \ge \min \lambda(L) \setminus A$. This means that, for our d-dimensional grid,

$$\begin{aligned}
\rho_{k+1} &\ge \min\ \lambda(L) \setminus \{\rho_{i_1,\dots,i_d} : i_1, \dots, i_d \le \tau\} \\
&= 4\sin^2(\pi\tau/(2N)),
\end{aligned}$$

where recall $N = n^{1/d}$, as explained by (A.39), in the proof of Lemma A.8.

Hence, we have established

$$\sup_{\theta_0 : \|D\theta_0\|_2 \le C_n'} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LE}}, \theta_0)\big] \le \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\tau^d + \frac{(C_n')^2}{n}\frac{4N^2}{\pi^2\tau^2}.$$

Choosing τ to balance the two terms on the right-hand side above results in $\tau^* = (2NC_n'/(\pi\sigma))^{\frac{2}{d+2}}$. Plugging in this choice of τ, while utilizing the bounds $1 \le \tau \le N$, very similar to the arguments given at the end of the proof of Theorem A.1, gives the result for Laplacian eigenmaps.

Laplacian smoothing. The smoother matrix for this estimator is $S_\lambda = (I + \lambda L)^{-1}$, for a tuning parameter $\lambda \ge 0$. From (A.38),

$$\mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LS}}, \theta_0)\big] = \frac{\sigma^2}{n}\sum_{i=1}^n \frac{1}{(1 + \lambda\rho_i)^2} + \frac{1}{n}\|(I - S_\lambda)\theta_0\|_2^2.$$

When $d = 1, 2$, or 3, the first term upper is bounded by $c_1\sigma^2/n + c_2\sigma^2/\lambda^{d/2}$, for some constants $c_1, c_2 > 0$, by Lemma A.9. As for the second term,

$$\begin{aligned}
\sup_{\theta_0 : \|D\theta_0\|_2 \le C_n'} \frac{1}{n}\|(I - S_\lambda)\theta_0\|_2^2 &= \sup_{z : \|z\|_2 \le C_n'} \|(I - S_\lambda)D^\dagger z\|_2^2 \\
&= \frac{(C_n')^2}{n}\sigma_{\max}^2\big((I - S_\lambda)D^\dagger\big) \\
&= \frac{(C_n')^2}{n}\max_{i=2,\dots,n}\left(1 - \frac{1}{1+\lambda\rho_i}\right)^2\frac{1}{\rho_i} \\
&= \frac{(C_n')^2}{n}\lambda\max_{i=2,\dots,n}\frac{\lambda\rho_i}{(1+\lambda\rho_i)^2} \\
&\le \frac{(C_n')^2\lambda}{4n}.
\end{aligned}$$

In the third equality we have used the fact the eigenvectors of $I - S_\lambda$ are the left singular vectors of $D^\dagger$, and in the last inequailty we have used the simple upper bound $f(x) = x/(1+x)^2 \le 1/4$ for $x \ge 0$ (this function being maximized at $x = 1$).

Therefore, from what we have shown,

$$\sup_{\theta_0 : \|D\theta_0\|_2 \le C_n'} \mathbb{E}\big[\mathrm{MSE}(\hat\theta^{\mathrm{LS}}, \theta_0)\big] \le \frac{c_1\sigma^2}{n} + \frac{c_2\sigma^2}{\lambda^{d/2}} + \frac{(C_n')^2\lambda}{4n}.$$

Choosing λ to balance the two terms on the right-hand side above gives $\lambda^* = c(n/(C_n')^2)^{2/(d+2)}$, for a constant $c > 0$. Plugging in this choice, and using upper bounds from the trivial cases $\lambda = 0$ and $\lambda = \infty$ when C_n' is very small or very large, respectively, gives the result for Laplacian smoothing. □

Remark A.3. When $d = 4$, Lemma A.9 gives a slightly worse upper bound on $\sum_{i=1}^n 1/(1+\lambda\rho_i)^2$, with an "extra" term $(nc_2/\lambda^{d/2}))\log(1 + c_3\lambda)$, for constants $c_2, c_3 > 0$. It is not hard to show, by tracing through the same arguments as given above that we can use this to establish an upper bound on the max risk of

$$\sup_{\theta_0 \in \mathcal{S}_d(C_n')} \mathbb{E}\big[\mathrm{MSE}(\hat\theta^{\mathrm{LE}}, \theta_0)\big] \le \frac{c}{n}\Big((n\sigma^2)^{\frac{2}{d+2}}(C_n')^{\frac{2d}{d+2}}\log(n/(C_n')^2) \wedge n\sigma^2 \wedge n^{2/d}(C_n')^2\Big) + \frac{c\sigma^2}{n},$$

only slightly worse than the minimax optimal rate, by a log factor.

When $d \ge 5$, our analysis provides a much worse bound for the max risk of Laplacian smoothing, as the integral denoted $I(d)$ in the proof of Lemma A.9 grows very large when $d \ge 5$. We conjecture that this not due to slack in our proof technique, but rather, to the Laplacian smoothing estimator itself, since all inequalities the proof are fairly tight.

A.10 Utility lemmas

This section contains some calculations on the partial sums of eigenvalues of the Laplacian matrix L, for d-dimensions grids.

A.10.1 Lemma A.6

The next lemma is the key driver for the sharp rate established in Theorem 2.2. Here and henceforth, denote $[i] = \{1, \dots, i\}$ for an integer $i \geq 1$.

Lemma A.6. Let $\xi_1 \leq \ldots \leq \xi_{n-1}$ be the nonzero singular values of the GTF operator $\Delta^{(k+1)}$ of order $k+1$. If $k = 0$, then for any $i_0 \in [n-1]$,

$$\sum_{i=i_0+1}^{n-1} \frac{1}{\xi_i^2} \leq cn \log(n/i_0).$$

for large enough n, where $c > 0$ is an absolute constant. If $k > 1$, then for any $i_0 \in [n-1]$,

$$\sum_{i=i_0+1}^{n-1} \frac{1}{\xi_i^2} \leq cn^{k+1}/i_0^k,$$

for large enough n, where now $c > 0$ is a constant depending only on k.

Proof. In the following, we denote by $c > 0$ a constant whose value may change from line to line, as needed.

Let us denote by $\lambda_1 \leq \ldots \leq \lambda_{n-1}$ the nonzero eigenvalues of the Laplacian of the 2d grid graph of size $N \times N$. As shown in Wang et al. (2016), the GTF operator $\Delta^{(k+1)}$ has squared singular values $\xi_i^2 = \lambda_i^{k+1}$, $i \in [n-1]$. We can index the eigenvalues of the Laplacian by 2d grid positions, and we note (as, e.g., in the proof of Corollary 8 in Wang et al. (2016)) that they may be written as

$$\lambda_{i_1,i_2} = 4 \sin^2 \left(\frac{\pi(i_1 - 1)}{2N} \right) + 4 \sin^2 \left(\frac{\pi(i_2 - 1)}{2N} \right), \quad i_1, i_2 \in [N].$$

For the first claim in the lemma, take $j_0 = \lfloor \sqrt{i_0} \rfloor$. Observe, using $\sin(x) \geq x/2$ for $x \in [0, \pi/2]$,

$$\begin{aligned}
\sum_{i=i_0+1}^{n-1} \frac{1}{\lambda_i} &\leq \sum_{\min\{i_1,i_2\} \geq j_0+1} \frac{1}{\lambda_{i_1,i_2}} \\
&\leq cn \sum_{\min\{i_1,i_2\} \geq j_0+1} \frac{1}{(i_1-1)^2 + (i_2-1)^2} \\
&\leq cn \sum_{i_1=j_0}^{N-1} \sum_{i_2=1}^{N-1} \frac{1}{i_1^2 + i_2^2} \\
&\leq cn \sum_{i_1=j_0}^{N-1} \int_0^{N-1} \frac{1}{i_1^2 + x^2} \, dx \\
&= cn \sum_{i_1=j_0}^{N-1} \frac{1}{i_1} \tan^{-1} \left(\frac{N-1}{i_1} \right) \\
&\leq cn \sum_{i_1=j_0}^{N-1} \frac{1}{i_1} \frac{\pi}{2}
\end{aligned}$$

$$\leq cn\log(N/j_0),$$

for sufficiently large n.

As for the second claim in the lemma, observe, again using $\sin(x) \geq x/2$ for $x \in [0, \pi/2]$,

$$\begin{aligned}
\sum_{i=i_0+1}^{n-1} \frac{1}{\lambda_i^{k+1}} &\leq \sum_{(i_1-1)^2+(i_2-1)^2 \geq i_0}^{n} \frac{1}{\lambda_{i_1,i_2}^{k+1}} \\
&\leq cn^{k+1} \sum_{(i_1-1)^2+(i_2-1)^2 \geq i_0} \frac{1}{((i_1-1)^2+(i_2-1)^2)^{k+1}} \\
&\leq cn^{k+1} \left(\int_{i_0 \leq x^2+y^2 \leq 2(n-1),\, x,y \geq 0} \frac{1}{(x^2+y^2)^{k+1}}\, dx\, dy + \sum_{(i_1-1)^2+(i_2-1)^2=i_0} \frac{1}{i_0^{k+1}} \right) \\
&\leq cn^{k+1} \left(\int_0^{\pi/2} \int_{\sqrt{i_0}}^{\sqrt{2(n-1)}} \frac{1}{r^{2(k+1)}} r\, dr\, d\theta + \frac{1}{i_0^{k+1/2}} \right) \\
&\leq cn^{k+1} \left(\frac{\pi}{2} \int_{i_0}^{2(n-1)} \frac{1}{u^{k+1}}\, du + \frac{1}{i_0^{k+1/2}} \right) \\
&= cn^{k+1} \left(\frac{\pi}{2} \left(\frac{1}{i_0^k} - \frac{1}{(2(n-1))^k} \right) + \frac{1}{i_0^{k+1/2}} \right) \\
&\leq cn^{k+1}/i_0^k.
\end{aligned}$$

□

A.10.2 Lemma A.7

This result slightly generalizes Lemma A.3 of Sadhanala et al. (2016).

Lemma A.7. Let $L \in \mathbb{R}^{n \times n}$ denote the Laplacian matrix of the d-dimensional grid graph with equal side lengths $N = n^{1/d}$, and let

$$\lambda_{i_1,\ldots,i_d} = 4 \sum_{j=1}^{d} \sin^2\left(\frac{\pi(i_j-1)}{2N}\right), \quad i_1, \ldots, i_d \in [N]$$

denote its eigenvalues. Then for any integer $k \geq 0$ and $\tau \in [N]$,

$$\sum_{i_1,\ldots,i_d \leq \tau} \lambda_{i_1,\ldots,i_d}^{k+1} \leq c \frac{\tau^{2k+2+d}}{N^{2k+2}},$$

for a constant $c > 0$ depending only on k and d.

Proof. The proof follows the same chain of arguments as that for Lemma A.3 in Sadhanala et al. (2016). Using the fact that $\sin(x) \leq x$ for all $x \geq 0$,

$$\sum_{i_1,\ldots,i_d \leq \tau} \lambda_{i_1,\ldots,i_d}^{k+1} \leq \frac{\pi^{2k+2}}{4^k N^{2k+2}} \sum_{i_1,\ldots,i_d \leq \tau} \Big((i_1-1)^2 + \ldots + (i_d-1)^2 \Big)^{k+1}$$

$$\leq \frac{\pi^{2k+2}}{4^k N^{2k+2}} \tau^{d-1} \sum_{i=1}^{\tau} (i-1)^{2k+2}$$
$$\leq c \frac{\tau^{2k+2+d}}{N^{2k+2}}.$$

□

Lemma A.8. Let $L \in \mathbb{R}^{n \times n}$ denote the graph Laplacian matrix of a d-dimensional grid graph, and $\rho_{i_1,\ldots,i_d}$, $(i_1,\ldots,i_d) \in \{1,\ldots,N\}^d$ be its eigenvalues, where $N = n^{1/d}$. Then there exists a constant $c > 0$ (dependent on d) such that, for any $1 \leq \tau \leq N$,

$$\sum_{(i_1,\ldots,i_d) \in \{1,\ldots,\tau\}^d} \rho_{i_1,\cdots,i_d} \leq c \frac{\tau^{d+2}}{N^2}.$$

Proof. The eigenvalues of L can be written explicitly as

$$\rho_i = 4\sin^2\Big(\frac{\pi(i_1-1)}{2N}\Big) + \ldots + 4\sin^2\Big(\frac{\pi(i_d-1)}{2N}\Big), \quad (i_1,\ldots,i_d) \in \{1,\ldots,N\}^d. \tag{A.39}$$

This follows from known facts about the eigenvalues for the Laplacian matrix of a 1d grid, and the fact that the Laplacian matrix for higher-dimensional grids can be expressed in terms of a Kronecker sum of the Laplacian matrix of an appropriate 1d grid (e.g., Conte & de Boor (1980), Kunsch (1994), Ng et al. (1999), Wang et al. (2008, 2016), Hutter & Rigollet (2016)). We now use the fact that $\sin(x) \leq x$ for all $x \geq 0$, which gives us the upper bound

$$\begin{aligned}
\sum_{(i_1,\ldots,i_d) \in \{1,\ldots,\tau\}^d} \rho_{i_1,\cdots,i_d} &\leq \frac{\pi^2}{N^2} \sum_{(i_1,\ldots,i_d) \in \{1,\ldots,\tau\}^d} \Big((i_1-1)^2 + \ldots + (i_d-1)^2\Big) \\
&\leq \frac{\pi^2 d}{N^2} \tau^{d-1} \sum_{i=1}^{\tau} (i-1)^2 \\
&\leq \frac{\pi^2 d}{N^2} \tau^{d-1} \tau^3 \\
&= \frac{\pi^2 d}{N^2} \tau^{d+2},
\end{aligned}$$

as desired. □

Lemma A.9. Let $L \in \mathbb{R}^{n \times n}$ denote the graph Laplacian matrix of a d-dimensional grid graph, and ρ_i, $i = 1,\ldots,n$ be its eigenvalues. Let $\lambda \geq 0$ be arbitrary. For $d = 1, 2$, or 3, there are constants $c_1, c_2 > 0$ such that

$$\sum_{i=1}^{n} \frac{1}{(1+\lambda\rho_i)^2} \leq c_1 + c_2 \frac{n}{\lambda^{d/2}}.$$

For $d = 4$, there are constants $c_1, c_2, c_3 > 0$ such that

$$\sum_{i=1}^{n} \frac{1}{(1+\lambda\rho_i)^2} \leq c_1 + c_2 \frac{n}{\lambda^{d/2}} \Big(1 + \log(1 + c_3\lambda)\Big).$$

Proof. We will use the explicit form of the eigenvalues as given in the proof of Lemma A.8. In the expressions below, we use $c > 0$ to denote a constant whose value may change from line to line. Using the inequality $\sin x \geq x/2$ for $x \in [0, \pi/2]$,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{1}{(1+\lambda\rho_i)^2} &\leq \sum_{(i_1,\ldots,i_d)\in\{1,\ldots,N\}^d} \frac{1}{\left(1+\lambda\frac{\pi^2}{4N^2}\sum_{j=1}^{d}(i_j-1)^2\right)^2} \\
&\leq 1 + \int_{[0,N]^d} \frac{1}{\left(1+\lambda\frac{\pi^2}{4}\sum_{j=1}^{d} x_j^2/N^2\right)^2}\,dx \\
&= 1 + c\int_0^{N\sqrt{d}} \frac{1}{\left(1+\lambda\frac{\pi^2}{4}r^2/N^2\right)^2} r^{d-1}\,dr \\
&= 1 + c\frac{n}{\lambda^{d/2}} \underbrace{\int_0^{\frac{\pi}{2}\sqrt{\lambda d}} \frac{u^{d-1}}{(1+u^2)^2}\,du}_{I(d)}.
\end{aligned}$$

In the second inequality, we used the fact that the right-endpoint Riemann sum is always an underestimate for the integral of a function that is monotone nonincreasing in each coordinate. In the third, we made a change to spherical coordinates, and suppressed all of the angular variables, as they contribute at most a constant factor. It remains to compute $I(d)$, which can be done by symbolic integration:

$$\begin{aligned}
I(1) &= \frac{\pi\sqrt{d}}{4\left(1+\frac{\pi^2}{4}\lambda d\right)} + \frac{1}{2}\tan^{-1}\left(\frac{\pi}{2}\sqrt{\lambda d}\right) \leq \frac{1}{4} + \frac{\pi}{4}, \\
I(2) &= \frac{1}{2} - \frac{1}{2\left(1+\frac{\pi^2}{4}\lambda d\right)} \leq \frac{1}{2}, \\
I(3) &= \frac{1}{2}\tan^{-1}\left(\frac{\pi}{2}\sqrt{\lambda d}\right) \leq \frac{\pi}{4}, \quad \text{and} \\
I(4) &= \frac{1}{2}\log\left(1+\frac{\pi^2}{4}\lambda d\right) + \frac{1}{2\left(1+\frac{\pi^2}{4}\lambda d\right)} - \frac{1}{2} \leq \frac{1}{2}\log\left(1+\frac{\pi^2}{4}\lambda d\right) + \frac{1}{2}.
\end{aligned}$$

This completes the proof. □

Lemma A.10. Consider the eigenvalues $\{\rho_i : i = (i_1, \cdots, i_d) \in [N]^d\}$ of the d-dimensional grid graph Laplacian with $n = N^d$ nodes. Let k be a non-negative integer and $\beta \in (2, \sqrt{d}N)$. Then,

$$\sum_{i\in[N]^d:\|i-1\|_2^2\geq\beta^2} \frac{1}{\rho_i^k} \leq c\begin{cases} n & 2k < d \\ n\log(N/\beta) & 2k = d \\ N^{2k}\beta^{d-2k} & 2k > d \end{cases}$$

Proof of Lemma A.10. Let I denote the summation on the left. Then

$$I = \sum_{i\in[N]^d:\|i-1\|_2^2\geq\beta^2} \frac{1}{\rho_i^k} = \sum_{\|i-1\|_2^2\geq\beta^2} \frac{1}{\left(\sum_{j=1}^{d} 4\sin^2\frac{\pi(i_j-1)}{2N}\right)^k}$$

$$
\begin{aligned}
&\leq \sum_{\|i-1\|_2^2 \geq \beta^2} \frac{1}{\left(\sum_{j=1}^d \frac{\pi^2 (i_j-1)^2}{4N^2}\right)^k} \\
&= cN^{2k} \sum_{\|i-1\|_2^2 \geq \beta^2} \frac{1}{\left(\sum_{j=1}^d (i_j-1)^2\right)^k} \\
&\leq cN^{2k} \int_{\beta/2 \leq \|x\|_2 \leq \sqrt{d}N} \frac{1}{\left(\sum_{j=1}^d x_i^2\right)^k}\, dx \\
&= cN^{2k} \int_{\beta/2 \leq r \leq \sqrt{d}N} \frac{1}{r^{2k}} r^{d-1}\, dr
\end{aligned}
$$

In the second line, we used the fact that $\sin x \geq x/2$ for $x \in [0, \pi/2]$.
If $d = 2k$, then

$$I = cN^{2k} \log(N/\beta) = cn \log(2N/\beta).$$

If $2k < d$, then

$$I = cN^{2k} \frac{1}{d-2k} \left((N\sqrt{d})^{d-2k} - (\beta/2)^{d-2k} \right) . \leq cN^d.$$

If $2k > d$, then

$$I = cN^{2k} \frac{1}{2k-d} \left((\beta/2)^{d-2k} - (N\sqrt{d})^{d-2k} \right) .$$

Treating d, k as constants, we write

$$I \leq cN^{2k} \beta^{d-2k}.$$

□

A.10.3 Lemma A.11

This lemma provides a result analogous to Lemma A.10, by tying together the singular values of the KTF and GTF operators.

Lemma A.11. Let $\xi_1 \leq \ldots \leq \xi_{n-(k+1)^2}$ be the nonzero singular values of the d-dimensional KTF operator $\widetilde{\Delta}^{(k+1)}$ of order $k+1$. For any $i_0 \in [n - (k+1)^d - 1]$,

$$\sum_{i=i_0+1}^{n-(k+1)^d} \frac{1}{\xi_i^2} \leq c \begin{cases} n & 2(k+1) < d \\ n \log(n/i_0) & 2(k+1) = d \\ n(n/i_0)^{(2k+2-d)/d} & 2(k+1) > d \end{cases}$$

for large enough n, where $c > 0$ is a constant depending only on k.

Proof. Abbreviate $D = D_{1\mathrm{d}}^{(k+1)}$, and write G for the GTF operator of order $k+1$ defined over a 1d chain of length N. Also let $N' = N - k - 1$, and $k' = \lfloor (k+1)/2 \rfloor$. Then D is given by removing the first k_1 rows and last k_2 rows of G, i.e.,

$$D = PG, \quad \text{where } P = \left[\begin{array}{ccc} 0_{N' \times k'} & I_{N'} & 0_{N' \times k'} \end{array} \right].$$

This means

$$DD^T = PGG^TP^T.$$

Let β_i, $i \in [N']$ be the eigenvalues of DD^T, and let α_i, $i \in [N]$ be the eigenvalues of GG^T. The Cauchy interlacing theorem now tells us that

$$\beta_i \geq \alpha_i^{k+1}, \quad i \in [N']. \tag{A.40}$$

This key property will allow us to relate the nonzero singular values of the KTF operator to those of the GTF operator, more specifically, to the eigenvalues of the Laplacian of the 2d grid graph.

The squared nonzero singular values of $\widetilde{\Delta}^{(k+1)}$ are the nonzero eigenvalues of $(\widetilde{\Delta}^{(k+1)})^T\widetilde{\Delta}^{(k+1)}$. We can index the eigenvalues of $(\widetilde{\Delta}^{(k+1)})^T\widetilde{\Delta}^{(k+1)}$ by 2d grid positions, as in

$$\psi_{i_1,\ldots,i_d} = \sum_{j=1}^d \rho_{i_j}, \quad i_1,\ldots,i_d \in [N],$$

where ρ_i, $i \in [N]$ denote the eigenvalues of D^TD, i.e., $\rho_1 = \cdots = \rho_{k+1} = 0$ and $\rho_{i+k+1} = \beta_i$, $i \in [N']$, where D, β_i, $i \in [N']$ are as above. Also, as in the proof of Lemma A.10, we can write the eigenvalues of the Laplacian matrix of the d-dimensional grid graph as

$$\lambda_{i_1,\ldots,i_d} = \sum_{j=1}^d \alpha_{i_j}, \quad i_1,\ldots,i_d \in [N]$$

where α_i, $i \in [N]$ is as above. For arbitrary $i \in [N]^d$ $[k+1]^d$, note that

$$\frac{1}{\psi_{i_1,\ldots,i_d}} = \frac{1}{\sum_{j=1}^d \beta_{i_j-k-1}} \leq \frac{1}{\sum_{j=1}^d \alpha_{i_j-k-1}^{k+1}} \leq \frac{d^{k+1}}{\lambda_{i_1-k-1,\ldots,i_d-k-1}^{k+1}},$$

where we use the convention $\beta_{-i} = 0$ and $\alpha_{-i} = 0$ for $i \leq 0$, the first inequality was due to the key property (A.40), and the second was due to the simple fact $(\sum_{j=1}^d a_i)^k \leq d^k \sum_{j=1}^d a_i^k$ for $k \geq 1$. The last display shows that to bound the sum of squared reciprocal nonzero singular values of the KTF operator, it suffices to bound the reciprocal kth power of Laplacian eigenvalues, as was the case for the GTF operator. Proceeding as in the proof of Lemma A.10 gives the result. □

A.11 Incoherence lemmas for 1d difference operators

In this section, the first two lemmas establish incoherence of the left and right singular vectors of $D_{1\mathrm{d}}^{(k+1)}$. They rely heavily on approximation results for eigenvectors of symmetric banded Toeplitz matrices in Bogoya et al. (2016). The third lemma relates the eigenvectors of $(D_{1\mathrm{d}}^{(k+1)})(D_{1\mathrm{d}}^{(k+1)})^T$ to those of its elementwise absolute value matrix. This is critical for the proof of the first lemma, since, curiously, $(D_{1\mathrm{d}}^{(k+1)})(D_{1\mathrm{d}}^{(k+1)})^T$ falls outside of the scope of matrices considered in Bogoya et al. (2016) (as well as related papers on eigenvector approximations for Toeplitz matrices), but the elementwise absolute value matrix does not.

Lemma A.12. The left singular vectors u_i, $i \in [N-k-1]$ of $D_{1\mathrm{d}}^{(k+1)} \in \mathbb{R}^{N\times(N-k-1)}$ are incoherent, i.e., there exists a constant $\mu > 0$ depending only on k such that

$$\|u_i\|_\infty \le \frac{\mu}{\sqrt{N}}, \quad i \in [N-k-1],$$

for a constant $\mu > 0$ depending only on k.

Proof. For $k = 0$, the result has already been proved in Wang et al. (2016). Assume $k \ge 1$ henceforth. As in Lemma A.14 and its proof, abbreviate $D = D_{1\mathrm{d}}^{(k+1)}$, and $N' = N-k-1$. The left singular vectors of D are the eigenvectors of DD^T, which is a symmetric banded Toeplitz matrix with entries

$$(DD^T)_{ij} = c_{|i-j|}, \quad i,j \in [N'],$$
$$\text{where } c_\ell = (-1)^\ell \binom{2k+2}{k+1+\ell}, \quad \ell = 0,\dots,k+1.$$

Let $\beta_1 \le \ldots \le \beta_{N'}$ be the eigenvalues of DD^T. Observe that $\beta_{N'} \le 4^{k+1}$ by the Gershgorin circle theorem.

Unfortunately, the approximation results on eigenvectors of Toeplitz matrices from Bogoya et al. (2016) are not applicable to DD^T, because DD^T does not satisfy their simple-loop assumption. However, the Toeplitz matrix

$$T = 4^{k+1}I - \mathrm{abs}(DD^T),$$

where $\mathrm{abs}(A)$ denotes the elementwise absolute value of a matrix A, does satisfy the simple-loop assumption, and its eigenvectors are the same as those of DD^T up to elementwise sign flips, as we show in Lemma A.14. Thus, it suffices to verify the incoherence property for T, which we pursue in the following.

To be concrete, T is a symmetric banded Toeplitz matrix with entries

$$T_{ij} = a_{|i-j|}, \quad i,j \in [N'],$$
$$\text{where } a_\ell = 4^{k+1}\cdot 1\{\ell = 0\} - \binom{2k+2}{k+1+\ell}, \quad \ell = 0,\dots,k+1.$$

We introduce some notation. Let $\mathbb{C}$ denote the complex plane and $\mathbb{T}$ the unit circle in $\mathbb{C}$. The symbol of T is the function $a : \mathbb{T} \to \mathbb{C}$ is defined by

$$a(t) = \sum_{\ell=-(k+1)}^{k+1} a_\ell t^\ell = 4^{k+1} - \left(2 + t + \frac{1}{t}\right)^{k+1}.$$

We define the function $g : [0, 2\pi) \to \mathbb{R}$ by $g(\sigma) = a(e^{\iota\sigma}) = 4^{k+1} - (2 + 2\cos\sigma)^{k+1}$. (Here we use $\iota = \sqrt{-1}$ for the imaginary unit, to differentiate it from the the index variable i.)

It is straightforward to check that a, g as defined above satisfy what Bogoya et al. (2016) refer to as the "simple-loop" conditions: a is real-valued, the range of g is contained in the bounded set $[0, 4^{k+1}]$, g satisfies $g(0) = g(2\pi) = 0$, $g''(0) = g''(2\pi) > 0$, and g reaches its maximum of 4^{k+1} at $\pi \in [0, 2\pi)$. Hence, in the notation of Bogoya et al. (2016), we have $a \in SL^\alpha$ for any $\alpha \ge 4$.

For an eigenvalue τ, the characteristic polynomial of T is given by

$$p_\tau(t) = a(t) - \tau,$$

whose $2k+2$ are denoted by $z_0(\tau), z_1(\tau), z_2(\tau), \ldots, z_k(\tau)$ and their inverses. Following Bogoya et al. (2016), we use a labeling convention such that $|z_0(\tau)| = 1$, and $|z_\kappa(\tau)| > 1$ for $\kappa \in [k]$. We also define the function $b : \mathbb{T} \times [0, \pi] \to (0, \infty)$ by

$$b(t,s) = \frac{a(t) - g(s)}{2\cos s - (1 + 1/t)} = \frac{(2 + t + 1/t)^{k+1} - (2 + 2\cos\sigma)^{k+1}}{(2 + 2\cos s) - (2 + t + 1/t)}.$$

(Here we are using the simplified form of b in Corollary 2.2 of Bogoya et al. (2016), due to symmetry of g.) As b is a rational function (ratio of two polynomials) in (t, s), denoted

$$b(t,s) = \frac{P(t,s)}{Q(t,s)}$$

it has a Wiener-Hopf factorization $b(t,s) = b_-(t,s)b_+(t,s)$, where

$$b_+(t,s) = b_0(s)\frac{\prod_{i=1}^{p}(1 - t/\nu_i(s))}{\prod_{i=1}^{q}(1 - t/\zeta_i(s))}$$

for a constant $b_0(s)$, where $\nu_i(s)$, $i \in [p]$ and $\zeta_i(s)$, $i \in [q]$ denote the roots of $P(\cdot, s)$ and $Q(\cdot, s)$, respectively, with complex moduli larger at least 1. (The term $b_-(t,s)$ has a similar representation, but the specific details are unimportant for our purposes.)

Because $a(t) - g(s)$ is the characteristic polynomial $p_\tau(t)$ with $\tau = g(s)$, the roots $\nu_i(s)$, $i \in [p]$ of $P(\cdot, s)$ are simply $z_0(g(s))$, $z_\kappa(g(s))$, $\kappa \in [k]$; moreover, according to Chapter 1 in Bottcher & Grudsky (2005), the positive Wiener-Hopf factor $b_+(t,s)$ in the last display can be simplified to

$$b_+(t,s) = \prod_{\kappa=1}^{k}\Big(t - z_\kappa(g(s))\Big). \tag{A.41}$$

We are now ready to state the eigenvector approximation result. Write $\tau_i, \tilde{u}_i$ for a pair of eigenvalue and (unit norm) eigenvector of T, for $i \in [N']$. Combining Theorem 2.5, Theorem 4.1, and Lemma 4.2 in Bogoya et al. (2016), for each $i \in [N']$, we can represent $\tilde{u}_i = e_i/\|e_i\|_2$, where

$$e_i = M_i + L_i + R_i + \delta_i, \tag{A.42}$$

and for each $j \in [N']$,

$$M_{ij} = \frac{z_{0i}^{\frac{N'-1}{2}-j+1}}{|b_{+i}(z_{0i})|} + (-1)^{N'-i}\frac{\overline{z}_{0i}^{\frac{N'-1}{2}-j+1}}{|b_{+i}(\overline{z}_{0i})|}, \tag{A.43}$$

$$L_{ij} = \frac{z_{0i}^{\frac{N'+1}{2}}(z_{0i} - \overline{z}_{0i})b_{+i}(z_{0i})}{|b_{+i}(z_{0i})|}\sum_{\kappa=1}^{k}\frac{z_\kappa(\tau_i)^{-j}}{\frac{\partial b_{+i}}{\partial t}(z_\kappa(\tau_i))(z_\kappa(\tau_i) - z_{0i})(z_\kappa(\tau_i) - \overline{z}_{0i})} \tag{A.44}$$

$$R_{ij} = \bar{L}_{i,N'+1-j}. \tag{A.45}$$

Here, $\delta_{ij} = o(1/N')$, uniformly over i, j, and we use the abbreviations $b_{+i}(t) = b_+(t, s_i)$, where s_i is such that $g(s_i) = \tau_i$, and $z_{0i} = z_0(\tau_i)$, $i \in [N']$.

The details of the approximation in (A.42)–(A.45) are important for the next lemma, Lemma A.13, but are not needed presently. By the triangle inequality, for each $i, j \in [N']$,

$$\frac{|e_{ij}|}{\|e_i\|_2} \leq \frac{|M_{ij}|}{\|e_i\|_2} + \frac{|e_{ij} - M_{ij}|}{\|e_i\|_2} \leq \frac{1/|b_{+i}(z_{0i})| + 1/|b_{+i}(\bar{z}_{0i})|}{\|e_i\|_2} + \frac{|e_{ij} - M_{ij}|}{\|e_i\|_2}, \tag{A.46}$$

the second inequality following as $|z_{0i}| = 1$. Furthermore, by Theorem 2.6 and Lemma 4.2 in Bogoya et al. (2016), we know that for each $i \in [N']$,

$$\|e_i\|_2 = \sqrt{N'}\Big(b_{+i}(z_{0i})^{-2} + b_{+i}(\bar{z}_{0i})^{-2}\Big)^{1/2} + O(1), \quad \text{and} \tag{A.47}$$

$$\frac{\|e_i - M_i\|_2}{\|e_i\|_2} = O\bigg(\frac{1}{\sqrt{N'}}\bigg), \tag{A.48}$$

where the $O(1), O(1/\sqrt{N'})$ terms in the above are uniform over i. Their Theorem 2.6 also shows that $(b_{+i}(z_{0i})^{-2} + b_{+i}(\bar{z}_{0i})^{-2})^{1/2} \asymp 1$, uniformly over i. Noting the equivalence of ℓ_1 and ℓ_2 norms in $\mathbb{R}^2$, we also have that $|b_{+i}(z_{0i})| + |b_{+i}(\bar{z}_{0i})| \asymp 1$, uniformly over i, and therefore, combining this with (A.46)–(A.48), we conclude

$$|\tilde{u}_{ij}| = \frac{|e_{ij}|}{\|e_i\|_2} \leq O\bigg(\frac{1}{\sqrt{N'}}\bigg),$$

uniformly over $i, j \in [N']$, which completes the proof. □

Lemma A.13. The right singular vectors v_i, $i \in [N - k - 1]$ of $D_{1\mathrm{d}}^{(k+1)} \in \mathbb{R}^{N \times (N-k-1)}$ are incoherent, i.e., there exists a constant $\mu > 0$ depending only on k such that

$$\|v_i\|_\infty \leq \frac{\mu}{\sqrt{N}}, \quad i \in [N - k - 1],$$

for a constant $\mu > 0$ depending only on k.

Proof. As before, abbreviate $D = D_{1\mathrm{d}}^{(k+1)}$, and $N' = N - k - 1$. Denote by β_i, u_i, v_i a triplet of nonzero singular value, left singular vector, and right singular vector of D, for $i \in [N']$. Also denote by $\tilde{u}_i$, $i \in [N']$ the eigenvectors of $T = 4^{k+1}I - \mathrm{abs}(DD^T)$.

Note that by Lemma A.14 we have the relationship

$$u_i = S\tilde{u}_i, \quad i \in [N'], \tag{A.49}$$

between the left singular vectors of D and eigenvectors of T, where S is the alternating sign diagonal matrix (as defined in the proof of the lemma). Note also the relationship

$$\sqrt{\beta_i} v_i = D^T u_i, \quad i \in [N'], \tag{A.50}$$

between the right and left singular vectors of D. We will bound the absolute entries of v_i, $i \in [N']$ over the interior and boundary coordinates separately.

Bounding the interior elements. Using (A.49), (A.50), we can translate the expansion in (A.42) for $\tilde{u}_i = e_i/\|e_i\|_2$, $i \in [N']$ into one for v_i, $i \in [N']$. Write $w_i = D_{1i}$, $i \in [k+2]$

for the $(k+1)$st order forward difference coefficients. Fix an arbitrary $i \in [N']$ and interior coordinate $j \in \{k+2, \dots, N'\}$. We have, abbreviating $j' = j - k - 2$,

$$\begin{aligned}\sqrt{\beta_i} v_{ij} &= (-1)^{k+1} \sum_{\ell=1}^{k+2} w_\ell u_{i,j'+\ell} \\ &= \frac{(-1)^{k+1}}{\|e_i\|_2} \sum_{\ell=1}^{k+2} (-1)^{j'+\ell+1} w_\ell e_{i,j'+\ell} \\ &= \frac{(-1)^{j'+1}}{\|e_i\|_2} \sum_{\ell=1}^{k+2} |w_\ell| \big(M_{i,j'+\ell} + L_{i,j'+\ell} + R_{i,j'+\ell} + \delta_{i,j'+\ell} \big). \end{aligned} \tag{A.51}$$

We first work on the terms in the above sum involving $M_{i,j'+\ell}$, $\ell \in [k+2]$. Note that for $t \in \mathbb{C}$,

$$\sum_{\ell=1}^{k+2} |w_\ell| t^{\ell-1} = (1+t)^{k+1} = t^{(k+1)/2} q(t), \quad \text{where } q(t) = (2 + t + 1/t)^{(k+1)/2}. \tag{A.52}$$

Therefore, recalling (A.43), we have

$$\begin{aligned}\sum_{\ell=1}^{k+2} |w_\ell| M_{i,j'+\ell} &= \frac{z_{0i}^{\frac{N'-1}{2} - j'}}{|b_{+i}(z_{0i})|} \sum_{\ell=1}^{k+2} |w_\ell| z_{0i}^{-(\ell-1)} + (-1)^{N'-i} \frac{\bar{z}_{0i}^{\frac{N'-1}{2} - j'}}{|b_{+i}(\bar{z}_{0i})|} \sum_{\ell=1}^{k+2} |w_\ell| \bar{z}_{0i}^{-(\ell-1)} \\ &= \frac{z_{0i}^{\frac{N'-1}{2} - j'}}{|b_{+i}(z_{0i})|} z_{0i}^{-(k+1)/2} q(z_{0i}) + (-1)^{N'-i} \frac{\bar{z}_{0i}^{\frac{N'-1}{2} - j'}}{|b_{+i}(\bar{z}_{0i})|} \bar{z}_{0i}^{-(k+1)/2} q(\bar{z}_{0i}), \end{aligned} \tag{A.53}$$

where in the last line we have used the fact that $q(t) = q(1/t)$. Recall also that $z_{0i} = z_0(\tau_i)$, where τ_i denotes the ith eigenvalue of T, i.e., $\tau_i = 4^{k+1} - \beta_i$. By definition, z_{0i} is a unit-modulus root of the characteristic polynomial

$$p_{\tau_i}(t) = 4^{k+1} - (2 + t + 1/t)^{k+1} - \tau_i = \beta_i - (2 + t + 1/t)^{k+1}, \tag{A.54}$$

and therefore it holds that $q(z_{0i}) = q(\bar{z}_{0i}) = \sqrt{\beta_i}$. Continuing on from (A.53), we have

$$\sum_{\ell=1}^{k+2} |w_\ell| M_{i,j'+\ell} = \sqrt{\beta_i} \big(z_{0i}^{-(k+1)/2} + \bar{z}_{0i}^{-(k+1)/2} \big) M_{i,j'-1}. \tag{A.55}$$

Similar logic holds for the terms in (A.51) involving $L_{i,j'+\ell}, R_{i,j'+\ell}$, $\ell \in [k+2]$. First, we reexpress the definition in (A.44) as

$$L_{ij} = \sum_{\kappa=1}^{k} L_{i\kappa} z_\kappa(\tau_i)^{-j}.$$

Then, again applying (A.52), we have

$$\sum_{\ell=1}^{k+2} |w_\ell| L_{i,j'+\ell} = \sum_{\kappa=1}^{k} L_{i\kappa} z_\kappa(\tau_i)^{-1} q(z_\kappa(\tau_i)).$$

For each $\kappa \in [k]$, recall that $z_\kappa(\tau_i)$ is a root of the characteristic polynomial in (A.54) with modulus larger than 1, and hence $q(z_\kappa(\tau_i)) = \pm\sqrt{\beta_i}$. From the last display, this means we can write

$$\sum_{\ell=1}^{k+2} |w_\ell| L_{i,j'+\ell} = \sqrt{\beta_i} \sum_{\kappa=1}^{k} \sigma_{i\kappa} L_{i\kappa} z_\kappa(\tau_i)^{-j'-1}, \tag{A.56}$$

for signs $\sigma_{i\kappa} \in \{-1, 1\}$, $\kappa \in [k]$. Based on its definition in (A.45), we also have

$$\sum_{\ell=1}^{k+2} |w_\ell| R_{i,j'+\ell} = \sqrt{\beta_i} \overline{\left(\sum_{\kappa=1}^{k} \sigma_{i\kappa} L_{i\kappa} z_\kappa(\tau_i)^{-(N'-j')} \right)}. \tag{A.57}$$

Putting together (A.51), (A.55), (A.56), (A.57), and canceling out the common factor of $\sqrt{\beta_i}$, we have

$$v_{ij} = \frac{(-1)^{j'+1}}{\|e_i\|_2} \Bigg[\big(z_{0i}^{-(k+1)/2} + \bar{z}_{0i}^{-(k+1)/2} \big) M_{i,j'-1} + \sum_{\kappa=1}^{k} \sigma_{i\kappa} L_{i\kappa} z_\kappa(\tau_i)^{-j'-1} + \overline{\left(\sum_{\kappa=1}^{k} \sigma_{i\kappa} L_{i\kappa} z_\kappa(\tau_i)^{-(N'-j')} \right)} + \frac{\delta_{ij}}{\sqrt{\beta_i}} \Bigg].$$

Thus, using the fact that $|z_{0i}| = 1$ and $|z_\kappa(\tau_i)| > 1$, $\kappa \in [k]$,

$$|v_{ij}| \le \frac{2}{\|e_i\|_2} \left(|M_{i,j'-1}| + \sum_{\kappa=1}^{k} |L_{i\kappa}| + \frac{|\delta_{ij}|}{\sqrt{\beta_i}} \right).$$

It can be shown from the form of the positive Wiener-Hopf factor $b_+(t, s)$ in (A.41) that $L_{i\kappa} = O(1)$, $\kappa \in [k]$, uniformly in i. Furthermore, as already shown in the proof of Lemma A.12, we know that $|M_{ij}|/\|e_i\|_2 = O(1/\sqrt{N'})$ uniformly over i, j, and also $\|e_i\|_2 = \Omega(\sqrt{N'})$, uniformly over i. Lastly, $|\delta_{ij}|/\sqrt{\beta_i} \le (2/\pi)^{2k+2} |\delta_{ij}| N^{2k+2}$, where we have lower bounded the smallest singular value of D using (A.40) and the inequality $\sin(x) \ge x/2$ for small enough x. This does not pose any problems, because the remainder term δ_{ij} is actually smaller than any polynomial in N, uniformly over i, j, according to Theorem 2.5 of Bogoya et al. (2016). Therefore, combining all of this with the last display, we have $|v_{ij}| = O(1/\sqrt{N'})$, uniformly over i and interior coordinates j.

Bounding the boundary elements. Consider the "inverse" relationship to (A.50),

$$Dv_i = \sqrt{\beta_i} u_i, \quad i \in [N']. \tag{A.58}$$

Since $\beta_i \le 4^{k+1}$, $i \in [N']$, and the vectors u_i, $i \in [N']$ are incoherent from Lemma A.12, we have

$$\|Dv_i\|_\infty \le \frac{\mu}{\sqrt{N'}}, \quad i \in [N'],$$

for a constant $\mu > 0$ depending only on k, or more explicitly,

$$\left| \sum_{\ell=1}^{k+2} w_\ell v_{i,j+\ell-1} \right| \le \frac{\mu}{\sqrt{N'}}, \quad i, j \in [N'].$$

Fix an arbitrary $i \in [N']$, and consider $j = k+1$. By the above display, the triangle inequality, and the observation that $|w_1| = 1$,

$$|v_{i,k+1}| \le \frac{\mu}{\sqrt{N'}} + \sum_{\ell=2}^{k+2} |w_\ell||v_{i,k+\ell}| \le \frac{c}{\sqrt{N'}},$$

for a constant $c > 0$ depending only on k, where in the second inequality we used the incoherence of the right singular vectors over the interior elements, as shown previously. Continuing on in the same manner verifies the incoherence property at all positions $j = k, \dots, 1$, and similarly, at all positions $j = N'+1, \dots, N$. This completes the proof. □

Lemma A.14. Abbreviate $D = D_{1\mathrm{d}}^{(k+1)} \in \mathbb{R}^{(N-k-1)\times N}$, and use the notation $\mathrm{abs}(A)$ to denote the elementwise absolute value of a matrix A. Consider eigendecompositions

$$DD^T = U\Lambda U^T, \quad \mathrm{abs}(DD^T) = U_+^T \Lambda U^T.$$

Then:

(a) $\Lambda = \Lambda_+$;

(b) $\mathrm{abs}(U) = \mathrm{abs}(U_+)$.

Proof. Denote $N' = N - k - 1$. Let $S \in \mathbb{R}^{N'\times N'}$ be the alternating sign diagonal matrix with diagonal elements $1, -1, 1, -1, \dots$. Note that $S^{-1} = S^T = S$. From the relationship

$$DD^T = S^{-1}\mathrm{abs}(DD^T)S$$

we conclude that DD^T and $\mathrm{abs}(DD^T)$ are similar, i.e., $\Lambda = \Lambda_+$. From their eigendecompositions,

$$U\Lambda U^T = SU_+\Lambda U_+^T S^T$$

we also see that $U = SU_+$ which implies $\mathrm{abs}(U) = \mathrm{abs}(U_+)$. □

A.11.1 Proof of Lemma 2.3

Denote

$$\widetilde{Z}_d = \big\{x = (x_1, \dots, x_d) \in Z_d : x_j \le 1 - (k+1)/N, j = 1, \dots, d\big\}.$$

Pick an arbitrary $\theta \in \mathcal{H}_d^{k+1}(L)$, corresponding to discretizations of $f \in H(k+1, L; [0,1]^d)$. The bound (2.16) holds at any $x \in \widetilde{Z}_d$, and the fact that $\delta(N) \le cL/N$ is verified by Lemma A.15. The KTF penalty is then

$$\|\widetilde{\Delta}^{(k+1)}\theta\|_1 = \sum_{x\in\widetilde{Z}_d} \big|\big(D_{x_j^{k+1}}\theta\big)(x)\big| \le cnLN^{-k-1} = cLn^{1-(k+1)/d},$$

recalling $N = n^{1/d}$.

A.11.2 Lemma A.15

The following lemma follows standard calculations in numerical analysis, e.g., as in Strikwerda (2004).

Lemma A.15. Let $f \in H(k+1, L; [0,1]^d)$. The kth order forward discrete difference along a unit direction $v \in \mathbb{R}^d$, with step size $h > 0$, obeys at any point $x \in [0,1]^d$,

$$\left| \frac{1}{h^k} (D_{v^k} \theta)(x) - \frac{\partial^k}{\partial v^k} f(x) \right| \leq cLh,$$

where $c > 0$ is a constant depending only on k, provided that $x + khv \in [0,1]^d$ (so that the discrete approximation is well-defined).

Proof. By Taylor expanding f around x at $x, x+hv, x+2hv, \ldots, x+khv$, we have

$$\begin{aligned}
f(x) &= f(x), \\
f(x+hv) &= f(x) + \frac{\partial}{\partial v} f(x) h + \frac{1}{2} \frac{\partial^2}{\partial v^2} f(x) h^2 + \ldots + \frac{1}{k!} \frac{\partial^k}{\partial v^k} f(x) h^k + r(h), \\
f(x+2hv) &= f(x) + \frac{\partial}{\partial v} f(x)(2h) + \frac{1}{2} \frac{\partial^2}{\partial v^2} f(x)(2h)^2 + \ldots + \frac{1}{k!} \frac{\partial^k}{\partial v^k} f(x)(2h)^k + r(2h), \\
&\vdots \\
f(x+khv) &= f(x) + \frac{\partial}{\partial v} f(x)(kh) + \frac{1}{2} \frac{\partial^2}{\partial v^2} f(x)(kh)^2 + \ldots + \frac{1}{k!} \frac{\partial^k}{\partial v^k} f(x)(kh)^k + r(kh),
\end{aligned}$$

where $r(ih)$ is integral form of the remainder in the expansion for $x + ihv$, satisfying

$$|r(ih)| = \left| \frac{1}{k!} \int_0^{ih} \frac{\partial^{k+1}}{\partial v^{k+1}} f(x+tv) t^k \, dt \right| \leq \frac{k^{k+1}}{(k+1)!} L h^{k+1}, \quad i = 1, \ldots, k.$$

(Note that such integrals are well-defined since Lipschitz continuity of $\partial^k f / \partial v^k$ implies that the $(k+1)$st derivative $\partial^{k+1} f / \partial v^{k+1}$ exists almost everywhere and is Lebesgue integrable, by Rademacher's theorem.) In the inequality above, we invoked the Holder property, recalling $f \in H(k+1, L; [0,1]^d)$.

Now denote the kth order forward difference coefficients by

$$w_i = (-1)^{k+i-1} \binom{k}{i-1}, \quad i = 1, \ldots, k+1.$$

Inverting the above $(k+1) \times (k+1)$ system of equations (from the $k+1$ Taylor expansions), and inspecting the last equality in the inverted system, gives

$$\frac{\partial^k}{\partial v^k} f(x) h^k = \sum_{i=1}^{k+1} w_i \Big(f(x + (i-1)hv) - r((i-1)h) \Big) = (D_{v^k} \theta)(x) - \sum_{i=1}^{k+1} w_i r((i-1)h).$$

Using our previous bound on the magnitude of remainders, we see

$$\left| (D_{v^k} \theta)(x) - \frac{\partial^k}{\partial v^k} f(x) h^k \right| \leq \frac{k^{k+1}}{(k+1)!} \sum_{i=1}^{k+1} |w_i| L h^{k+1},$$

and dividing through by h^k gives the claimed result. □

A.11.3 Proof of Lemma 2.4

We need only to construct a single counterexample for each $k, d \geq 1$. We give such a construction for $d = 2$ and $k = 1$; all other cases follows similarly. Consider a function $f : [0,1]^d \to \mathbb{R}$ defined by $f(x) = Mx_1 + x_2$, and let $\theta \in \mathbb{R}^n$ contain the evaluations of f over the grid Z_2. As f is linear, it is clearly an element of $H(2, 1; [0,1]^2)$. But, for any x on the left boundary of Z_2,

$$\|\Delta^{(2)}\theta\|_1 \geq \left| f\left(x + \frac{e_1}{N}\right) + f\left(x - \frac{e_2}{N}\right) + f\left(x + \frac{e_2}{N}\right) - 3f(x) \right| = \left| f\left(x + \frac{e_1}{N}\right) - f(x) \right| = Mn^{1/2},$$

Since M can be arbitrary, this proves the result.

A.12 Additional experiments comparing TV denoising and Laplacian smoothing for piecewise constant functions

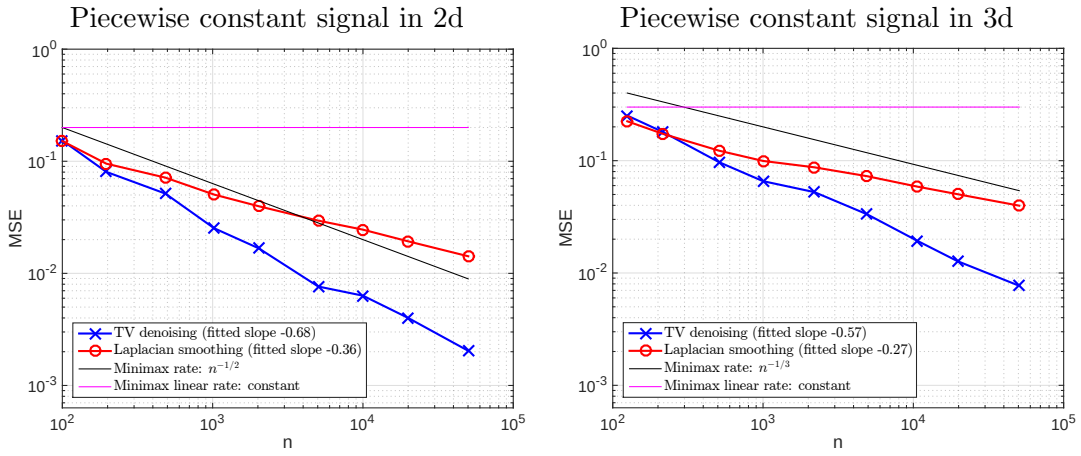

Figure A.2: *MSE curves for estimating a "piecewise constant" signal, having a single elevated region, over 2d and 3d grids. For each n, the results were averaged over 5 repetitions, and the Laplacian smoothing and TV denoising estimators were tuned for best average MSE performance. We set θ_0 to satisfy $\|D\theta_0\|_1 \asymp n^{1-1/d}$, matching the canonical scaling. Note that all estimators achieve better performance than that dictated by their minimax rates.*

Appendix B

Appendix for Additive models with Trend Filtering

B.1 Fast extrapolation

We discuss extrapolation using the fitted functions $\hat{f}_j$, $j = 1, \dots, d$ from additive trend filtering (3.6), as in (3.9). We must compute the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$ whose block form is given in (3.10), (3.11). Clearly, the computation of $\hat{b}_j$ in (3.11) requires $O(n)$ operations (owing to the bandedness of $D^{(X_j,k+1)}$, and treating k as a constant). As for $\hat{a}_j$ in (3.10), it can be seen from the structure of $C^{(X_j,k+1)}$ as described in Wang et al. (2014) that

$$
\begin{gathered}
(\hat{a}_j)_1 = (S_j \hat{\theta}_j)_1, \\
(\hat{a}_j)_\ell = \frac{1}{(\ell-1)!} \left[\operatorname{diag}\left(\frac{1}{X_j^\ell - X_j^1}, \dots, \frac{1}{X_j^n - X_j^{n-\ell+1}} \right) D^{(X_j,\ell-1)} S_j \hat{\theta}_j \right]_1, \quad \ell = 2, \dots, k+1,
\end{gathered}
$$

which takes only $O(1)$ operations (again treating k as constant, and now using the bandedness of each $D_j^{(X_j,\ell-1)}$, $\ell = 2, \dots, k+1$). In total then, computing the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$ requires $O(n)$ operations, and computing $\hat{\alpha} = (\hat{\alpha}_1, \dots, \alpha_d)$ requires $O(nd)$ operations.

After having computed $\hat{\alpha} = (\hat{\alpha}_1, \dots, \alpha_d)$, which only needs to be done once, a prediction at a new point $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ with the additive trend filtering fit $\hat{f}$ is given by

$$
\hat{f}(x) = \bar{Y} + \sum_{j=1}^{d} \sum_{\ell=1}^{n} \hat{\alpha}_j^\ell h_\ell^{(X_j)}(x_j),
$$

This requires $O(d + \sum_{j=1}^{d} \sum_{\ell=k+2}^{n} 1\{\hat{\alpha}_j^\ell \neq 0\})$ operations, utilizing the sparsity of the components in $\hat{\alpha}$ not associated with the polynomial basis functions.

B.2 Proof of Lemma 3.2

We begin by eliminating the constraint in the additive trend filtering problem (3.4), rewriting it as

$$\min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\theta_j\Big\|_2^2 + \lambda \sum_{j=1}^d \big\|D^{(X_j,k+1)} S_j M\theta_j\big\|_1,$$

where $M = I - \mathbb{1}\mathbb{1}^T/n$. Noting that $D^{(X_j,k+1)}\mathbb{1} = 0$ for $j = 1,\ldots,d$, we can replace the penalty term above by $\sum_{j=1}^d \|D^{(X_j,k+1)} S_j \theta_j\|_1$. Reparametrizing using the falling factorial basis, as in Lemma 3.1, yields the problem

$$\min_{a \in \mathbb{R}^{(k+1)d},\, b \in \mathbb{R}^{(n-k-1)d}} \frac{1}{2}\Big\|MY - M\sum_{j=1}^d P_j a_j - M\sum_{j=1}^d K_j b_j\Big\|_2^2 + \lambda k! \sum_{j=1}^d \|b_j\|_1,$$

where we have used the abbreviation $P_j = P^{(X_j,k)}$ and $K_j = K^{(X_j,k)}$, as well as the block representation $\alpha_j = (a_j, b_j) \in \mathbb{R}^{(k+1)} \times \mathbb{R}^{(n-k-1)}$, for $j = 1,\ldots,d$. Since each P_j, $j = 1,\ldots d$ has $\mathbb{1}$ for its first column, the above problem is equivalent to

$$\min_{a \in \mathbb{R}^{kd},\, b \in \mathbb{R}^{(n-k-1)d}} \frac{1}{2}\Big\|MY - M\sum_{j=1}^d \tilde{P}_j a_j - M\sum_{j=1}^d K_j b_j\Big\|_2^2 + \lambda k! \sum_{j=1}^d \|b_j\|_1,$$

where $\tilde{P}_j$ denotes P_j with the first column removed, for $j = 1,\ldots,d$. To be clear, solutions in the above problem and the original trend filtering formulation (3.4) are related by

$$\hat{\theta}_j = \tilde{P}_j \hat{a}_j + K_j \hat{b}_j, \quad j = 1,\ldots,d.$$

Furthermore, we can see that $\hat{a} = (\hat{a}_1, \ldots \hat{a}_d)$ solves

$$\min_{a \in \mathbb{R}^{kd}} \frac{1}{2}\Big\|\Big(MY - M\sum_{j=1}^d K_j \hat{b}_j\Big) - \tilde{P}a\Big\|_2^2, \tag{B.1}$$

where $\tilde{P}$ is as defined in (3.12), and $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_d)$ solves

$$\min_{b \in \mathbb{R}^{(n-k-1)d}} \frac{1}{2}\Big\|UU^T MY - UU^T M\sum_{j=1}^d K_j b_j\Big\|_2^2 + \lambda k! \|b\|_1,$$

where UU^T is the projection orthogonal to the column space of $\tilde{P}$, i.e., it solves

$$\min_{b \in \mathbb{R}^{(n-k-1)d}} \frac{1}{2}\big\|U^T MY - \tilde{K}b\big\|_2^2 + \lambda k! \|b\|_1, \tag{B.2}$$

where $\tilde{K}$ is as in (3.13). Since problem (B.2) is a standard lasso problem, existing results on the lasso (e.g., Tibshirani (2013)) imply that the solution $\hat{b}$ is unique whenever $\tilde{K}$ has columns in general position. This proves the first part of the lemma. For the second part of the lemma, note that the solution $\hat{a}$ in the least squares problem (B.1) is just given by the regression of $MY - M\sum_{j=1}^d K_j \hat{b}_j$ onto $\tilde{P}$, which is unique whenever $\tilde{P}$ has full column rank. This completes the proof.

B.3 Derivation of additive trend filtering dual

As in the proof of Lemma 3.2, we begin by rewriting the problem (3.4) as

$$\min_{\theta_1,\ldots,\theta_d\in\mathbb{R}^n} \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\theta_j\Big\|_2^2 + \lambda\sum_{j=1}^d \|D_jS_jM\theta_j\|_1,$$

where $M = I - \mathbb{1}\mathbb{1}^T/n$. Then, we reparametrize the above problem,

$$\begin{aligned}\min_{\substack{\theta_1,\ldots,\theta_d\in\mathbb{R}^n\\ w\in\mathbb{R}^n,\, z\in\mathbb{R}^{md}}} \quad & \frac{1}{2}\|MY - w\|_2^2 + \lambda\sum_{j=1}^d \|z_j\|_1\\ \text{subject to} \quad & w = \sum_{j=1}^d M\theta_j, \quad z_j = D_jS_jM\theta_j, \quad j = 1,\ldots,d,\end{aligned}$$

and form the Lagrangian

$$L(\theta, w, z, u, v) = \frac{1}{2}\|MY - w\|_2^2 + \lambda\sum_{j=1}^d \|z_j\|_1 + u^T\Big(w - \sum_{j=1}^d M\theta_j\Big) + \sum_{j=1}^d v_j^T(D_jS_jM\theta_j - z_j).$$

Minimizing the Lagrangian L over all θ, z yields the dual problem

$$\begin{aligned}\max_{\substack{u\in\mathbb{R}^n\\ v_1,\ldots,v_d\in\mathbb{R}^m}} \quad & \frac{1}{2}\|MY\|_2^2 - \frac{1}{2}\|MY - u\|_2^2\\ \text{subject to} \quad & u = S_jD_j^Tv_j, \quad \|v_j\|_\infty \le \lambda, \quad j = 1,\ldots,d.\end{aligned}$$

The claimed dual problem (3.14) is just the above, rewritten in an equivalent form.

B.4 Proof of Lemma 3.3

We first eliminate the equality constraint in (3.4), rewriting this problem, as was done in the proof of Lemma 3.2, as

$$\min_{\theta_1,\ldots,\theta_d\in\mathbb{R}^d} \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\theta_j\Big\|_2^2 + \lambda\sum_{j=1}^d \|D_jS_j\theta_j\|_1,$$

where $M = I - \mathbb{1}\mathbb{1}^T/n$, and $D_j = D^{(X_j,k+1)}$, $j = 1,\ldots,d$. This is a generalized lasso problem with a design matrix $T \in \mathbb{R}^{n\times nd}$ that has d copies of M stacked along its columns, and a penalty matrix $D \in \mathbb{R}^{nd\times nd}$ that is block diagonal in the blocks D_j, $j = 1,\ldots,d$. Applying Theorem 3 of Tibshirani & Taylor (2012), we see that

$$\mathrm{df}(T\hat{\theta}) = \mathbb{E}\big[\dim\big(T\mathrm{null}(D_{-A})\big)\big],$$

where $A = \mathrm{supp}(D\hat{\theta})$, and where D_{-A} denotes the matrix D with rows removed that correspond to the set A. The conditions for uniqueness in the lemma now precisely imply that

$$\dim\big(T\mathrm{null}(D_{-A})\big) = \left(\sum_{j=1}^{d} |A_j|\right) + kd,$$

where A_j denotes the subset of A corresponding to the block of rows occupied by D_j, and $|A_j|$ its cardinality, for $j = 1, \ldots, d$. This can be verified by transforming to the basis perspective as utilized in the proofs of Lemmas 3.1 and 3.2. The desired result is obtained by noting that, for $j = 1, \ldots, d$, the component $\hat{\theta}_j$ exhibits a knot for each element in A_j.

B.5 Preliminaries for the proof of Theorem 3.1

Before the proof of Theorem 3.1, we collect important preliminary results. We start with a result on orthonormal polynomials. We thank Dejan Slepcev for his help with the next lemma.

Lemma B.1. Given an integer $\kappa \geq 0$, and a continuous measure Λ on $[0, 1]$, whose Radon-Nikodym derivative λ is bounded below and above by constants $b_1, b_2 > 0$, respectively. Denote by $\phi_0, \ldots, \phi_\kappa$ an orthonormal basis for the space of polynomials of degree κ on $[0, 1]$, given by running the Gram-Schmidt procedure on the polynomials $1, t, \ldots, t^\kappa$, with respect to the $L_2(\Lambda)$ inner product. Hence, for $\ell = 0, \ldots, \kappa$, ϕ_ℓ is an ℓth degree polynomial, orthogonal (in $L_2(\Lambda)$) to all polynomials of degree less than ℓ, and we denote its leading coefficient by $a_\ell > 0$. Now define, for $t \in [0, 1]$:

$$\begin{aligned} \Phi_{\kappa,0}(t) &= \phi_\kappa(t)\lambda(t), \\ \Phi_{\kappa,\ell+1}(t) &= \int_0^t \Phi_{\kappa,\ell}(u)\, du, \quad \ell = 0, \ldots, \kappa. \end{aligned}$$

Then the following two relations hold:

$$\Phi_{\kappa,\ell}(1) = \begin{cases} 0 & \text{for } \ell = 1, \ldots, \kappa, \\ \frac{(-1)^\kappa}{a_\kappa \kappa!} & \text{for } \ell = \kappa + 1, \end{cases} \tag{B.3}$$

and

$$a_\kappa \kappa! |\Phi_{\kappa,\kappa}(t)| \leq \binom{2\kappa}{\kappa} \sqrt{\frac{b_2}{b_1}}, \quad t \in [0, 1]. \tag{B.4}$$

Proof. First, we use induction to show that for $t \in [0, 1]$,

$$\Phi_{\kappa,\ell}(t) = \int_0^t \phi_\kappa(u) \frac{(t-u)^{\ell-1}}{(\ell-1)!} \lambda(u)\, du, \quad \ell = 1, \ldots, \kappa + 1. \tag{B.5}$$

This statement holds for $\ell = 1$ by definition of $\Phi_{\kappa,0}, \Phi_{\kappa,1}$. Assume it holds at some $\ell > 1$. Then

$$\Phi_{\kappa,\ell+1}(t) = \int_0^t \int_0^u \phi_\kappa(v) \frac{(u-v)^{\ell-1}}{(\ell-1)!} \lambda(v)\, dv\, du$$

$$= \int_0^t \phi_\kappa(v) \Big(\int_v^t \frac{(u-v)^{\ell-1}}{(\ell-1)!}\, du \Big) \lambda(v)\, dv$$
$$= \int_0^t \phi_\kappa(v) \frac{(t-v)^\ell}{\ell!} \lambda(v)\, dv,$$

where we used inductive hypothesis in the first line and Fubini's theorem in the second line, which completes the inductive proof.

Now, the relation in (B.5) shows that $\Phi_{\kappa,\ell}(1)$ is the $L_2(\Lambda)$ inner product of ϕ_κ and an $(\ell-1)$st degree polynomial, for $\ell = 1, \ldots, \kappa$. As ϕ_κ is orthogonal to all polynomials of degree less than κ, we have $\Phi_{\kappa,\ell}(1) = 0$, $\ell = 1, \ldots, \kappa$. For $\ell = \kappa + 1$, note that this same orthogonality along with (B.5) also shows

$$\Phi_{\kappa,\kappa+1}(1) = \Big\langle \phi_\kappa, \frac{(-1)^\kappa}{a_\kappa \kappa!} \phi_\kappa \Big\rangle_2 = \frac{(-1)^\kappa}{a_\kappa \kappa!}.$$

where $\langle \cdot, \cdot \rangle_2$ is the $L_2(\Lambda)$ inner product. This establishes the statement in (B.3).

As for (B.4), note that if $\kappa = 0$ then the statement holds, because the left-hand side is 1 and the right-hand side is always larger than 1. Hence consider $\kappa \geq 1$. From (B.5), we have, for any $t \in [0, 1]$,

$$\begin{aligned} |\Phi_{\kappa,\kappa}(t)| &\leq \int_0^t |\phi_\kappa(u)| \frac{(t-u)^{\kappa-1}}{(\kappa-1)!} \lambda(u)\, du \\ &\leq \Big(\int_0^t \phi_\kappa^2(u) \lambda(u)\, du \Big)^{1/2} \Big(\int_0^t \frac{(t-u)^{2\kappa-2}}{(\kappa-1)!^2} \lambda(u)\, du \Big)^{1/2} \\ &\leq \frac{\sqrt{b_2}}{(\kappa-1)! \sqrt{2\kappa-1}}, \end{aligned} \tag{B.6}$$

where in the second line we used Cauchy-Schwartz, and in the third line we used the fact that ϕ_κ has unit norm, and the upper bound b_2 on λ. Next we bound a_κ. Let p be the projection of x^κ onto the space of polynomials of degree $\kappa - 1$, with respect to the $L_2(\Lambda)$ inner product. Then we have $\phi_\kappa = (x^\kappa - p)/\|x^\kappa - p\|_2$, thus its leading coefficient is $a_\kappa = 1/\|x^\kappa - p\|_2$, where $\|\cdot\|_2$ is the $L_2(\Lambda)$ norm. Consider

$$\begin{aligned} \|x^\kappa - p\|_2 &\geq \sqrt{b_1} \Big(\int_0^1 (x^\kappa - p)^2(t)\, dt \Big)^{1/2} \\ &\geq \sqrt{b_1} \Big(\int_0^1 P_\kappa^2(t)\, dt \Big)^{1/2} = \frac{\sqrt{b_1}}{\sqrt{2\kappa+1} \binom{2\kappa}{\kappa}}. \end{aligned} \tag{B.7}$$

In the first line we used the lower bound b_1 on λ. In the second we used the fact the Legendre polynomial P_κ of degree κ, shifted to $[0, 1]$ but unnormalized, is the result from projecting out $1, t, \ldots, t^{\kappa-1}$ from t^κ, with respect to the uniform measure. In the last step we used the fact that P_κ has norm $1/(\sqrt{2\kappa+1}\binom{2\kappa}{\kappa})$. Combining (B.6) and (B.7) gives the result (B.4). □

Remark B.1 (**Special case: uniform measure and Rodrigues' formula**)**.** In the case of $\Lambda = U$, the uniform measure on $[0, 1]$, we can just take $\phi_0, \ldots, \phi_\kappa$ to be the Legendre

polynomials, shifted to $[0,1]$ and normalized appropriately. Invoking Rodrigues' formula to express these functions,

$$\phi_\ell(t) = \frac{\sqrt{2\ell+1}}{\ell!}\frac{d^\ell}{dt^\ell}(t^2-t)^\ell, \quad \ell = 0, \dots \kappa,$$

the results in Lemma B.1 can be directly verified.

We use Lemma B.1 to construct a sup norm bound on functions in $B_J(1)$ that are orthogonal (in $L_2(\Lambda)$) to all polynomials of degree k. We again Dejan Slepcev for his help with the next lemma.

Lemma B.2. Given an integer $k \geq 0$, and a continuous measure Λ on $[0,1]$, whose Radon-Nikodym derivative λ is bounded below and above by constants $b_1, b_2 > 0$, respectively. Let J be a functional satisfying Assumptions C1 and C2, for a constant $L > 0$. There is a constant $R_0 > 0$, that depends only on k, b_1, b_2, L, such that

$$\|g\|_\infty \leq R_0, \quad \text{for all } g \in B_J(1), \text{ such that } \langle g, p\rangle_2 = 0 \text{ for all polynomials } p \text{ of degree } k,$$

where $\langle \cdot, \cdot \rangle_2$ denotes the $L_2(\Lambda)$ inner product.

Proof. Fix an arbitrary function $g \in B_J(1)$, orthogonal (in $L_2(\Lambda)$) to all polynomials of degree k. Using integration by parts, and repeated application of Lemma B.1, we have

$$0 = a_\ell \ell! \cdot \langle g, \phi_\ell \rangle_2 = \int_0^1 g^{(\ell)}(t) w_\ell(t)\, dt, \quad \ell = 0, \dots, k, \tag{B.8}$$

where $w_\ell(t) = (-1)^\ell a_\ell \ell! \Phi_{\ell,\ell}(t)$, $\ell = 0, \dots, k$, and by properties (B.3), (B.4) of Lemma B.1,

$$\int_0^1 w_\ell(t)\, dt = 1, \quad \int_0^1 |w_\ell(t)|\, dt \leq \binom{2\ell}{\ell}\sqrt{\frac{b_2}{b_1}}, \quad \ell = 0, \dots, k. \tag{B.9}$$

Now, we will prove the following by induction:

$$\|g^{(\ell)}\|_\infty \leq L\left(\frac{b_2}{b_1}\right)^{(k-\ell+1)/2} \prod_{i=\ell}^{k} \binom{2i}{i}, \quad \ell = 0, \dots, k. \tag{B.10}$$

Starting at $\ell = k$, the statement holds because, using (B.8), for almost every $t \in [0,1]$,

$$\begin{aligned} |g^{(k)}(t)| &= \left| g^{(k)}(t) - \int_0^1 g^{(k)}(u) w_k(u)\, du \right| \\ &= \left| \int_0^1 \Big(g^{(k)}(t) - g^{(k)}(u) \Big) w_k(u)\, du \right| \\ &\leq L\binom{2k}{k}\sqrt{\frac{b_2}{b_1}}, \end{aligned}$$

where in the second line we used the fact that the weight function integrates to 1 from (B.9), and in the third we used Assumption C2 and the upper bound on the integrated

absolute weights from (B.9). Assume the statement holds at some $\ell < k$. Then again by (B.8), (B.9), for almost every $t \in [0,1]$,

$$\begin{aligned}
|g^{(\ell-1)}(t)| &= \left| \int_0^1 \Big(g^{(\ell-1)}(t) - g^{(\ell-1)}(u) \Big) w_{\ell-1}(u)\, du \right| \\
&\le \left(\operatorname*{ess\,sup}_{0 \le u < v \le 1} |g^{(\ell-1)}(v) - g^{(\ell-1)}(u)| \right) \binom{2\ell-2}{\ell-1} \sqrt{\frac{b_2}{b_1}} \\
&= \left(\operatorname*{ess\,sup}_{0 \le u < v \le 1} \left| \int_u^v g^{(\ell)}(s)\, ds \right| \right) \binom{2\ell-2}{\ell-1} \sqrt{\frac{b_2}{b_1}} \\
&\le L \left(\frac{b_2}{b_1} \right)^{(k-\ell+2)/2} \prod_{i=\ell-1}^{k} \binom{2i}{i},
\end{aligned}$$

the last line using ess $\sup_{0 \le u < v \le 1} |\int_u^v g^{(\ell)}(s)\, ds| \le \|g^{(\ell)}\|_\infty$ and the inductive hypothesis. This verifies (B.10). Taking $\ell = 0$ in (B.10) and defining $R_0 = L(b_2/b_1)^{(k+1)/2} \prod_{i=0}^{k} \binom{2i}{i}$ proves the lemma. □

We study the minimum eigenvalue of the (uncentered) empirical covariance matrix of a certain basis for additive kth degree polynomials in $\mathbb{R}^{kd}$. We thank Mathias Drton for his help with part (a) of the next lemma.

Lemma B.3. Let X^i, $i = 1, \dots, n$ denote an i.i.d. sample from a continuous distribution Q on $[0,1]^d$. For an integer $k \ge 0$, let $V \in \mathbb{R}^{n \times kd}$ be a matrix whose ith row is given by

$$V^i = \Big(X_1^i, (X_1^i)^2, \dots, (X_1^i)^k, \dots, X_d^i, (X_d^i)^2, \dots, (X_d^i)^k \Big) \in \mathbb{R}^{kd}, \tag{B.11}$$

for $i = 1, \dots, n$. Let

$$\nu_n^2 = \lambda_{\min}\Big(\frac{1}{n} V^T V \Big), \quad \text{and} \quad \nu_0^2 = \lambda_{\min}\Big(\frac{1}{n} \mathbb{E}[V^T V] \Big),$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of its argument. Assuming that $n \ge kd$, the following properties hold:

(a) $\nu_n > 0$, almost surely with respect to Q;

(b) $\nu_0 > 0$;

(c) for any $0 \le t \le 1$, $\mathbb{P}(\nu_n^2 > t\nu_0^2)$ with probability at least $1 - d \exp\Big(- \dfrac{(1-t)^2 \nu_0 n}{2(kd)^2} \Big)$.

Proof. For part (a), if the claim holds for $n = kd$, then it holds for all $n > kd$, so we may assume without a loss of generality that $n = kd$. Note that the determinant of $V \in \mathbb{R}^{n \times kd}$ is a polynomial function, call it $q(X)$, of the elements X_j^i, $i = 1, \dots, n$, $j = 1, \dots, d$. By Lemma 1 of Okamoto (1973), the roots of any polynomial—that is not identically zero—form a set of Lebesgue measure zero. To check that the polynomial q in question is not

identically zero, it suffices to show that it is nonzero at a single realization of X. To this end, consider an input matrix defined by

$$X = \begin{bmatrix} \alpha_1 I \\ \vdots \\ \alpha_k I \end{bmatrix} \in \mathbb{R}^{n \times kd},$$

the rowwise concatenation of $\alpha_\ell I \in \mathbb{R}^{d \times d}$, $\ell = 1, \ldots, k$. By the blockwise Vandermonde structure of the corresponding basis matrix V, we have that $q(X) \neq 0$ provided the coefficients α_ℓ, $\ell = 1, \ldots, k$ are all distinct. Therefore q is not identically zero, and with respect to the continuous distribution Q, the determinant of V is nonzero, i.e., $\nu_n > 0$, almost surely.

For part (b), given any $a \in \mathbb{R}^{kd}$ with $a \neq 0$, we know that $Va \neq 0$ almost surely, since $\nu_n > 0$ almost surely, by part (a). Thus

$$a^T \mathbb{E}[V^T V] a = \mathbb{E}\|Va\|_2^2 > 0,$$

which proves that $\nu_0 > 0$.

Part (c) is an application of a matrix Chernoff bound from Tropp (2012). In order to apply this result, we must obtain an almost sure upper bound R on $\lambda_{\max}(V^i(V^i)^T)$, with V^i as in (B.11) and $\lambda_{\max}(\cdot)$ denoting the maximum eigenvalue of its argument. This follows as

$$\lambda_{\max}\big(V^i(V^i)^T\big) \leq \sum_{j=1}^{kd} \sum_{\ell=1}^{kd} (V_j^i V_\ell^i)^2 \leq (kd)^2,$$

as each component of V^i has absolute magnitude at most 1 (recalling that Q is supported on $[0,1]^d$). Taking $R = (kd)^2$ and applying Corollary 5.2 of Tropp (2012) (to be specific, applying the form of the Chernoff bound given in Remark 5.3 of this paper) gives the result. □

The next lemma pertains to the additive function space

$$\mathcal{M}_n(\delta) = \left\{ \sum_{i=1}^d m_j : \sum_{j=1}^d J(m_j) \leq \delta, \text{ and } \langle m_j, 1 \rangle_n = 0, \ j = 1, \ldots, d \right\}. \tag{B.12}$$

We give a sup norm bound on the components of functions in $\mathcal{M}_n(1) \cap B_n(\rho)$. The proof combines Lemmas B.2 and B.3, and uses a general strategy that follows the arguments given in Example 2.1(ii) of van de Geer (1990).

Lemma B.4. Let X^i, $i = 1, \ldots, n$ denote an i.i.d. sample from a continuous distribution Q on $[0,1]^d$, and let J be a seminorm satisfying Assumptions C1 and C2. There are constants $R_1, R_2, c_0, n_0 > 0$, depending only on d, k, L, such that for all $\rho > 0$ and $n \geq n_0$,

$$\|m_j\|_\infty \leq R_1 \rho + R_2, \quad \text{for all } j = 1, \ldots, d \text{ and } \sum_{j=1}^d m_j \in \mathcal{M}_n(1) \cap B_n(\rho),$$

with probabilty at least $1 - \exp(-c_0 n)$, where $\mathcal{M}_n(1)$ is the function space in (B.12).

Proof. Fix an arbitrary $m = \sum_{j=1}^d m_j \in \mathcal{M}_n(1) \cap B_n(\rho)$. For each $j = 1, \ldots, d$, decompose

$$m_j = \langle m_j, 1\rangle_n + p_j + g_j,$$

where p_j is a polynomial of degree k such that $\langle p_j, 1\rangle_n = 0$, and g_j is orthogonal to all polynomials of degree k, with respect to the $L_2(U)$ inner product, with U the uniform distribution on $[0,1]$; in fact, by definition of $\mathcal{M}_n(1)$, we know that $\langle m_j, 1\rangle_n = 0$ so

$$m_j = p_j + g_j.$$

By the triangle inequality and Lemma B.2 applied to the measure $\Lambda = U$ (whose density is of course lower and upper bounded with $b_1 = b_2 = 1$), we have, for each $j = 1, \ldots, d$,

$$\Big\| \sum_{j=1}^d g_j \Big\|_\infty \le \sum_{j=1}^d \|g_j\|_\infty \le R_0 \sum_{j=1}^d J(g_j) \le R_0, \tag{B.13}$$

where $R_0 > 0$ is the constant from Lemma B.2, and we have used $J(m_j) = J(g_j)$, for $j = 1, \ldots, d$, as the null space of J contains kth degree polynomials.

The triangle inequality and (B.13) now imply

$$\|p\|_n \le \|m\|_n + \|g\|_n \le \rho + R_0. \tag{B.14}$$

Write

$$p(x) = \sum_{j=1}^d \sum_{\ell=1}^k \alpha_{j\ell} x_j^\ell, \quad \text{for } x \in [0,1]^d,$$

for some coefficients $\alpha_{j\ell}$, $j = 1, \ldots,$ $\ell = 1, \ldots, k$. For $V \in \mathbb{R}^{n \times kd}$ the basis matrix as in Lemma B.3, and $\alpha = (\alpha_{11}, \ldots, \alpha_{1k}, \ldots, \alpha_{d1}, \ldots, \alpha_{dk}) \in \mathbb{R}^{kd}$, we have

$$\|p\|_n = \frac{1}{\sqrt{n}} \|V\alpha\|_2.$$

Furthermore, noting

$$\|p\|_n \ge \sqrt{\lambda_{\min}\Big(\frac{1}{n} V^T V\Big)}\, \|\alpha\|_2,$$

we have

$$\|\alpha\|_2 \le \frac{\rho + R_0}{\nu_n},$$

where $\nu_n^2 = \lambda_{\min}(V^T V/n)$, as in Lemma B.3, and we have used the upper bound in (B.14). Using part (c) of Lemma B.3, with $t = 1/2$, we have

$$\|\alpha\|_2 \le \frac{2(\rho + R_0)}{\nu_0},$$

with probability at least $1 - d\exp(-\nu_0 n/(8(kd)^2))$, where $\nu_0^2 = \lambda_{\min}(\mathbb{E}[V^T V]/n)$, as in Lemma B.3. Therefore, using the triangle inequality and the fact that Q is supported on $[0,1]^d$, we have for each $j = 1, \ldots, d$, and any $x_j \in [0,1]$,

$$|p_j(x_j)| \le \sum_{\ell=1}^k |\alpha_{j\ell}| \le \|\alpha\|_1 \le \frac{2\sqrt{kd}(\rho + R_0)}{\nu_0},$$

with probability at least $1 - d\exp(-\nu_0 n/(8(kd)^2))$. Finally, for each $j = 1, \ldots, d$, using the triangle inequality, and the sup norm bound from Lemma B.2 once again,

$$\|m_j\|_\infty \leq \|p_j\|_\infty + \|g_j\|_\infty \leq \frac{2\sqrt{kd}(\rho + R_0)}{\nu_0} + R_0,$$

with probability $1 - d\exp(-\nu_0 n/(8(kd)^2))$, completing the proof. □

We give a simple bound on the entropy of an arbitrary sum of sets in terms of the entropies of the original sets.

Lemma B.5. Given sets $S_1, \ldots, S_d$ and a norm $\|\cdot\|$, it holds that

$$\log N(\delta, \|\cdot\|, S_1 + \cdots + S_d) \leq \sum_{j=1}^{d} \log N(\delta/d, \|\cdot\|, S_j).$$

Proof. For $j = 1, \ldots, d$, suppose that S_j can be covered in N_j balls of radius δ/d, with centers at $s_j^1, \ldots, s_j^{N_j}$. Take an arbitrary $s \in S_1 + \cdots + S_d$, and write $s = \sum_{j=1}^d s_j$, with $s_j \in S_j$, $j = 1, \ldots, d$. For each $j = 1, \ldots, d$, there is some $s_j^{\ell_j}$ such that $\|s_j - s_j^{\ell}\| \leq \delta/d$, and so by the triangle inequality

$$\left\| \sum_{j=1}^{d} s_j - \sum_{j=1}^{d} s_j^{\ell_j} \right\| \leq \delta.$$

That is, we have shown that $\prod_{j=1}^d N_j$ balls of radius δ with centers at

$$\sum_{j=1}^{d} s_j^{\ell_j}, \quad \text{for } (\ell_1, \ldots, \ell_d) \in \{1, \ldots, N_1\} \times \cdots \times \{1, \ldots, N_d\},$$

cover S. This completes the proof. □

The next result represents our main tool from empirical process theory that will be used in the proof of Theorem 3.1. It is essentially an application of Lemma 3.5 in van de Geer (1990) (see also van de Geer (2000)).

Lemma B.6. Let X^i, $i = 1, \ldots, n$ denote an i.i.d. sample from a continuous distribution Q on $[0,1]^d$. Let ϵ^i, $i = 1, \ldots, n$ be uniformly sub-Gaussian random variates that have variance proxy $\sigma^2 > 0$ and are independent of X^i, $i = 1, \ldots, n$. Let J be a seminorm satisfying Assumptions C1, C2, C3, and let $\rho > 0$ be arbitrary. Then there are constants $c_1, c_2, c_3, n_0 > 0$, depending only on $d, \sigma, k, L, K, w, \rho$, such that for all $c \geq c_1$ and $n \geq n_0$,

$$\sup_{m \in \mathcal{M}_n(1) \cap B_n(\rho)} \frac{\frac{1}{n}\sum_{i=1}^n \epsilon^i m(X^i)}{\|m\|_n^{1-w/2}} \leq \frac{c}{\sqrt{n}},$$

with probabilty at least $1 - \exp(-c_2 c^2) - \exp(-c_3 n)$.

Proof. Let Ω_1 denote the event on which the conclusion in Lemma B.4 holds, which has probability at least $1 - \exp(-c_3 n)$ for $n \geq n_1$, for constants $c_3, n_1 > 0$. Also let $R_0 = R_1\rho + R_2$, where $R_1, R_2 > 0$ are the constants from the lemma. Denote

$$B_\infty^d(\delta) = \bigg\{ \sum_{j=1}^d f_j : \|f_j\|_\infty \leq \delta,\ j = 1, \ldots, d \bigg\}.$$

On Ω_1, consider

$$\begin{aligned}
\log N\big(\delta, \|\cdot\|_n, \mathcal{M}_n(1) \cap B_n(\rho)\big) &\leq \log N\big(\delta, \|\cdot\|_n, \mathcal{M}_n(1) \cap B_\infty^d(R_0)\big) && \text{(B.15)}\\
&\leq \sum_{j=1}^d \log N\big(\delta/d, \|\cdot\|_n, B_J(1) \cap B_\infty(R_0)\big) && \text{(B.16)}\\
&\leq \sum_{j=1}^d \log N\big(\delta/(R_0 d), \|\cdot\|_n, B_J(1) \cap B_\infty(1)\big) && \text{(B.17)}\\
&\leq K d^{1+w} (R_0)^w \delta^{-w}. && \text{(B.18)}
\end{aligned}$$

The first inequality (B.15) uses the sup norm bound from Lemma B.4; the second inequality (B.16) uses

$$\mathcal{M}_n(1) \cap B_\infty^d(R_0) \subseteq \bigg\{ \sum_{j=1}^d m_j : m_j \in B_J(1) \cap B_\infty(R_0),\ j = 1, \ldots, d \bigg\},$$

and applies Lemma B.5 to the space on the right-hand side above. The third inequality (B.17) just uses the fact we may assume $R_0 \geq 1$, without a loss of generality; and the last inequality (B.18) uses Assumption C3. The entropy bound established in (B.18) allows us to apply Lemma 3.5 van de Geer (1990) (see also Lemma 8.4 in van de Geer (2000)), which gives constants $c_1, c_2, n_2 > 0$, depending only on d, σ, k, R_0, K, w, such that for all $c \geq c_1$ and $n \geq n_1$,

$$\sup_{m \in \mathcal{M}_n(1) \cap B_n(\rho)} \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon^i m(X^i)}{\|m\|_n^{1-w/2}} \leq c$$

on an event Ω_2 with probability at least $1 - \exp(-c_2 c^2)$. The desired result in the lemma therefore holds for all $c \geq c_1$ and $n \geq n_0 = \max\{n_1, n_2\}$, on $\Omega_1 \cap \Omega_2$. □

We finish with two simple results, on shifting around exponents in sums and products.

Lemma B.7. For any $a, b \geq 0$, and any $0 < q < 1$,

$$(a+b)^q \leq a^q + b^q.$$

Proof. The function $f(t) = (1+t)^q - (1+t^q)$ has derivative $f'(t) = q(1+t)^{q-1} - qt^{q-1} < 0$ for all $t > 0$, and so $f(t) < f(0) = 0$ for all $t > 0$. Plugging in $t = a/b$ and rearranging gives the claim. □

Lemma B.8. 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 $ab^{1-w/2} \leq a^{2/(1+w/2)}$. □

B.6 Proof of Theorem 3.1

This proof roughly follows the ideas in the proof of Theorem 9 in Mammen & van de Geer (1997), though it differs in a few key ways. We use $c > 0$ to denote a constant that will multiply our final estimation error bound; it will also control the probability with which our final result holds. Some steps will only hold for sufficiently large n, but we do not always make this explicit. Lastly, we will occasionally abuse our notation for the empirical norms and empirical inner products by using them with vector arguments, to be interpreted in the appropriate sense (e.g., $\langle m, v\rangle_n = \frac{1}{n}\sum_{i=1}^n v^i m(X^i)$ for a function m and vector $v \in \mathbb{R}^n$).

We break down the presentation of our proof into mini sections for readability.

Deriving a basic inequality. Denote by $\hat{f} = \sum_{j=1}^d \hat{f}_j$ the total additive fit in (3.18). Let $\mathcal{S}$ denote feasible set in (3.18). For any $f \in \mathcal{S}$, note that by orthogonality,

$$\|Y - \bar{Y}\mathbb{1} - f\|_n^2 = \|(f_0 + \epsilon - \bar{\epsilon}\mathbb{1}) - f\|_n^2 + (\bar{\epsilon})^2$$

where $\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^n \epsilon^i$. Therefore $\hat{f}$ must also be optimal for the problem

$$\min_{f \in \mathcal{S}} \ \frac{1}{2}\|W - f\|_n^2 + \lambda_n J_d(f),$$

where $W^i = f_0(X^i) + \epsilon^i - \bar{\epsilon}$, $i = 1, \ldots, n$, and we denote $\lambda_n = \lambda/n$ and $J_d(f) = \sum_{j=1}^d J(f_j)$. Standard arguments (from first-order optimality) show that any solution $\hat{f}$ in the above satisfies

$$\langle W - \hat{f}, \tilde{f} - \hat{f}\rangle_n \leq \lambda_n\big(J_d(\tilde{f}) - \lambda_n J_d(\hat{f})\big),$$

for any feasible $\tilde{f} = \sum_{j=1}^d \tilde{f}_j \in \mathcal{S}$. Expanding the definition of W and rearranging gives

$$\langle \hat{f} - f_0, \hat{f} - \tilde{f}\rangle_n \leq \langle \epsilon - \bar{\epsilon}\mathbb{1}, \hat{f} - \tilde{f}\rangle_n + \lambda_n\big(J_d(\tilde{f}) - \lambda_n J_d(\hat{f})\big).$$

Using the polarization identity $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a - b\|_2^2)$ for an inner product $\langle \cdot, \cdot \rangle$ and its corresponding norm $\|\cdot\|$,

$$\|\hat{f} - f_0\|_n^2 + \|\hat{f} - \tilde{f}\|_n^2 \leq 2\langle \epsilon - \bar{\epsilon}\mathbb{1}, \hat{f} - \tilde{f}\rangle_n + 2\lambda_n\big(J_d(\tilde{f}) - \lambda_n J_d(\hat{f})\big) + \|\tilde{f} - f_0\|_n^2.$$

Abbreviating $\hat{\Delta} = \hat{f} - \tilde{f}$, $\hat{J} = J_d(\hat{f})$, and $\tilde{J} = J_d(\tilde{f})$, and using $\langle \bar{\epsilon}\mathbb{1}, \hat{\Delta}\rangle = 0$, this becomes

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq 2\langle \epsilon, \hat{\Delta}\rangle_n + 2\lambda_n(\tilde{J} - \hat{J}) + \|\tilde{f} - f_0\|_n^2, \tag{B.19}$$

which is our basic inequality. In what follows, we will restrict our attention to feasible $\tilde{f}$ such that $\|\tilde{f} - f_0\|_n \leq \max\{C_n, \tilde{J}\}$.

Localizing the error vector. We prove that $\hat{\Delta}$ is appropriately bounded in the empirical norm. By the tail bound for quadratic forms of sub-Gaussian random variates in Theorem 2.1 of Hsu et al. (2012), for all $t > 0$,

$$\mathbb{P}\left(\|\epsilon\|_n^2 > \sigma^2\left(1 + \frac{2\sqrt{t}}{\sqrt{n}} + \frac{2t}{n}\right)\right) \leq e^{-t},$$

and hence taking $t = \sqrt{n}$,

$$\|\epsilon\|_n^2 \leq 5\sigma^2,$$

on an event Ω_1 with probability at least $1 - \exp(-\sqrt{n})$. Thus returning to the basic inequality (B.19), using the Cauchy-Schwartz inequality, and the above bound, we have

$$\|\hat{\Delta}\|_n^2 \leq 2\sqrt{5}\sigma\|\hat{\Delta}\|_n + 2\lambda_n\tilde{J} + \|\tilde{f} - f_0\|_n^2,$$

on Ω_1. This is a quadratic inequality of the form $x^2 \leq bx + c$ in $x = \|\hat{\Delta}\|_n$, so we can upper bound x by the larger of the two roots, $x \leq (b + \sqrt{b^2 + 4c})/2 \leq b + \sqrt{c}$, i.e.,

$$\|\hat{\Delta}\|_n \leq 2\sqrt{5}\sigma + \sqrt{2\lambda_n\tilde{J} + \|\tilde{f} - f_0\|_n^2},$$

on Ω_1. Abbreviating $J^* = \max\{C_n, \tilde{J}\}$, and using $J^* \geq 1$ (as $C_n \geq 1$ by assumption),

$$\|\hat{\Delta}\|_n \leq J^*\Big(2\sqrt{5}\sigma + \sqrt{2\lambda_n + \|\tilde{f} - f_0\|_n^2/(J^*)^2}\Big),$$

on Ω_1. Recalling $\|\tilde{f} - f_0\|_n \leq J^*$, and using the fact that $\lambda_n = o(1)$ for our eventual choice of λ_n, we have that for sufficiently large n,

$$\|\hat{\Delta}\|_n \leq J^*(2\sqrt{5}\sigma + \sqrt{2}), \tag{B.20}$$

on Ω_1.

Bounding the sub-Gaussian complexity term. We focus on the first term on the right-hand side in (B.19), i.e., the sub-Gaussian complexity term. Let $m = \hat{\Delta}/(\hat{J} + J^*)$. By construction, we have $J(m) \leq 1$, and from (B.20), we have $\|m\|_n \leq 2\sqrt{5}\sigma + \sqrt{2}$ on Ω_1. Then, applying Lemma B.6, with the choice $\rho = 2\sqrt{5}\sigma + \sqrt{2}$, we see that there are constants $c_1, c_2, c_3 > 0$ such that for all $c \geq c_1$,

$$\frac{2\langle \epsilon, m\rangle_n}{\|m\|_n^{1-w/2}} \leq \frac{c}{\sqrt{n}},$$

on $\Omega_1 \cap \Omega_2$, where Ω_2 is an event with probability at least $1 - \exp(-c_2c^2) - \exp(-c_3n)$. Plugging this into (B.19) gives

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq \frac{c}{\sqrt{n}}(\hat{J} + J^*)^{w/2}\|\hat{\Delta}\|_n^{1-w/2} + 2\lambda_n(\tilde{J} - \hat{J}) + \|\tilde{f} - f_0\|_n^2,$$

on $\Omega_1 \cap \Omega_2$. By the inequality in Lemma B.8, applied to the first term on the right-hand side above, with $a = n^{-1/2}(\hat{J} + J^*)^{w/2}$ and $b = \|\hat{\Delta}\|_n$,

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq cr_n(\hat{J} + J^*)^{w/(2+w)}\|\hat{\Delta}\|_n + cr_n^2(\hat{J} + J^*)^{2w/(2+w)} + 2\lambda_n(\tilde{J} - \hat{J}) + \|\tilde{f} - f_0\|_n^2,$$

on $\Omega_1 \cap \Omega_2$, where we abbreviate $r_n = n^{-1/(2+w)}$. Applying the simple inequality $2ab \leq a^2 + b^2$ to the first term on the right-hand side, with $a = cr_n(\hat{J} + J^*)^{w/(2+w)}$ and $b = \|\hat{\Delta}\|_n$, and subtracting $\|\hat{\Delta}\|_n^2/2$ from both sides,

$$\|\hat{f} - f_0\|_n^2 + \frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \frac{3}{2}c^2r_n^2(\hat{J} + J^*)^{2w/(2+w)} + 2\lambda_n(\tilde{J} - \hat{J}) + \|\tilde{f} - f_0\|_n^2, \tag{B.21}$$

on $\Omega_1 \cap \Omega_2$ (where we have assumed without a loss of generality that $c \geq 1$).

Controlling the effect of the penalty terms. Now we handle the appearances of the achieved penalty term $\hat{J}$. First, set $\lambda_n \geq (3/4)c^2 r_n^2 / C_n^{(2-w)/(2+w)}$, and denote

$$a = \frac{3}{2} c^2 r_n^2 (\hat{J} + J^*)^{2w/(2+w)} + 2\lambda_n (\tilde{J} - \hat{J}).$$

Consider the case $\hat{J} \geq C_n$. Then $-1/C_n^{(2-w)/(2+w)} \geq -1/\hat{J}^{(2-w)/(2+w)}$, and

$$2\lambda_n (\tilde{J} - \hat{J}) \leq 2\lambda_n \tilde{J} - (3/2) c^2 r_n^2 \hat{J}^{2w/(2+w)},$$

thus, using the simple inequality in Lemma B.7, we have $a \leq 4\lambda_n J^*$. In the case $\hat{J} < C_n$, we have by Lemma B.7 again,

$$a \leq \frac{3}{2} c^2 r_n^2 \Big(C_n^{2w/(2+w)} + (J^*)^{2w/(2+w)} \Big) + 2\lambda_n \tilde{J} \leq 6\lambda_n J^*.$$

Therefore, altogether, we conclude that $a \leq 6\lambda_n J^*$, and plugging this into (B.21) gives

$$\|\hat{f} - f_0\|_n^2 + \frac{1}{2} \|\hat{\Delta}\|_n^2 \leq 6\lambda_n J^* + \|\tilde{f} - f_0\|_n^2,$$

on $\Omega_1 \cap \Omega_2$. The statement (3.19) as made in the theorem follows by dropping the nonnegative term $\|\hat{\Delta}\|_n^2/2$ from the left-hand side, and adjusting the constants $c, c_1, c_2, c_3 > 0$ as needed.

B.7 Proof of the best additive approximation bound in (3.22)

We follow the exact same arguments as in the proof of Theorem 3.1, up until the last part, in which we control the achieved penalty terms $\hat{J}$. Now we deviate from the previous arguments, slightly. Set $\lambda_n \geq (3/2)c^2 r_n^2 / C_n^{(2-w)/(2+w)}$, and denote

$$a = \frac{3}{2} c^2 r_n^2 (\hat{J} + J^*)^{2w/(2+w)} + \lambda_n (\tilde{J} - \hat{J}).$$

By the same logic as in the proof of Theorem 3.1, we have $a \leq 3\lambda_n J^*$. Plugging this into (B.21) gives

$$\|\hat{f} - f_0\|_n^2 + \frac{1}{2} \|\hat{\Delta}\|_n^2 \leq 3\lambda_n J^* + \lambda_n (\tilde{J} - \hat{J}) + \|\tilde{f} - f_0\|_n^2,$$

on $\Omega_1 \cap \Omega_2$. Rearranging,

$$\frac{1}{2} \|\hat{\Delta}\|_n^2 \leq 3\lambda_n J^* + \Big(\|\tilde{f} - f_0\|_n^2 + \lambda_n \tilde{J} - \|\hat{f} - f_0\|_n^2 - \lambda_n \hat{J} \Big),$$

on $\Omega_1 \cap \Omega_2$. But, setting $\tilde{f} = f^{\text{best}}$, the bracketed term on the right-hand side above is nonpositive (by definition of f^{best} in (3.21)). This leads to (3.22), after adjusting $c, c_1, c_2, c_3 > 0$ as needed.

B.8 Preliminaries for the proof of Corollary 3.1

The following two lemmas will be helpful for the proof of Corollary 3.1.

Lemma B.9. Given $f = \sum_{j=1}^d f_j$, whose component functions are each k times weakly differentiable, there exists an additive spline approximant $\check{f} = \sum_{j=1}^d \check{f}_j$, where $\check{f}_j \in \mathcal{G}_j$, the set of kth order splines with knots in the set T_j defined in (3.17), for $j = 1, \ldots, d$, such that

(i) $\mathrm{TV}(\check{f}_j^{(k)}) \leq a_k \mathrm{TV}(f_j^{(k)})$, for $j = 1, \ldots, d$; and

(ii) $\|\check{f}_j - f_j\|_\infty \leq a_k W_{\max}^k \mathrm{TV}(f_j^{(k)})$, for $j = 1, \ldots, d$.

Above, $a_k \geq 1$ is a constant depending only on k, and we define $W_{\max} = \max_{j=1,\ldots,d} W_j$, where

$$W_j = \max_{i=1,\ldots,n-1} |X_j^{(i)} - X_j^{(i+1)}|, \quad j = 1, \ldots, d.$$

When the input points are drawn from a distribution Q that satisfies Assumptions A1, A2, there are universal constants $c_0, n_0 > 0$ such that for $n \geq n_0$, we have $W_{\max} \leq (c_0/b_0) \log n/n$ with probability at least $1 - 2b_0 d/n$, and so the bounds in (ii) become

$$\|\check{f}_j - f_j\|_\infty \leq \frac{c_0^k a_k}{b_0^k} \left(\frac{\log n}{n}\right)^k \mathrm{TV}(f_j^{(k)}), \quad \text{for } j = 1, \ldots, d, \tag{B.22}$$

with probability at least $1 - 2b_0 d/n$.

Proof. Parts (i) and (ii) are simply a componentwise application of Proposition 7 in Mammen & van de Geer (1997). In particular, from their result, we know that for $j = 1, \ldots, d$, there is a kth degree spline function $\check{f}_j$ whose knots lie in T_j in (3.17), with $\mathrm{TV}(\check{f}_j^{(k)}) \leq a_k \mathrm{TV}(f_j^{(k)})$ and

$$\|\check{f}_j - f_j\|_\infty \leq a_k W_j^k \mathrm{TV}(f_j^{(k)}),$$

where $a_k \geq 1$ depends only on k. (This result follows from strong quasi-interpolating properties of spline functions, from de Boor (1978).) This proves parts (i) and (ii).

When we consider random inputs drawn from a distribution Q satisfying Assumptions A1, A2, the densities of the marginals Q_j, $j = 1, \ldots, d$ will be bounded below by $b_0 > 0$, and thus there are universal constants $c_0, n_0 > 0$ such that for $n \geq n_0$, we have $W_j \leq (c_0/b_0) \log n/n$ with probability at least $1 - 2b_0/n$ (see, e.g., Lemma 5 in Wang et al. (2014)), for $j = 1, \ldots, d$, and hence applying a union bound gives the result for $W_{\max}$. □

Lemma B.10. Given $f = \sum_{j=1}^d f_j$, whose component functions are each k times weakly differentiable, there is an additive falling factorial approximant $\check{f} = \sum_{j=1}^d \check{f}_j$, where $\check{f}_j \in \mathcal{H}_j$, the set of kth order falling factorial functions defined over $X_j^1, \ldots, X_j^n$, for each $j = 1, \ldots, d$, such that

(i) $\mathrm{TV}(\check{f}_j^{(k)}) \leq a_k \mathrm{TV}(f_j^{(k)})$, for $j = 1, \ldots, d$; and

(ii) $\|\check{f}_j - f_j\|_\infty \leq a_k (W_{\max}^k + 2k^2 W_{\max}) \mathrm{TV}(f_j^{(k)})$, for $j = 1, \ldots, d$.

Again, $a_k \geq 1$ is a constant depending only on k, and $W_{\max}$ is as defined in Lemma B.9. When the inputs are drawn from a distribution Q satisfying Assumptions A1, A2, the bound in (ii) become

$$\|\check{f}_j - f_j\|_\infty \leq a_k \left(\frac{c_0^k}{b_0^k} \left(\frac{\log n}{n} \right)^k + 2k^2 \frac{c_0}{b_0} \frac{\log n}{n} \right) \mathrm{TV}(f_j^{(k)}), \quad \text{for } j = 1, \ldots, d, \tag{B.23}$$

with probability at least $1 - 2b_0 d/n$.

Proof. First we apply Lemma B.9 to produce an additive spline approximant, call it $f^* = \sum_{j=1}^d f_j^*$, to the given $f = \sum_{j=1}^d f_j$. Next, we parametrize the spline component functions in a helpful way:

$$f_j^* = \sum_{\ell=1}^n \alpha_j^\ell g_{j\ell}, \quad j = 1, \ldots, d.$$

where $\alpha_j^1, \ldots, \alpha_j^n \in \mathbb{R}$ are coefficients and $g_{j1} \ldots, g_{jn}$ are the truncated power basis functions over the knot set T_j defined in (3.17), and we write $g_{j\ell}(t) = t^{\ell-1}$, $\ell = 1, \ldots, k$ without a loss of generality, for $j = 1, \ldots, d$. It is not hard to check that $\mathrm{TV}((f_j^*)^{(k)}) = \sum_{\ell=k+2}^n |\alpha_j^\ell|$, for $j = 1, \ldots, d$.

We now define $\check{f} = \sum_{j=1}^d \check{f}_j$, our falling factorial approximant, to have component functions

$$\check{f}_j = \sum_{\ell=1}^{k+1} \alpha_j^\ell g_{j\ell} + \sum_{\ell=k+2}^n \alpha_j^\ell h_{j\ell}, \quad j = 1, \ldots, d.$$

where $h_{j1}, \ldots, h_{jn}$ are the falling factorial basis functions defined over $X_j^1, \ldots, X_j^n$, for $j = 1, \ldots, d$. (Note that $\check{f}_j$ preserves the polynomial part of f_j^* exactly, for $j = 1, \ldots, d$.) Again, it is straightforward to check that $\mathrm{TV}(\check{f}_j^{(k)}) = \sum_{\ell=k+2}^n |\alpha_j^\ell|$, for $j = 1, \ldots, d$, i.e.,

$$\mathrm{TV}(\check{f}_j^{(k)}) = \mathrm{TV}\big((f_j^*)^{(k)}\big) \leq a_k \mathrm{TV}(f_j^{(k)}), \quad \text{for } j = 1, \ldots, d,$$

the inequality coming from part (i) of of Lemma B.9. This verifies part (i) of the current lemma. As for part (ii), we note that Lemma 4 of Wang et al. (2014) shows that

$$|h_{j\ell}(X_j^i) - g_{j\ell}(X_j^i)| \leq k^2 W_j, \quad \text{for } \ell = k+2, \ldots, n,\ i = 1, \ldots, n,\ j = 1, \ldots, d,$$

where recall W_j is the maximum gap between sorted input points along the jth dimension, $j = 1, \ldots, d$, as defined in Lemma B.9. In fact, a straightforward modification of their proof can be used to strengthen this result to

$$\|h_{j\ell} - g_{j\ell}\|_\infty \leq 2k^2 W_j, \quad \text{for } \ell = k+2, \ldots, n,\ j = 1, \ldots, d,$$

which means that by Holder's inequality,

$$\|\check{f}_j - f_j^*\|_\infty \leq 2k^2 W_j \sum_{\ell=k+2}^n |\alpha_j^\ell| \leq 2k^2 a_k W_j \mathrm{TV}(f_j^{(k)}) \quad \text{for } j = 1, \ldots, d.$$

Then, by the triangle inequality,

$$\|\check{f}_j - f_j\|_\infty \leq \|\check{f}_j - f_j^*\|_\infty + \|f_j^* - f_j\|_\infty \leq a_k \Big(W_{\max}^k + 2k^2 W_{\max} \Big) \mathrm{TV}(f_j^{(k)}), \quad \text{for } j = 1, \ldots, d,$$

where we have used part (ii) of Lemma B.9. This verifies part (ii) of the current lemma.

Lastly, for random inputs drawn from a distribution Q satisfying Assumptions A1, A2, the proof of (B.23) follows the same arguments as the proof of (B.22). □

B.9 Proof of Corollary 3.1

We consider first the statement in part (a). We must check that Assumptions C1, C2, C3 hold for our choice of regularizer $J(g) = \mathrm{TV}(g^{(k)})$, and then we can apply Theorem 3.1. Assumptions C1, C2 are immediate. As for Assumption C3, consider the univariate function class

$$\mathcal{W}_{k+1} = \bigg\{ f : \int_0^1 |f^{(k+1)}(t)|\, dt \leq 1,\ \|f\|_\infty \leq 1 \bigg\}.$$

The results in Birman & Solomyak (1967) imply that for any set $Z_n = \{z^1, \ldots, z^n\} \subseteq [0,1]$,

$$\log N(\delta, \|\cdot\|_{Z_n}, \mathcal{W}_{k+1}) \leq K\delta^{-1/(k+1)},$$

for a universal constant $K > 0$. As explained in Mammen (1991), Mammen & van de Geer (1997), this confirms that Assumption C3 holds for our choice of regularizer, with $w = 1/(k+1)$. Applying Theorem 3.1, with $\tilde{f} = f_0$, gives the result in (3.23).

For the statement in part (b), note first that we can consider $k \geq 2$ without a loss of generality, as pointed out in Remark 3.4 following the corollary. Using Lemma B.9, can choose an additive spline approximant $\check{f}$ to f_0, with components $\check{f}_j \in \mathcal{G}_j$, $j = 1, \ldots, d$. Define $\tilde{f}_j$ to be the centered version of $\check{f}_j$, with zero empirical mean, $j = 1, \ldots, d$. By the fact that centering does not change the penalty, and part (i) of the lemma, $\sum_{j=1}^d \mathrm{TV}(\tilde{f}_j^{(k)}) \leq a_k \sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)})$. Also, using the fact that centering cannot increase the empirical norm, the triangle inequality, and (B.22), we get that with probability least $1 - 2b_0 d/n$,

$$\begin{aligned}
\bigg\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_{0j} \bigg\|_n &\leq \bigg\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_{0j} \bigg\|_n \\
&\leq \sum_{j=1}^d \|\tilde{f}_j - \sum_{j=1}^d f_{0j}\|_\infty \\
&\leq \frac{c_0^k a_k}{b_0^k} \bigg(\frac{\log n}{n} \bigg)^k \sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}),
\end{aligned}$$

When $\sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}) \leq C_n$, we see that $\|\tilde{f} - f_0\|_n$ is bounded by C_n for large enough n. This meets required condition for Theorem 3.1, by the above display, the approximation error in (3.19) satisfies

$$\bigg\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_{0j} \bigg\|_n^2 \leq \bigg(\frac{c_0^k a_k}{b_0^k} \bigg)^2 \bigg(\frac{\log n}{n} \bigg)^{2k} C_n^2.$$

But when $n/(\log n)^{1+1/k} \geq n_0 C_n^{(2k+2)/(2k^2+2k-1)}$, the right-hand side above is upper bounded by $a_0 n^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}$, for a constant $a_0 > 0$. This establishes the result in (3.23) for restricted additive locally adaptive splines.

For the statement in part (c), we can again consider $k \geq 2$ without a loss of generality. Then the same arguments as given for part (b) apply here, but now we use Lemma B.10 for the additive falling factorial approximant $\check{f}$ to f_0, and we require $n/(\log n)^{2k+3} \geq n_0 C_n^{4k+4}$ for the approximation error to be bounded by the estimation error.

B.10 Preliminaries for the proof of Theorem 3.2

Our first lemma is similar to Lemma B.6, but concerns univariate functions. As in Lemma B.6, this result relies on Lemma 3.5 in van de Geer (1990) (see also van de Geer (2000)).

Lemma B.11. Let ϵ^i, $i = 1, \dots, n$ be uniformly sub-Gaussian random variables having variance proxy $\sigma^2 > 0$. Let J be a seminorm satisfying Assumption C3, and let $\rho > 0$ be arbitrary. Then there exist constants $c_1, c_2, n_0 > 0$, depending only on σ, K, w, ρ, such that for all $c \geq c_1$ and $n \geq n_0$,

$$\sup_{Z_n = \{z^1, \dots, z^n\} \subseteq [0,1]} \sup_{g \in B_J(1) \cap B_\infty(\rho)} \frac{\frac{1}{n} \sum_{i=1}^n \epsilon^i g(z^i)}{\|g\|_{Z_n}^{1-w/2}} \leq \frac{c}{\sqrt{n}},$$

with probabilty at least $1 - \exp(-c_2 c^2)$, where we write $\|\cdot\|_{Z_n}$ for the empirical norm defined over a set of univariate points $Z_n = \{z^1, \dots, z^n\} \subseteq [0, 1]$.

Proof. Assume without a loss of generality that $\rho \geq 1$. Note that for any $Z_n = \{z^1, \dots, z^n\} \subseteq [0, 1]$,

$$\log N\big(\delta, \|\cdot\|_{Z_n}, B_J(1) \cap B_\infty(\rho)\big) \leq K \rho^w \delta^{-w},$$

by Assumption C3. As the right-hand side in the above entropy bound does not depend on Z_n, we can apply Lemma 3.5 in van de Geer (1990) to get the desired uniform control over all subsets. □

We give a coupling between the empirical and L_2 norms over $B_J(1) \cap B_\infty(\rho)$, using Theorem 14.1 in Wainwright (2019) (see also van de Geer (2000), Bartlett et al. (2005), Raskutti et al. (2012)).

Lemma B.12. Let z^i, $i = 1, \dots, n$ denote an i.i.d. sample from a distribution Λ on $[0, 1]$. Write $\|\cdot\|_2$ for the $L_2(\Lambda)$ norm, and $\|\cdot\|_n$ for the $L_2(\Lambda_n)$ norm. Let J satisfy Assumption C3, and let $\rho > 0$ be arbitrary. Then there are constants $c_1, c_2, c_3, n_0 > 0$, that depend only on K, w, ρ, such that for any $t \geq c_1 n^{-1/(2+w)}$ and $n \geq n_0$,

$$\big|\|g\|_n^2 - \|g\|_2^2\big| \leq \frac{1}{2}\|g\|_2^2 + \frac{t^2}{2}, \quad \text{for all } g \in B_J(1) \cap B_\infty(\rho),$$

with probability at least $1 - c_2 \exp(-c_3 n t^2)$.

Proof. Abbreviate $\mathcal{F} = B_J(1) \cap B_\infty(\rho)$. We will analyze the local Rademacher complexity

$$\mathcal{R}\big(\mathcal{F} \cap B_2(t)\big) = \mathbb{E}_{z,\sigma}\left[\sup_{g \in \mathcal{F} \cap B_2(t)} \frac{1}{n}\left|\sum_{i=1}^n \sigma^i g(z^i)\right|\right],$$

the expectation being taken over i.i.d. draws z^i, $i = 1, \dots, n$ from Λ and i.i.d. Rademacher variables σ^i, $i = 1, \dots, n$, as usual. Define the critical radius $\tau_n > 0$ to the smallest solution of the equation

$$\frac{\mathcal{R}\big(\mathcal{F} \cap B_2(t)\big)}{t} = \frac{t}{\rho}.$$

We will prove $\tau_n \le c_1 n^{-1/(2+w)}$ for a constant $c_1 > 0$. Applying Theorem 14.1 in Wainwright (2019) would then give the result.

In what follows, we will use $c > 0$ to denote a constant whole value may change from line to line (but does not depend on z^i, $i = 1, \dots, n$). Consider the empirical local Rademacher complexity

$$\mathcal{R}_n\big(\mathcal{F} \cap B_2(t)\big) = \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{F} \cap B_2(t)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i g(z^i) \right| \right].$$

As we are considering $t \ge \tau_n$, Corollary 2.2 of Bartlett et al. (2005) gives

$$\mathcal{F} \cap B_2(t) \subseteq \mathcal{F} \cap B_n(\sqrt{2}t),$$

with probability at least $1 - 1/n$. Denote by $\mathcal{E}$ the event that this occurs. Then on $\mathcal{E}$,

$$\begin{aligned}
\mathcal{R}_n\big(\mathcal{F} \cap B_2(t)\big) &\le \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{F} \cap B_n(\sqrt{2}t)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i g(z^i) \right| \right] \\
&\le \frac{c}{\sqrt{n}} \int_0^{\sqrt{2}t} \sqrt{\log N(\delta, \|\cdot\|_n, \mathcal{F})}\, d\delta \\
&\le \frac{c\sqrt{K}\rho^{w/2}}{\sqrt{n}} \int_0^{\sqrt{2}t} \delta^{-w/2}\, d\delta = \frac{c}{\sqrt{n}} t^{1-w/2},
\end{aligned}$$

where in second line we used Dudley's entropy integral (Dudley 1967), and in the third line we used Assumption C3. On $\mathcal{E}^c$, note that we have the trivial bound $\mathcal{R}_n(\mathcal{F} \cap B_2(t)) \le \rho$. Therefore we can upper bound the local Rademacher complexity, splitting the expectation over $\mathcal{E}$ and $\mathcal{E}^c$,

$$\mathcal{R}\big(\mathcal{F} \cap B_2(t)\big) = \mathbb{E}_z \mathcal{R}_n\big(\mathcal{F} \cap B_2(t)\big) \le \frac{ct^{1-w/2}}{\sqrt{n}} + \frac{\rho}{n} \le \frac{ct^{1-w/2}}{\sqrt{n}},$$

where the second inequality holds when n is large enough, as we may assume $t \ge n^{-1/2}$ without a loss of generality. An upper bound on the critical radius τ_n is thus given by the solution of

$$\frac{ct^{-w/2}}{\sqrt{n}} = \frac{t}{\rho},$$

which is $t = cn^{-1/(2+w)}$. This completes the proof. □

We extend Lemma B.2 to give a uniform sup norm bound on the functions in $B_J(1) \cap B_2(\rho)$.

Lemma B.13. Assume the conditions of Lemma B.2. Then there are constants $R_1, R_2 > 0$ that depend only on k, b_1, b_2, L, such that

$$\|m\|_\infty \le R_1 \rho + R_2, \quad \text{for all } m \in B_J(1) \cap B_2(\rho).$$

Proof. For $m \in B_J(1) \cap B_2(\rho)$, decompose $m = p + g$ where p is a polynomial of degree k, and g is orthogonal to all polynomials of degree k with respect to the $L_2(\Lambda)$ inner product. By Lemma B.2, we have $\|g\|_\infty \le R_0$ for a constant $R_0 > 0$, and by the triangle inequality,

$$\|p\|_2 \le \|m\|_2 + \|g\|_2 \le \rho + R_0.$$

Now write

$$p(x) = \sum_{\ell=1}^{k+1} \alpha_\ell \phi_\ell(x), \quad \text{for } x \in [0,1]^d,$$

where ϕ_ℓ, $\ell = 1, \ldots, k+1$ are orthonormal polynomials on $[0,1]$ with respect to the $L_2(\Lambda)$ inner product. Then $\|\alpha\|_2 = \|p\|_2 \leq \rho + R_0$, from the second to last display, and $\|\alpha\|_2 \leq \sqrt{k+1}(\rho + R_0)$, so for any $x \in [0,1]$,

$$|p(x)| \leq \|\alpha\|_1 \max_{\ell=1,\ldots,k+1} |\phi_\ell(x)| \leq c_k \sqrt{k+1}(\rho + R_0),$$

where $c_k = \max_{\ell=1,\ldots,k+1} \|\phi_\ell\|_\infty$ is a constant that depends only on k, b_1 from Aptekarev et al. (2016). Therefore

$$\|m\|_\infty \leq \|p\|_\infty + \|g\|_\infty \leq c_k \sqrt{k+1}(\rho + R_0) + R_0,$$

and defining $R_1, R_2 > 0$ appropriately, this is of the desired form, and completes the proof. □

Our last two lemmas pertain to the function space

$$\mathcal{M}_2(\delta) = \bigg\{ \sum_{i=1}^d m_j : J(m_j) \leq \delta, \text{ and } \langle m_j, 1 \rangle_2 = 0, \ j = 1, \ldots, d \bigg\}. \tag{B.24}$$

We derive a one-sided bound on the L_2 norm in terms of the empirical norm, over $\mathcal{M}_2(1)$. Our proof uses Theorem 14.2 in Wainwright (2019), which is a somewhat unique theorem, because it does not require a global sup norm bound on the function class in consideration (unlike many standard results of this type).

Lemma B.14. Let X^i, $i = 1, \ldots, n$ denote an i.i.d. sample from a distribution Q on $[0,1]^d$ satisfying Assumption A3, and let J satisfy Assumption C3. Then there are constants $c_1, c_2, c_3, n_0 > 0$, that depend only on b_1, b_2, k, L, K, w, such that for any $c_1 \sqrt{d} n^{-1/(2+w)} \leq t \leq 1$ and $n \geq n_0$,

$$\|m\|_2^2 \leq 2\|m\|_n^2 + t^2, \quad \text{for all } m \in \mathcal{M}_2(1),$$

with probability at least $1 - c_2 \exp(-c_3 n t^2)$, where $\mathcal{M}_2(1)$ is the space in (B.24).

Proof. Let $m \in \mathcal{M}_2(1)$ with $\|m\|_2 \leq 1$. Then as $\|m\|_2^2 = \sum_{j=1}^d \|m_j\|_2^2$, it follows that $\|m_j\|_2 \leq 1$, $j = 1, \ldots, d$, and by Lemma B.13, we have $\|m_j\|_\infty \leq R_1 + R_2$, $j = 1, \ldots, d$. From the calculation in Example 14.6 of Wainwright (2019), we have the property

$$\|m^2\|_2^2 \leq C^2 \|m\|_2^4, \quad \text{for all } m \in \mathcal{M}_2(1) \cap B_2(1),$$

where $C^2 = (R_1 + R_2)^2 + 6$. Abbreviating $\mathcal{F} = \mathcal{M}_2(1)$, we will study the local Rademacher complexity

$$\mathcal{R}\big(\mathcal{F} \cap B_2(t)\big) = \mathbb{E}_{z,\sigma} \left[\sup_{m \in \mathcal{F} \cap B_2(t)} \frac{1}{n} \bigg| \sum_{i=1}^n \sigma^i m(z^i) \bigg| \right],$$

and the associated critical radius $\tau_n > 0$, defined as usual to be the smallest solution of

$$\frac{\mathcal{R}\big(\mathcal{F} \cap B_2(t)\big)}{t} = \frac{t}{C}.$$

We will establish $\tau_n \leq c_1 \sqrt{d} n^{-1/(2+w)}$ for a constant $c_1 > 0$. Applying Theorem 14.2 in Wainwright (2019) would then give the result.

For the rest of the proof, we will use $c > 0$ for a constant whose value may change from line to line; also, many statements will hold for large enough n, but this will not always be made explicit. Fix some $0 < t \leq 1$. By L_2 orthogonality of the components of functions in $\mathcal{F}$,

$$\begin{aligned}
\sup_{m \in \mathcal{F} \cap B_2(t)} \frac{1}{n} \bigg| \sum_{i=1}^n \sigma^i m(X^i) \bigg| &\leq \sup_{\|\beta\|_2 \leq t} \sup_{\substack{m_j \in B_J(1) \cap B_2(|\beta_j|), \\ j=1,\ldots,d}} \bigg| \sum_{i=1}^n \sigma^i \sum_{j=1}^d m_j(X_j^i) \bigg| \\
&\leq \sup_{\|\beta\|_2 \leq t} \sum_{j=1}^d \sup_{m_j \in B_J(1) \cap B_2(|\beta_j|)} \frac{1}{n} \bigg| \sum_{i=1}^n \sigma^i m_j(X_j^i) \bigg|.
\end{aligned}$$

We now bound the inner supremum above, for an arbitrary $j = 1, \ldots, d$. Denote by τ_{nj} the critical radius of $B_J(1) \cap B_2(|\beta_j|)$, denote $r_n = n^{-1/(2+w)}$, and define the abbreviation $a \vee b = \max\{a, b\}$. Observe

$$\begin{aligned}
&\sup_{m_j \in B_J(1) \cap B_2(|\beta_j|)} \frac{1}{n} \bigg| \sum_{i=1}^n \sigma^i m_j(X_j^i) \bigg| \\
&\quad \leq c \bigg(\mathcal{R}_n\big(B_J(1) \cap B_2(|\beta_j|)\big) + \sqrt{\frac{\log n}{n}} \Big(\sup_{m_j \in B_J(1) \cap B_2(|\beta_j|)} \|m_j\|_n \Big) \bigg) \\
&\quad \leq c \bigg(\mathcal{R}\big(B_J(1) \cap B_2(|\beta_j|)\big) + \frac{\log n}{n} + \sqrt{\frac{\log n}{n}} \Big(\sup_{m_j \in B_J(1) \cap B_2(|\beta_j|)} \|m_j\|_n \Big) \bigg) \\
&\quad \leq c \bigg(\mathcal{R}\big(B_J(1) \cap B_2(|\beta_j|)\big) + \frac{\log n}{n} + \sqrt{\frac{\log n}{n}} \sqrt{2} (|\beta_j| \vee \tau_{nj}) \bigg) \\
&\quad \leq c \bigg(\frac{|\beta_j|^{1-w/2}}{\sqrt{n}} + \frac{\log n}{n} + (|\beta_j| \vee \tau_{nj}) \sqrt{\frac{\log n}{n}} \bigg) \\
&\quad \leq c \bigg(\frac{|\beta_j|^{1-w/2}}{\sqrt{n}} + (|\beta_j| \vee r_n) \sqrt{\frac{\log n}{n}} \bigg).
\end{aligned}$$

The first three inequalities above hold with probability at least $1 - 1/3n^2$ each. The first inequality is by Theorem 3.6 in Wainwright (2019) (see also Example 3.9 in Wainwright (2019)); the second and third are by Lemma A.4 and Lemma 3.6 in Bartlett et al. (2005), respectively. The fourth upper bounds the local Rademacher complexity of $B_J(1) \cap B_2(|\beta_j|)$, and the fifth upper bounds the critical radius τ_{nj} of this class, both following the proof of Lemma B.12 (recall, the functions in $B_J(1) \cap B_2(|\beta_j|)$ have a uniform sup norm bound of $\rho = R_1 + R_2$, by Lemma B.13). The last step also uses $\log n / n \leq r_n \sqrt{\log n / n}$ for n sufficiently large. The final result of the above display holds with probability at least

$1 - 1/n^2$; by a union bound, it holds with probability at least $1 - d/n^2$ simultaneously over $j = 1, \ldots, d$. Call this event $\mathcal{E}$. Then on $\mathcal{E}$,

$$\sup_{m \in \mathcal{F} \cap B_2(t)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i m(X^i) \right| \leq c \sum_{j=1}^d \left(\frac{|\beta_j|^{1-w/2}}{\sqrt{n}} + (|\beta_j| \vee r_n) \sqrt{\frac{\log n}{n}} \right)$$
$$\leq c \left(\frac{d^{(2+w)/4} t^{1-w/2}}{\sqrt{n}} + \sqrt{\frac{d \log n}{n}} t + d r_n^2 \right). \tag{B.25}$$

In the second line, we use Holder's inequality $a^T b \leq \|a\|_p \|b\|_q$ for the first term, with $p = 4/(2+w)$ and $q = 4/(2-w)$; we use $a \vee b \leq a + b$ for the second term, along the bound $\|\beta\|_1 \leq \sqrt{d} t$, and the fact that $r_n \sqrt{\log n / n} \leq r_n^2$ for large enough n.

Meanwhile, on $\mathcal{E}^c$, we can apply the simple bound $\|m\|_\infty \leq \sum_{j=1}^d \|m_j\|_\infty \leq \rho d$ for functions in $\mathcal{F} \cap B_2(t)$, where $\rho = R_1 + R_2$ (owing to Lemma B.13), and thus

$$\sup_{m \in \mathcal{F} \cap B_2(t)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i m(X^i) \right| \leq \rho d. \tag{B.26}$$

Splitting the expectation defining the local Rademacher complexity over $\mathcal{E}, \mathcal{E}^c$, and using (B.25), (B.26),

$$\mathcal{R}\big(\mathcal{F} \cap B_2(t)\big) = \mathbb{E}_{X,\sigma} \left[\sup_{m \in \mathcal{F} \cap B_2(t)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i m(X^i) \right| \right]$$
$$\leq c \left(\frac{d^{(2+w)/4} t^{1-w/2}}{\sqrt{n}} + \sqrt{\frac{d \log n}{n}} t + d r_n^2 \right) + \frac{\rho d^2}{n^2}. \tag{B.27}$$

It can be easily verified that for $t = c\sqrt{d} r_n^2$, the upper bound in (B.27) is at most t^2/C. Therefore this is an upper bound on the critical radius of $\mathcal{F}$, which completes the proof. □

Lastly, we bound the gap in the empirical and L_2 means of functions in $\mathcal{M}_2(1) \cap B_2(t)$, for small enough t. The proof uses Theorem 2.1 in Bartlett et al. (2005).

Lemma B.15. Let X^i, $i = 1, \ldots, n$ denote an i.i.d. sample from a distribution Q on $[0,1]^d$ satisfying Assumption A3, and let J satisfy Assumption C3. There are constants $c_0, n_0 > 0$, that depend only on b_1, b_2, k, L, K, w, such that for any $0 < t \leq 1$ and $n \geq n_0$,

$$\big| \langle m, 1 \rangle_n - \langle m, 1 \rangle_2 \big| \leq c_0 \left(\frac{d^{(2+w)/4} t^{1-w/2}}{\sqrt{n}} + \sqrt{\frac{d \log n}{n}} t + d n^{-2/(2+w)} \right), \quad \text{for all } m \in \mathcal{M}_2(1) \cap B_2(t),$$

with probability at least $1 - 1/n$, where $\mathcal{M}_2(1)$ is the space in (B.24).

Proof. This follows by combining the local Rademacher bound in (B.27) from the proof of Lemma B.14 with Theorem 2.1 in Bartlett et al. (2005), and simplifying by keeping the dominant terms for large enough n. □

B.11 Proof of Theorem 3.2

At a high-level, the difference between this proof and that of Theorem 3.1 is that here we do not try to directly control the sub-Gaussian complexity term (as this would lead to a poor dependence on the dimension d). Instead, we reduce the problem to controlling univariate sub-Gaussian complexities, and then assemble the result using ties between the empirical and L_2 norms, and the decomposition property (3.25). We will use the same general notation as in the proof of Theorem 3.1: $c > 0$ denotes a constant that will multiply our final bound, and will control the probability with which the final result holds; we will use the empirical norms and inner products with vector arguments, to be interpreted appropriately; we use the abbreviations $r_n, \hat{\Delta}$, and so on. Finally, in many lines that follow, we will redefine c by absorbing constant factors into it, without explicit notice.

The same arguments that led us to (B.19) yield the basic inequality

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq 2\langle \epsilon, \hat{\Delta} \rangle_n + \|\tilde{f} - f_0\|_n^2 = 2 \sum_{j=1}^d \langle \epsilon, \hat{\Delta}_j \rangle_n + \|\tilde{f} - f_0\|_n^2, \tag{B.28}$$

where we write $\hat{\Delta} = \sum_{j=1}^d \hat{\Delta}_j$.

Bounding the sub-Gaussian complexity terms. We now bound the univariate sub-Gaussian complexity terms, appearing in the sum on the right-hand side in (B.28). For $j = 1, \ldots, d$, define $g_j = \hat{\Delta}_j / (2\delta + \|\hat{\Delta}_j\|_n)$, and note that by construction $J(g_j) \leq 1$ and $\|g_j\|_n \leq 1$. By Lemma B.4, there are constants $c_0, R > 0$ such that $\|g_j\|_\infty \leq R$ on an event whose probability is at least $1 - \exp(-c_0 n)$. Thus by Lemma B.11, there are constants $c_1, c_2 > 0$ such that for all $c \geq c_1$,

$$\frac{2\langle \epsilon, g_j \rangle_n}{\|g_j\|_n^{1-w/2}} \leq \frac{c}{\sqrt{n}}, \quad \text{for all } j = 1, \ldots, d,$$

on an event Ω_1 with probability at least $1 - \exp(c_0 n) - \exp(-c_2 c^2)$. Plugging this into (B.28) gives

$$\begin{aligned} \|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 &\leq \frac{c}{\sqrt{n}} \sum_{j=1}^d (2\delta + \|\hat{\Delta}_j\|_n)^{w/2} \|\hat{\Delta}_j\|_n^{1-w/2} + \|\tilde{f} - f_0\|_n^2, \\ &\leq \frac{c\delta^{w/2}}{\sqrt{n}} \sum_{j=1}^d \|\hat{\Delta}_j\|_n^{1-w/2} + \frac{c}{\sqrt{n}} \sum_{j=1}^d \|\hat{\Delta}_j\|_n + \|\tilde{f} - f_0\|_n^2, \end{aligned} \tag{B.29}$$

on Ω_1, where we used Lemma B.7 in the second inequality.

Converting empirical norms into L_2 norms. For each $j = 1, \ldots, d$, let $\bar{\Delta}_j = \langle \hat{\Delta}_j, 1 \rangle_2$ be the L_2 mean of $\hat{\Delta}_j$, and $\tilde{\Delta}_j = \hat{\Delta}_j - \bar{\Delta}_j$ the L_2 centered version of $\hat{\Delta}_j$. Note that, for each $j = 1, \ldots, d$, we have by empirical orthogonality $\|\tilde{\Delta}_j\|_n^2 = \|\hat{\Delta}_j\|_n^2 + |\bar{\Delta}_j|^2$, which implies $\|\hat{\Delta}_j\|_n \leq \|\tilde{\Delta}_j\|_n$. Applying this to upper bound the right-hand side in (B.29) gives

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq \frac{c\delta^{w/2}}{\sqrt{n}} \sum_{j=1}^d \|\tilde{\Delta}_j\|_n^{1-w/2} + \frac{c}{\sqrt{n}} \sum_{j=1}^d \|\tilde{\Delta}_j\|_n + \|\tilde{f} - f_0\|_n^2, \tag{B.30}$$

on Ω_1. We bound each empirical norm in the sum on the right-hand side in (B.30) by its L_2 norm counterpart. Now, for each $j = 1, \ldots, d$, define $g_j = \tilde{\Delta}_j/(2\delta + \|\tilde{\Delta}_j\|_2)$. Since $J(g_j) \leq 1$ and $\|g_j\|_2 \leq 1$, by Lemma B.13, there is a constant $R > 0$ such that $\|g_j\|_\infty \leq R$. We can hence apply Lemma B.12 to the measure $\Lambda = Q_j$, which gives constants $c_3, c_4, c_5 > 0$ such that

$$\|g_j\|_n \leq \sqrt{\frac{3}{2}}\|g_j\|_2 + c_3 r_n, \quad \text{for all } j = 1, \ldots, d,$$

on an event Ω_2 with probability at least $1 - c_4 d \exp(-c_5 n r_n^2)$, where recall $r_n = n^{-1/(2+w)}$, i.e.,

$$\|\tilde{\Delta}_j\|_n \leq 2\sqrt{\frac{3}{2}}\|\tilde{\Delta}_j\|_2 + 2c_3 r_n \delta, \quad \text{for all } j = 1, \ldots, d,$$

on Ω_2, where we assume n is large enough so that $c_3 r_n \leq \sqrt{3/2}$. Returning to (B.30), and using the simple inequality in Lemma B.7, we have

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq \frac{c\delta^{w/2}}{\sqrt{n}} \sum_{j=1}^d \|\tilde{\Delta}_j\|_2^{1-w/2} + \frac{c}{\sqrt{n}} \sum_{j=1}^d \|\tilde{\Delta}_j\|_2 + c d r_n^2 \delta + \|\tilde{f} - f_0\|_n^2, \quad \text{(B.31)}$$

on $\Omega_1 \cap \Omega_2$.

Invoking L_2 decomposability. We recall the key L_2 decomposition property (3.25), of additive functions with L_2 mean zero components. Using Holder's inequality $a^T b \leq \|a\|_p \|b\|_q$ to bound the first sum on the right-hand side in (B.31), with $p = 4/(2+w)$ and $q = 4/(2-w)$, and Cauchy-Schwartz to bound the second sum in (B.31), we get

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq \frac{c d^{(2+w)/4} \delta^{w/2}}{\sqrt{n}} \|\tilde{\Delta}\|_2^{1-w/2} + c\sqrt{\frac{d}{n}} \|\tilde{\Delta}\|_2 + c d r_n^2 \delta + \|\tilde{f} - f_0\|_n^2, \quad \text{(B.32)}$$

on $\Omega_1 \cap \Omega_2$, where we denote $\tilde{\Delta} = \sum_{j=1}^d \tilde{\Delta}_j$.

Converting back to empirical norm. We bound the L_2 norm of the centered error vector on the right-hand side in (B.32) with its empirical norm counterpart. By Lemma B.14 applied to $m = \tilde{\Delta}/(2\delta)$, provided n is large enough so that $c_6 \sqrt{d} r_n \leq 1$, there are constants $c_6, c_7, c_8 > 0$ such that

$$\|\tilde{\Delta}\|_2 \leq \sqrt{2}\|\tilde{\Delta}\|_n + 2c_6 \sqrt{d} r_n \delta, \quad \text{(B.33)}$$

on an event Ω_3 with probability at least $1 - c_7 \exp(-c_8 d n r_n^2)$. Plugging this into the right-hand side in (B.32), and using Lemma B.7, we have

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq \frac{c d^{(2+w)/4} \delta^{w/2}}{\sqrt{n}} \|\tilde{\Delta}\|_n^{1-w/2} + c\sqrt{\frac{d}{n}} \|\tilde{\Delta}\|_n + c d r_n^2 \delta + \|\tilde{f} - f_0\|_n^2,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3$. Using Lemma B.8 on the first term above, with $a = d^{(2+w)/4} \delta^{w/2}/\sqrt{n}$ and $b = \|\tilde{\Delta}\|_n$, and simplifying, gives

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq c\sqrt{d} r_n \sqrt{\delta} \|\tilde{\Delta}\|_n + c d r_n^2 \delta + \|\tilde{f} - f_0\|_n^2, \quad \text{(B.34)}$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3$.

Deriving an empirical norm error bound. Note that in (B.34), we have $\|\hat{\Delta}\|_n$ on the left-hand side and $\|\tilde{\Delta}\|_n$ on the right-hand side, where $\tilde{\Delta} = \hat{\Delta} - \bar{\Delta}$ is the centered error vector, and we are abbreviating $\bar{\Delta} = \sum_{j=1}^d \bar{\Delta}_j$. We seek to bound $|\bar{\Delta}|$. Define $t = c_6\sqrt{d}r_n$, where c_6 is the constant in (B.33), and define

$$m = \frac{t\tilde{\Delta}/(2\delta)}{\sqrt{2}\|\tilde{\Delta}\|_n/(2\delta) + t}.$$

Note that $J(m_j) \leq J(\tilde{\Delta}_j)/(2\delta) \leq 1$, for $j = 1, \ldots, d$, by construction, and also

$$\|m\|_2 = \frac{t\|\tilde{\Delta}\|_2/(2\delta)}{\sqrt{2}\|\tilde{\Delta}\|_n/(2\delta) + t} \leq t,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3$, recalling (B.33). By Lemma B.15 applied to m, provided n is large enough such that $t = c_6\sqrt{d}r_n \leq 1$, there is a constant $c_9 > 0$ such that $|\langle m, 1\rangle_n| \leq c_9 t^2$ on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$, where Ω_4 is an event with probability at least $1 - 1/n$, i.e.,

$$|\langle 1, \tilde{\Delta}\rangle_n|/(2\delta) \leq c_9 t\big(\sqrt{2}\|\tilde{\Delta}\|_n/(2\delta) + t\big),$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$, i.e.,

$$|\bar{\Delta}| \leq \sqrt{2}c_9 t\|\tilde{\Delta}\|_n + 2c_9 t^2\delta,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. Thus, by empirical orthogonality,

$$\|\tilde{\Delta}\|_n^2 = \|\hat{\Delta}\|_n^2 + |\bar{\Delta}|^2 \leq \|\hat{\Delta}\|_n^2 + 2(\sqrt{2}c_9 t)^2\|\tilde{\Delta}\|_n^2 + 2(2c_9 t^2\delta)^2,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$, and assuming n is large enough so that $2(\sqrt{2}c_9 t)^2 \leq 1/2$ and $2(2c_9)^2 t^2\delta \leq 1$, this becomes

$$\frac{1}{2}\|\tilde{\Delta}\|_n^2 \leq \|\hat{\Delta}\|_n^2 + t^2\delta, \tag{B.35}$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. Using this on the right-hand side in (B.34) gives

$$\|\hat{f} - f_0\|_n^2 + \|\hat{\Delta}\|_n^2 \leq c\sqrt{d}r_n\sqrt{\delta}\|\hat{\Delta}\|_n + cdr_n^2\delta + \|\tilde{f} - f_0\|_n^2,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. Using the simple inequality $2ab \leq a^2 + b^2$ on the first term on the right-hand side above, with $a = c\sqrt{d}r_n\sqrt{\delta}$ and $b = \|\hat{\Delta}\|_n$, gives

$$\|\hat{f} - f_0\|_n^2 + \frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \|\tilde{f} - f_0\|_n^2 + c^2 dr_n^2\delta, \tag{B.36}$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. The empirical norm result in (3.26) in the theorem follows by dropping the nonnegative term $\|\hat{\Delta}\|_n^2/2$ from the left-hand side, and adjusting the constants $c, c_1, c_2, c_3 > 0$ as needed.

Deriving an L_2 norm error bound. Note that (B.36) also implies

$$\frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \|\tilde{f} - f_0\|_n^2 + c^2 dr_n^2\delta,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. Recalling (B.35), this gives

$$\|\tilde{\Delta}\|_n^2 \leq 4\|\tilde{f} - f_0\|_n^2 + c^2 d r_n^2 \delta, \tag{B.37}$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. By L_2 orthogonality,

$$\begin{aligned}
\|\hat{\Delta}\|_2^2 &= \|\tilde{\Delta}\|_2^2 + |\bar{\Delta}|^2 \\
&\leq 3\|\tilde{\Delta}\|_n^2 + t^2\delta^2 \\
&\leq 12\|\tilde{f} - f_0\|_n^2 + c^2 d r_n^2 \delta^2,
\end{aligned}$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$, where in the second line we used (B.33) and $|\bar{\Delta}| \leq \|\tilde{\Delta}\|_n$, and in the third line we used (B.37). Finally,

$$\|\hat{f} - f_0\|_2^2 \leq 2\|\hat{f} - \tilde{f}\|_2^2 + 2\|\tilde{f} - f_0\|_2^2 \leq 24\|\tilde{f} - f_0\|_n^2 + 2\|\tilde{f} - f_0\|_2^2 + c^2 d r_n^2 \delta^2,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. The L_2 norm result in (3.27) in the theorem follows by simply adjusting the constants $c, c_1, c_2, c_3 > 0$ as needed.

B.12 Proof of Corollary 3.2

The proof of the statement in part (a) is exactly as in the proof of part (a) in Corollary 3.1.

For part (b), we can consider $k \geq 2$ without a loss of generality, and start with an additive spline approximant $\check{f}$ to f_0 from Lemma B.9. Let $\tilde{f}$ denote the result of centering each component of $\check{f}$ to have zero empirical mean. Then $\mathrm{TV}(\tilde{f}_j^{(k)}) \leq a_k c_n = \delta$, $j = 1, \ldots, d$, and just as in the proof of part (b) in Corollary 3.1, letting $\|\cdot\|$ denote either the empirical or L_2 norm, we have

$$\Big\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_{0j} \Big\|^2 \leq \Big(\frac{c_0^k a_k}{b_0^k} \Big)^2 \Big(\frac{\log n}{n} \Big)^{2k} d^2 c_n^2.$$

But when $n \geq n_0 (d c_n)^{(2k+3)/(2k+2)}$, the right-hand side above is bounded by $a_0 d n^{-(2k+2)/(2k+3)} c_n$ for a constant $a_0 > 0$, which shows the approximation error terms in (3.26), (3.27) are of the desired order. This proves the desired result for restricted locally adaptive splines.

For part (c), we follow the same arguments, the only difference being that we construct a falling factorial approximant $\check{f}$ to f_0 from Lemma B.10.

B.13 Preliminaries for the proof of Theorem 3.3

The next two results in this subsection are helper lemmas for the last lemma.

Lemma B.16. Let J be a functional that satisfies Assumptions C1, C2, C4. Then there are constants $\tilde{K}_1, \tilde{\delta}_1 > 0$, that depend only on k, L, K_1, w, such that for all $0 < \delta \leq \tilde{\delta}_1$,

$$\log M\Big(\delta, \|\cdot\|_2, \Pi_k^{\perp}\big(B_J(1)\big)\Big) \geq \tilde{K}_1 \delta^{-w},$$

where $\|\cdot\|_2$ is the $L_2(U)$ norm, with U the uniform distribution on $[0,1]$, and $\Pi_k^{\perp}$ is defined by

$$\Pi_k^{\perp}(g) = g - \Pi_k(g), \quad \text{where} \quad \Pi_k(g) = \operatorname*{argmin}_{p\in\mathcal{P}_k} \|g-p\|_2,$$

with $\mathcal{P}_k$ denoting the space of polynomials of degree k. In other words, $\Pi_k^{\perp}$ is the projection operator onto the space orthogonal (in $L_2(U)$) to the polynomials of degree k.

Proof. Let $R_0 > 0$ be the constant from Lemma B.2, when we take $\Lambda = U$. Note that

$$B_J(1) \cap B_\infty(R_0) = \Pi_k^{\perp}\big(B_J(1)\big) + \big(\mathcal{P}_k \cap B_\infty(R_0)\big). \tag{B.38}$$

In general, for $S = S_1 + S_2$ and a norm $\|\cdot\|$, observe that, from basic relationships between covering and packing numbers,

$$M(4\delta, \|\cdot\|, S) \le N(2\delta, \|\cdot\|, S) \le N(\delta, \|\cdot\|, S_1) N(\delta, \|\cdot\|, S_2) \le M(\delta, \|\cdot\|, S_1) N(\delta, \|\cdot\|, S_2),$$

so that

$$\log M(\delta, \|\cdot\|, S_1) \ge \log \frac{M(4\delta, \|\cdot\|, S)}{N(\delta, \|\cdot\|, S_2)}.$$

Applying this to our decomposition in (B.38),

$$\begin{aligned}\log M\Big(\delta, \|\cdot\|_2, \Pi_k^{\perp}\big(B_J(1)\big)\Big) &\ge \log \frac{M\big(4\delta, \|\cdot\|_2, B_J(1)\cap B_\infty(R_0)\big)}{N\big(\delta, \|\cdot\|_2, \mathcal{P}_k \cap B_\infty(R_0)\big)} \\ &\ge K_1 R_0^w 4^{-w} \delta^{-w} - A(k+1)\log(1/\delta),\end{aligned}$$

where in the second inequality we used Assumption C4 (assuming without a loss of generality that $R_0 \ge 1$), and a well-known entropy bound for a finite-dimensional ball (e.g., Mammen (1991)), with $A > 0$ being a constant that depends only on R_0. For small enough δ, the right-hand side above is of the desired order, and this completes the proof. □

Lemma B.17. Let $d, M > 0$ be integers, and $I = \{1, \dots, M\}$. Denote by $H(u,v) = \sum_{j=1}^d 1\{u_j \ne v_j\}$ the Hamming distance between $u, v \in I^d$. Then there is a subset $S \subseteq I^d$ with $|S| \ge (M/4)^{d/2}$ such that $H(u,v) \ge d/2$ for any $u, v \in S$.

Proof. Let $\Omega_0 = I^d$, $u_0 = (1, \dots, 1) \in \Omega_0$. For $j = 0, 1, \dots$, recursively define

$$\Omega_{j+1} = \{u \in \Omega_j : H(u, u_j) > a = \lceil d/2 \rceil\},$$

where u_{j+1} is arbitrarily chosen from Ω_{j+1}. The procedure is stopped when Ω_{j+1} is empty; denote the last set defined in this procedure by Ω_E, and denote $S = \{u_0, \dots, u_E\}$. For $0 \le i, j \le E$, by construction, $H(u_i, u_j) > a$. For $j = 0, \dots, E$,

$$\begin{aligned} n_j = |\Omega_j - \Omega_{j+1}| &= |\{u \in \Omega_j : H(u, u_j) \le a\}| \\ &\le |\{u \in I^d : H(u, u_j) \le a\}| \\ &= \binom{d}{d-a} M^a \end{aligned}$$

The last step is true because we can choose $d-a$ positions in which u matches u_j in $\binom{d}{d-a}$ ways, and the rest of the a positions can be filled arbitrarily in M ways. Also note $M^d = n_0 + \cdots + n_E$. Therefore

$$M^d \leq (E+1)\binom{d}{d-a} M^a,$$

which implies

$$E + 1 \geq \frac{M^{d-a}}{\binom{d}{d-a}} \geq \frac{M^{d-a}}{2^d} \geq (M/4)^{d/2}.$$

□

The lemma below gives a key technical result used in the proof of Theorem 3.3.

Lemma B.18. Let J be a functional that satisfies Assumptions C1, C2, C4. Then there are constants $\bar{K}_1, \bar{\delta}_1 > 0$, that depend only on $w, \tilde{K}_1, \tilde{\delta}_1$, where $\tilde{K}_1, \tilde{\delta}_1 > 0$ are the constants from Lemma B.16, such that for all $0 < \delta \leq \bar{\delta}_1$,

$$\log M\Big(\delta, \|\cdot\|_2, \Pi_{k,d}^{\perp}\big(B_J^d(1)\big)\Big) \geq \bar{K}_1 d^{1+w/2} \delta^{-w},$$

where $\|\cdot\|_2$ is the $L_2(U)$ norm, with U the uniform distribution on $[0,1]^d$, and $\Pi_{k,d}^{\perp}$ is defined by

$$\Pi_{k,d}^{\perp}(g) = g - \Pi_{k,d}(g), \quad \text{where} \quad \Pi_{k,d}(g) = \operatorname*{argmin}_{p \in \mathcal{P}_{k,d}} \|g - p\|_2,$$

and $\mathcal{P}_{k,d}$ contains all functions of the form $p(x) = \sum_{j=1}^d p_j(x_j)$, for polynomials p_j, $j = 1, \ldots, d$ of degree k. In other words, $\Pi_{k,d}^{\perp}$ is the projection operator onto the space orthogonal (in $L_2(U)$) to the space $\mathcal{P}_{k,d}$ of additive polynomials of degree k.

Proof. It is easy to check that the decomposability property of the $L_2(U)$ norm, in (3.25), implies a certain decomposability of the $L_2(U)$ projection operators $\Pi_{k,d}, \Pi_{k,d}^{\perp}$ over additive functions:

$$\Pi_{k,d}\Big(\sum_{j=1}^d m_j\Big) = \sum_{j=1}^d \Pi_k(m_j), \quad \Pi_{k,d}^{\perp}\Big(\sum_{j=1}^d m_j\Big) = \sum_{j=1}^d \Pi_k^{\perp}(m_j),$$

where $\Pi_k, \Pi_k^{\perp}$ are projection operators onto $\mathcal{P}_k$ and its orthocomplement, respectively, as defined in Lemma B.16. The decomposability result for $\Pi_{k,d}^{\perp}$ in particular implies that

$$\Pi_{k,d}^{\perp}\big(B_J^d(1)\big) = \Big\{ \sum_{j=1}^d f_j : f_j \in \Pi_k^{\perp}\big(B_J(1)\big),\ j = 1, \ldots, d \Big\}. \tag{B.39}$$

Abbreviate $M = M(\delta/\sqrt{d/2}, \|\cdot\|_2, \Pi_k^{\perp}(B_J(1)))$. By Lemma B.16, we have for small enough δ,

$$\log M \geq \tilde{K}_1 2^{-w/2} d^{w/2} \delta^{-w}.$$

Now let $g_1, \ldots, g_M$ denote a $(\delta/\sqrt{d/2})$-packing of $\Pi_k^{\perp}(B_J(1))$. Let $I = \{1, \ldots, M\}$, and for $u \in I^d$, define $f_u \in \Pi_k^{\perp}(B_J^d(1))$ by

$$f_u = \sum_{j=1}^d g_{u_j},$$

i.e., f_u is an additive function with components g_{u_j}, $j = 1, \ldots, d$. If the Hamming distance between indices u, v satisfies $H(u, v) \geq d/2$, then

$$\|f_u - f_v\|_2^2 = \sum_{j=1}^{d} \|g_{u_j} - g_{v_j}\|_2^2 \geq H(u, v) \frac{\delta^2}{d/2} \geq \delta^2,$$

where we have again used the $L_2(U)$ decomposability property in (3.25). Thus, it is sufficient to find a subset S of I^d such that $u, v \in S \Rightarrow H(u, v) \geq d/2$. By Lemma B.17, we can choose such an S with $|S| \geq (M/4)^{d/2}$. For small enough δ, such that $M \geq 16$, this gives the desired result because

$$\log |S| \geq \frac{d}{2} \log \frac{M}{4} \geq \frac{d}{4} \log M \geq \tilde{K}_1 2^{-w/2-2} d^{1+w/2} \delta^{-w}.$$

□

B.14 Proof of Theorem 3.3

Clearly, by orthogonality, for any functions $\hat{f}, f_0$,

$$\|\hat{f} - f_0\|_2^2 = \|\Pi_{k,d}(\hat{f}) - \Pi_{k,d}(f_0)\|_2^2 + \|\Pi_{k,d}^{\perp}(\hat{f}) - \Pi_{k,d}^{\perp}(f_0)\|_2^2 \geq \|\Pi_{k,d}^{\perp}(\hat{f}) - \Pi_{k,d}^{\perp}(f_0)\|_2^2,$$

where $\Pi_{k,d}, \Pi_{k,d}^{\perp}$ are projection operators onto $\mathcal{P}_{k,d}$ and its orthocomplement, respectively, defined in Lemma B.16. Thus it suffices to consider the minimax error over $\Pi_{k,d}^{\perp}(B_J^d(c_n))$.

First, we lower bound the packing number and upper bound the covering number of the class $\Pi_{k,d}^{\perp}(B_J^d(c_n))$. The upper bound is more straightforward:

$$\begin{aligned} \log N\Big(\epsilon, \|\cdot\|_2, \Pi_{k,d}^{\perp}\big(B_J^d(c_n)\big)\Big) &= \log N\Big(\epsilon/c_n, \|\cdot\|_2, \Pi_{k,d}^{\perp}\big(B_J^d(1)\big)\Big) \\ &\leq \sum_{j=1}^{d} \log N\Big(\epsilon/(c_n\sqrt{d}), \|\cdot\|_2, \Pi_k^{\perp}\big(B_J(1)\big)\Big) \\ &\leq K_2 c_n^w d^{1+w/2} \epsilon^{-w}. \end{aligned} \tag{B.40}$$

The second inequality follows from property (B.39) in the proof of Lemma B.18 and similar arguments to those in the proof of Lemma B.5—except that we leverage the decomposability of the L_2 norm, as in (3.25), instead of using the triangle inequality. The third inequality follows from Assumption C4.

The lower bound is less straightforward, and is given by Lemma B.18:

$$\begin{aligned} \log M\Big(\delta, \|\cdot\|_2, \Pi_{k,d}^{\perp}\big(B_J^d(c_n)\big)\Big) &= \log M\Big(\delta/c_n, \|\cdot\|_2, \Pi_{k,d}^{\perp}\big(B_J^d(1)\big)\Big) \\ &\geq \bar{K}_1 c_n^w d^{1+w/2} \delta^{-w}. \end{aligned} \tag{B.41}$$

We note that (B.41) holds for $0 < \delta \leq \bar{\delta}_1$, where $\bar{\delta}_1 > 0$ is the constant from Lemma B.18.

Now, following the strategy in Yang & Barron (1999), we use these bounds on the packing and covering numbers, along with Fano's inequality, to establish the desired result.

Let $f_1, f_2, \ldots, f_M$ be a δ_n-packing of $\Pi^{\perp}_{k,d}(B_J(c_n))$, for $\delta_n > 0$ to be specified later. Fix an arbitrary estimator $\hat{f}$, and let

$$\hat{Z} = \operatorname*{argmin}_{j \in \{1,\ldots,M\}} \|\hat{f} - f_j\|_2.$$

We will use $P_{X,f}$ and $\mathbb{E}_{X,f}$ to denote the probability and expectation operators, respectively, over i.i.d. draws $X^i \sim U$, $i = 1, \ldots, n$ (where U is the uniform distribution on $[0,1]^d$), and i.i.d. draws $Y^i | X^i \sim N(f(X^i), \sigma^2)$, $i = 1, \ldots, n$. Then

$$\begin{aligned}
\sup_{f_0 \in \Pi^{\perp}_{k,d}(B^d_J(c_n))} \mathbb{E}_{X,f_0} \|\hat{f} - f_0\|_2^2 &\geq \sup_{f_0 \in \{f_1,\ldots,f_M\}} \mathbb{E}_{X,f_0} \|\hat{f} - f_0\|_2^2 \\
&\geq \frac{1}{M} \mathbb{E}_X \sum_{j=1}^M \mathbb{E}_{f_j} \|\hat{f} - f_j\|_2^2 \\
&= \frac{1}{M} \mathbb{E}_X \sum_{j=1}^M \Big(\mathbb{P}_{f_j}(\hat{Z} \neq j) \mathbb{E}_{f_j}\big(\|\hat{f} - f_j\|_2^2 \,\big|\, \hat{Z} \neq j\big) + \mathbb{P}_{f_j}(\hat{Z} = j) \mathbb{E}_{f_j}\big(\|\hat{f} - f_j\|_2^2 \,\big|\, \hat{Z} = j\big) \Big) \\
&\geq \frac{1}{M} \mathbb{E}_X \sum_{j=1}^M \mathbb{P}_{f_j}(\hat{Z} \neq j) \mathbb{E}_{f_j}\big(\|\hat{f} - f_j\|_2^2 \,\big|\, \hat{Z} \neq j\big) \\
&\geq \frac{1}{M} \mathbb{E}_X \sum_{j=1}^M \mathbb{P}_{f_j}(\hat{Z} \neq j) \frac{\delta_n^2}{4}, \qquad \text{(B.42)}
\end{aligned}$$

where in the last inequality we have used the fact that if $\hat{Z} \neq j$, then $\hat{f}$ must be at least $\delta_n/2$ away from f_j, for each $j = 1, \ldots, M$.

Abbreviate q_j for the distribution P_{f_j}, $j = 1, \ldots, M$, and define the mixture $\bar{q} = \frac{1}{M} \sum_{j=1}^M q_j$. By Fano's inequality,

$$\frac{1}{M} \mathbb{E}_X \sum_{j=1}^M \mathbb{P}_{f_j}(\hat{Z} \neq j) \geq 1 - \frac{\frac{1}{M} \sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{q}) + \log 2}{\log M}, \qquad \text{(B.43)}$$

where $\mathrm{KL}(P_1 \,\|\, P_2)$ denotes the Kullback-Leibler (KL) divergence between distributions P_1, P_2. Let $g_1, g_2, \ldots, g_N$ be an ϵ_n-covering of $\Pi^{\perp}_{k,d}(B^d_J(c_n))$, for $\epsilon_n > 0$ to be determined shortly. Abbreviate s_ℓ for the distribution P_{g_ℓ}, $\ell = 1, \ldots, N$, and $\bar{s} = \frac{1}{N} \sum_{\ell=1}^N s_\ell$. Also, write $p(N(f(X), \sigma^2 I))$ for the density of a $N(f(X), \sigma^2 I)$ random variable, where $f(X) = (f(X^1), \ldots, f(X^n)) \in \mathbb{R}^n$. Then

$$\begin{aligned}
\frac{1}{M} \sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{q}) &\leq \frac{1}{M} \sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{s}) \\
&= \frac{1}{M} \sum_{j=1}^M \mathbb{E}_{X,f_j} \log \frac{p\big(N(f_j(X), \sigma^2 I)\big)}{\frac{1}{N} \sum_{\ell=1}^N p\big(N(g_\ell(X), \sigma^2 I)\big)} \\
&\leq \frac{1}{M} \sum_{j=1}^M \Big(\log N + \mathbb{E}_X \min_{\ell=1,\ldots,N} \mathrm{KL}(q_j \,\|\, s_\ell) \Big)
\end{aligned}$$

$$\leq \frac{1}{M}\sum_{j=1}^{M}\left(\log N + \frac{n\epsilon_n^2}{2\sigma^2}\right)$$

$$\leq K_2 c_n^w d^{1+w/2}\epsilon_n^{-w} + \frac{n\epsilon_n^2}{2\sigma^2}. \tag{B.44}$$

In the first line above, we used the fact that $\sum_{j=1}^M \mathrm{KL}(q_j \,\|\, \bar{q}) \leq \sum_{j=1}^M \mathrm{KL}(q_j \,\|\, s)$ for any other distribution s; in the second and third, we explicitly expressed and manipulated the definition of KL divergence; in the fourth, we used $\mathrm{KL}(q_j \,\|\, s_\ell) = \|f_j(X) - g_\ell(X)\|_2^2/(2\sigma^2)$, and for each j, there is at least one ℓ such that $\mathbb{E}_X\|f_j(X) - g_\ell(X)\|_2^2 = \|f_j - g_\ell\|_2^2 \leq \epsilon_n^2$; in the fifth line, we used the entropy bound from (B.40). Minimizing (B.44) over $\epsilon_n > 0$ gives

$$\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_X\mathrm{KL}(q_j \,\|\, \bar{q}) \leq \bar{K}_2 d n^{w/(2+w)} c_n^{2w/(2+w)},$$

for a constant $\bar{K}_2 > 0$. Returning to Fano's inequality (B.42), (B.43), we see that a lower bound on the minimax error is

$$\frac{\delta_n^2}{4}\left(1 - \frac{\bar{K}_2 d n^{w/(2+w)} c_n^{2w/(2+w)} + \log 2}{\log M}\right),$$

Therefore, a lower bound on the minimax error is $\delta_n^2/8$, for any $\delta_n > 0$ such that

$$\log M \geq 2\bar{K}_2 d n^{w/(2+w)} c_n^{2w/(2+w)} + 2\log 2,$$

and for large enough n, the first term on the right-hand side above will be larger than $2\log 2$, so it suffices to have

$$\log M \geq 4\bar{K}_2 d n^{w/(2+w)} c_n^{2w/(2+w)}. \tag{B.45}$$

Set $\delta_n = (\bar{K}_1/4\bar{K}_2)^{1/w}\sqrt{d}n^{-1/(2+w)}c_n^{w/(2+w)}$. Provided that $\delta_n \leq \bar{\delta}_1$, our log packing bound (B.41) is applicable, and ensures that (B.45) will be satisfied. This completes the proof.

B.15 Proof of Corollary 3.3

We only need to check Assumption C4 for $J(g) = \mathrm{TV}(g^{(k)})$, $w = 1/(k+1)$, and then we can apply Theorem 3.3. As before, the entropy bound upper bound is implied by results in Birman & Solomyak (1967) (see Mammen (1991) for an explanation and discussion). The packing number lower bound is verified as follows. For f a $(k+1)$ times weakly differentiable function on $[0,1]$,

$$\mathrm{TV}(f^{(k)}) = \int_0^1 |f^{(k+1)}(t)|\,dt \leq \left(\int_0^1 |f^{(k+1)}(t)|^2\,dt\right)^{1/2}.$$

Hence

$$\left\{f : \mathrm{TV}(f^{(k)}) \leq 1,\ \|f\|_\infty \leq 1\right\} \supseteq \left\{f : \int_0^1 |f^{(k+1)}(t)|^2\,dt \leq 1,\ \|f\|_\infty \leq 1\right\}.$$

Results in Kolmogorov & Tikhomirov (1959) imply that the space on the right-hand side satisfies the desired log packing number lower bound. This proves the result.

B.16 Proof of the linear smoother lower bound in (3.34)

We may assume without a loss of generality that each f_{0j}, $j = 1, \ldots, d$ has L_2 mean zero (since f_0 does). By the decomposability property of the L_2 norm over additive functions with L_2 mean zero components, as in (3.25), we have for any additive linear smoother $\hat{f} = \sum_{j=1}^d \hat{f}_j$,

$$\|\hat{f} - f_0\|_2^2 = \bigg(\sum_{j=1}^d \bar{f}_j\bigg)^2 + \sum_{j=1}^d \|(\hat{f}_j - \bar{f}_j) - f_{0j}\|_2^2$$

where $\bar{f}_j$ denotes the L_2 mean of $\hat{f}_j$, $j = 1, \ldots, d$. Note that the estimator $\hat{f}_j - \bar{f}_j$ is itself a linear smoother, for each $j = 1, \ldots, d$, since if we write $\hat{f}_j(x_j) = w_j(x_j)^T Y$ for a weight function w_j over $x_j \in [0,1]$, then $\hat{f}_j(x_j) - \bar{f}_j = \tilde{w}_j(x_j)^T Y$ for a weight function $\tilde{w}_j(x_j) = w_j(x_j) - \int_0^1 w_j(t)\,dt$. This, and the last display, imply that

$$\inf_{\hat{f}\,\text{additive linear}} \sup_{f_0 \in \mathcal{F}_k^d(c_n)} \mathbb{E}\|\hat{f} - f_0\|_2^2 = \sum_{j=1}^d \inf_{\hat{f}_j\,\text{linear}} \sup_{f_0 \in \mathcal{F}_k^d(c_n)} \mathbb{E}\|\hat{f}_j(Y) - f_{0j}\|_2^2. \tag{B.46}$$

Now fix an arbitrary $j = 1, \ldots, d$, and consider the jth term in the sum on the right-hand side above. Here we are looking at a linear smoother $\hat{f}_j$ fit to data

$$Y^i = \mu + f_{0j}(X_j^i) + \sum_{\ell \neq j} f_{0\ell}(X_\ell^i) + \epsilon^i, \quad i = 1, \ldots, n. \tag{B.47}$$

which depends on the components $f_{0\ell}$, for $\ell \neq j$. This is why the supremum in the jth term of the sum on the right-hand side in (B.46) must be taken over $f_0 \in \mathcal{F}_k^d(c_n)$, rather than $f_{0j} \in \mathcal{F}_k(c_n)$. Our notation $\hat{f}_j(Y)$ is used as a reminder to emphasize the dependence on the full data vector in (B.47).

A simple reformulation, by appropriate averaging over the lattice, helps untangle this supremum. Write $\hat{f}_j(x_j) = w_j(x_j)^T Y$ for a weight function w_j over $x_j \in [0,1]$, and for each $v = 1, \ldots, N$, let I_j^v be the set of indices i such that $X_j^i = v/N$. Also let

$$\bar{Y}_j^v = \frac{1}{N^{d-1}} \sum_{i \in I_j^v} Y^i, \quad v = 1, \ldots, N,$$

and $\bar{Y}_j = (\bar{Y}_j^1, \ldots, \bar{Y}_j^N) \in \mathbb{R}^N$. Then note that we can also write $\hat{f}_j(x_j) = \bar{w}_j(x_j)^T \bar{Y}_j$ for a suitably defined weight function $\bar{w}_j$, i.e., note that we can think of $\hat{f}_j$ as a linear smoother fit to data $\bar{Y}_j$, whose components follow the distribution

$$\bar{Y}_j^v = \mu_j + f_{0j}(v/N) + \bar{\epsilon}_j^v, \quad v = 1, \ldots, N, \tag{B.48}$$

where we let $\mu_j = \mu + \frac{1}{N}\sum_{\ell \neq j} \sum_{u=1}^n f_{0\ell}(u/N)$, and $\bar{\epsilon}_j^v$, $v = 1, \ldots, n$ are i.i.d. $N(0, \sigma^2/N^{d-1})$. Recalling that $f_{0j} \in \mathcal{F}_k(c_n)$, we are in a position to invoke univariate minimax results from Donoho & Johnstone (1998). As shown in Section 5.1 of Tibshirani (2014), the space $\mathcal{F}_k(c_n)$ contains the Besov space $B_{1,1}^{k+1}(c_n')$, for a radius c_n' that differs from c_n only by a constant factor. Therefore, by Theorem 1 of Donoho & Johnstone (1998) on the minimax risk of

linear smoothers fit to data from the model (B.48), we see that for large enough N and a constant $c_0 > 0$,

$$\begin{aligned}\inf_{\hat{f}_j \text{ linear}} \sup_{f_{0j} \in \mathcal{F}_k(c_n)} \mathbb{E}\|\hat{f}_j(\bar{Y}_j) - f_{0j}\|_2^2 &\geq c_0 (c_n N^{(d-1)/2})^{2/(2k+2)} \frac{N^{-(2k+1)/(2k+2)}}{N^{d-1}} \\ &= c_0 N^{-d(2k+1)/(2k+2)} c_n^{2/(2k+2)} \\ &= c_0 n^{-(2k+1)/(2k+2)} c_n^{2/(2k+2)}. \end{aligned} \tag{B.49}$$

As we have reduced the lower bound to the minimax risk of linear smoothers over a Besov ball, we can see that the same result (B.49) indeed holds simultaneously over all $j = 1, \ldots, d$. Combining this with (B.46) gives the desired result (3.34).

B.17 Proof of Theorem 3.4 and derivation details for Algorithm 2

We show that the dual of (3.37) is equivalent to the additive trend filtering problem (3.4), and further, the Lagrange multipliers corresponding to the constraints $u_0 = u_j$, for $j = 1, \ldots, d$, are equivalent to the primal variables θ_j, $j = 1, \ldots, d$. Let $M = I - \mathbb{1}\mathbb{1}^T/n$, and rewrite problem (3.37) as

$$\begin{aligned} \min_{u_0, u_1, \ldots, u_d \in \mathbb{R}^n} \quad & \frac{1}{2}\|MY - Mu_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) \\ \text{subject to} \quad & Mu_0 = Mu_1, \; Mu_0 = Mu_2, \; \ldots, \; Mu_0 = Mu_d, \end{aligned}$$

We can write the Lagrangian of this problem as

$$L(u_0, u_1, \ldots, u_d, \theta_1, \ldots, \theta_d) = \frac{1}{2}\|MY - Mu_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) + \sum_{j=1}^d \theta_j^T M(u_0 - u_j).$$

and we want to minimize this over $u_0, \ldots, u_d$ to form the dual of (3.37). This gives

$$\max_{\theta_1, \ldots, \theta_d \in \mathbb{R}^n} \; \frac{1}{2}\|MY\|_2^2 - \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\theta_j\Big\|_2^2 - \sum_{j=1}^d \Big(\max_{u_j \in U_j} \; u_j^T M\theta_j\Big). \tag{B.50}$$

We use the fact that the support function of U_j is just ℓ_1 penalty composed with $S_j D_j$ (invoking the duality between ℓ_∞ and ℓ_1 norms),

$$\max_{u_j \in U_j} u_j^T M\theta_j = \max_{\|v_j\|_\leq \lambda} v_j^T D_j S_j M\theta_j = \lambda \|D_j S_j M\theta_j\|_1,$$

where recall we abbreviate $D_j = D^{(X_j, k+1)}$, for $j = 1, \ldots, d$, and this allows us to rewrite the above problem (B.50) as

$$\min_{\theta_1, \ldots, \theta_d \in \mathbb{R}^n} \; \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\theta_j\Big\|_2^2 + \lambda \sum_{j=1}^d \|D_j S_j M\theta_j\|_1,$$

which is precisely the same as the original additive trend filtering problem (3.4).

This realization has important consequences. In the ADMM iterations (3.38), the scaled parameters $\rho\gamma_j$, $j = 1, \ldots, d$ correspond to dual variables θ_j, $j = 1, \ldots, d$ in problem (3.37), which from the above calculation, are precisely primal variables in (3.4). Under weak conditions, ADMM is known to produce convergent dual iterates, e.g., Section 3.2 of Boyd et al. (2011) shows that if (i) the criterion is a sum of closed, convex functions and (ii) strong duality holds, then the dual iterates from ADMM converge to optimal dual solutions. (Convergence of primal iterates requires stronger assumptions.) Our problem (3.37) satisfies these two conditions, and so for the ADMM algorithm outlined in (3.38), the scaled iterates $\rho\gamma_j^{(t)}$, $j = 1, \ldots, d$ converge to optimal solutions in the dual of (3.37), i.e., optimal solutions in the additive trend filtering problem (3.4). This proves the first part of the theorem.

As for the second part of the theorem, it remains to show that Algorithm 2 is equivalent to the ADMM iterations (3.36). This follows by notationally swapping γ_j, $j = 1, \ldots, d$ for θ_j/ρ, $j = 1, \ldots, d$, rewriting the updates

$$\theta_j^{(t)}/\rho = u_0^{(t)} + \theta_j^{(t-1)}/\rho - u_j^{(t)}, \quad j = 1, \ldots, d,$$

as

$$\theta_j^{(t)} = \rho \cdot \mathrm{TF}_\lambda\big(u_0^{(t)} + \theta_j^{(t-1)}/\rho, X_j\big), \quad j = 1, \ldots, d,$$

using (3.35), and lastly, eliminating u_j, $j = 1, \ldots, d$ from the u_0 update by solving for these variables in terms of terms of θ_j, $j = 1, \ldots, d$, i.e., by using

$$u_j^{(t-1)} = u_0^{(t-1)} + \theta_j^{(t-2)}/\rho - \theta_j^{(t-1)}/\rho, \quad j = 1, \ldots, d.$$

B.18 Cyclic versus parallel backfitting

We compare the performances of the usual cyclic backfitting method in Algorithm 1 to the parallel version in Algorithm 2, on a simulated data set generated as in Section 3.5.1, except with $n = 2000$ and $d = 24$. We computed the additive trend filtering estimate (3.4) (of quadratic order), at a fixed value of λ lying somewhere near the middle of the regularization path, by running the cyclic and parallel backfitting algorithms until each obtained a suboptimality of 10^{-8} in terms of the achieved criterion value (the optimal criterion value here was determined by running Algorithm 1 for a very large number of iterations). We used simply $\rho = 1$ in Algorithm 2.

Figure B.1 shows the progress of the two algorithms, plotting the suboptimality of the criterion value across the iterations. The two panels, left and right, differ in how iterations are counted for the parallel method. On the left, one full cycle of d component updates is counted as one iteration for the parallel method—this corresponds to running the parallel algorithm in "naive" serial mode, where each component update is actually performed in sequence. On the right, d full cycles of d component updates is counted as one iteration for the parallel method—this corresponds to running the parallel algorithm in an "ideal" parallel mode with d parallel processors. In both panels, one full cycle of d component updates is counted as one iteration for the cyclic method. We see that, if parallelization is fully utilized, the parallel method cuts down the iteration cost by about a factor of 2,

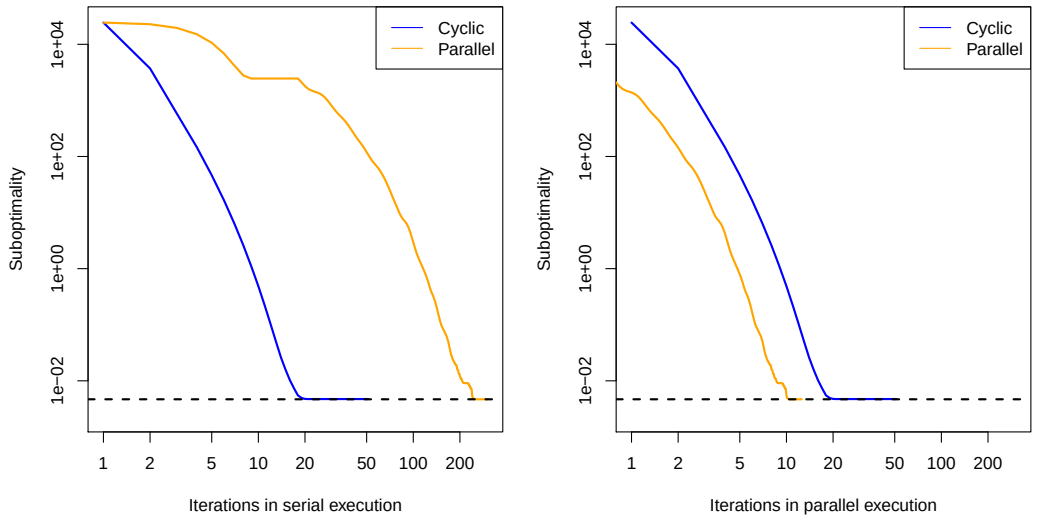

Figure B.1: *Suboptimality in criterion value versus iteration number for the cyclic (Algorithm 1) and parallel (Algorithm 2) backfitting methods, on a synthetic data set with* $n = 2000$ *and* $d = 24$*. On the left, iterations for the parallel method are counted as if "ideal" parallelization is used, where the* d *component updates are performed by* d *processors, at the total cost of one update, and on the right, iterations for the parallel method are counted as if "naive" serialization is used, where the component updates are performed in sequence. To avoid zeros on the y-axis (log scale), we added a small value to all the suboptimalities (dotted line).*

compared to the cyclic method. We should expect these computational gains to be even larger as the number of components d grows.

B.19 Simulated homogeneously-smooth data

Figure B.2 shows the results of a homogeneous simulation, as in Section 3.5.1 and Figure 3.4, except that for the base component trends we used sinusoids of equal (and spatially-constant) frequency:

$$g_{0j}(x_j) = \sin(10\pi x_i), \quad j = 1, \ldots, 10,$$

and we defined the component functions as $f_{0j} = a_j g_{0j} - b_j$, $j = 1, \ldots, d$, where a_j, b_j were chosen to standardize f_{0j} (give it zero empirical mean and unit empirical norm), for $j = 1, \ldots, d$.

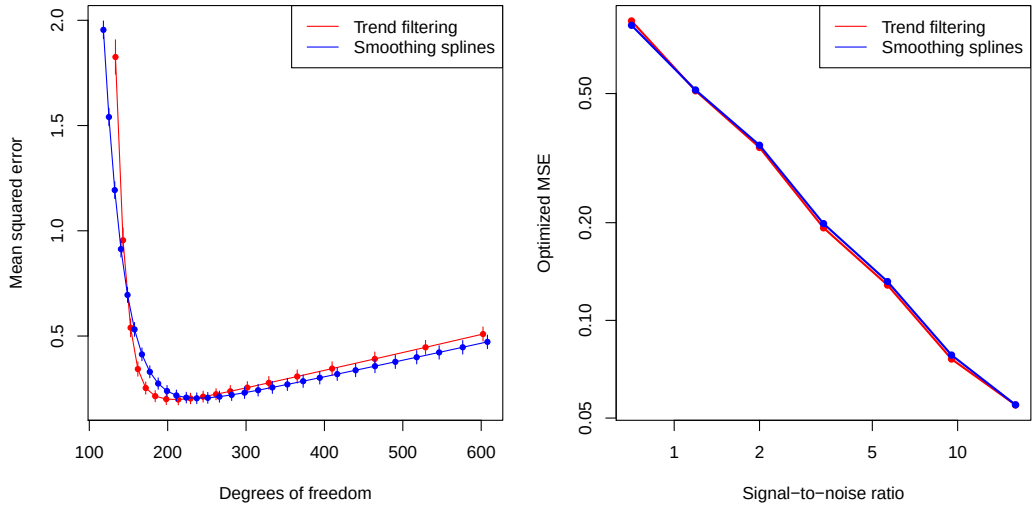

Figure B.2: *Results from a simulation setup identical to that described in Section 3.5.1, i.e., identical to that used to produce Figure 3.4, except with homogeneous smoothness in the underlying component functions.*

Appendix C

Appendix for Higher-Order Kolmogorov-Smirnov Test

C.0.1 Comparing the Test in Wang et al. (2014)

The test statistic in Wang et al. (2014) can be expressed as

$$T^{**} = \max_{t \in Z_{(N)}} |(\mathbb{P}_m - \mathbb{Q}_n) g_t^+| = \max_{t \in Z_{(N)}} \left| \frac{1}{m} \sum_{i=1}^{m} (x_i - t)_+^k - \frac{1}{n} \sum_{i=1}^{n} (y_i - t)_+^k \right|. \tag{C.1}$$

This is very close to our approximate statistic T^* in (4.9). The only difference is that we replace $g_t^+(x) = (x-t)_+^k$ by $g_t^-(x) = (t-x)_+^k$ for $t \le 0$.

Our exact (not approximate) statistic is in (4.6). This has the advantage having an equivalent variational form (4.5), and the latter form is important because it shows the statistic to be a metric.

C.1 Proof of Proposition 4.1

We first claim that $F(x) = |x|^k/k!$ is an envelope function for $\mathcal{F}_k$, meaning $f \le F$ for all $f \in \mathcal{F}_k$. To see this, note each $f \in \mathcal{F}_k$ has kth weak derivative with left or right limit of 0 at 0, so $|f^{(k)}(x)| \le \mathrm{TV}(f^{(k)}) \le 1$; repeatedly integrating and applying the derivative constraints yields the claim. Now due to the envelope function, if P, Q have k moments, then the IPM is well-defined: $|\mathbb{P}f| < \infty$, $|\mathbb{Q}f| < \infty$ for all $f \in \mathcal{F}_k$. Thus if $P = Q$, then clearly $\rho(P, Q; \mathcal{F}_k) = 0$.

For the other direction, suppose that $\rho(P, Q; \mathcal{F}_k) = 0$. By simple rescaling, for any f, if $\mathrm{TV}(f^{(k)}) = R > 0$, then $\mathrm{TV}((f/R)^{(k)}) \le 1$. Therefore $\rho(P, Q; \mathcal{F}_k) = 0$ implies $\rho(P, Q; \widetilde{\mathcal{F}}_k) = 0$, where

$$\widetilde{\mathcal{F}}_k = \big\{ f : \mathrm{TV}(f^{(k)}) < \infty,\ f^{(j)}(0) = 0,\ j \in \{0\} \cup [k-1],\ \text{and } f^{(k)}(0+) = 0 \text{ or } f^{(k)}(0-) = 0 \big\}.$$

This also implies $\rho(P, Q; \widetilde{\mathcal{F}}_k^+) = 0$, where

$$\widetilde{\mathcal{F}}_k^+ = \{ f : \mathrm{TV}(f^{(k)}) < \infty,\ f(x) = 0 \text{ for } x \le 0 \}.$$

As the class $\widetilde{\mathcal{F}}_k^+$ contains $C_c^\infty(\mathbb{R}_+)$, where $\mathbb{R}_+ = \{x : x > 0\}$ (and $C_c^\infty(\mathbb{R}_+)$ is the class of infinitely differentiable, compactly supported functions on $\mathbb{R}_+$), we have by Lemma C.1 that $P(A \cap \mathbb{R}_+) = Q(A \cap \mathbb{R}_+)$ for all open sets A. By similar arguments, we also get that $P(A \cap \mathbb{R}_-) = Q(A \cap \mathbb{R}_-)$, for all open sets A, where $\mathbb{R}_- = \{x : x < 0\}$. This implies that $P(\{0\}) = Q(\{0\})$ (as $1 - P(\mathbb{R}_+) - P(\mathbb{R}_-)$, and the same for Q), and finally, $P(A) = Q(A)$ for all open sets A, which means that $P = Q$.

C.2 Statement and Proof of Lemma C.1

Lemma C.1. For any two distributions P, Q supported on an open set Ω, if $\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{Y\sim Q}[f(Y)]$ for all $f \in C_c^\infty(\Omega)$, then $P = Q$.

Proof. It suffices to show that $P(A) = Q(A)$ for every open set $A \subseteq \Omega$. As P, Q are probability measures and hence Radon measures, there exists a sequence of compact sets $K_n \subseteq A$, $n = 1, 2, 3, \ldots$ such that $\lim_{n\to\infty} P(K_n) = P(A)$ and $\lim_{n\to\infty} Q(K_n) = Q(A)$. Let f_n, $n = 1, 2, 3, \ldots$ be smooth compactly supported functions with values in $[0, 1]$ such that $f_n = 1$ on K_n and $f_n = 0$ outside of A. (Such functions can be obtained by applying Urysohn's Lemma on appropriate sets containing K_n and A and convolving the resulting continuous function with a bump function.) Then $P(K_n) \le E_P(f_n) = E_Q(f_n) \le Q(A)$ (where the equality by the main assumption in the lemma). Taking $n \to \infty$ gives $P(A) \le Q(A)$. By reversing the roles of P, Q, we also get $Q(A) \le P(A)$. Thus $P(A) = Q(A)$. □

C.3 Proof of Theorem 4.1

Let $\mathcal{G}_k$ be as in (4.10). Noting that $G_k \subseteq \mathcal{F}_k$, it is sufficient to show

$$\sup_{f\in\mathcal{F}_k} |\mathbb{P}_m f - \mathbb{Q}_n f| \le \sup_{g\in\mathcal{G}_k} |\mathbb{P}_m g - \mathbb{Q}_n g|.$$

Fix any $f \in \mathcal{F}_k$. Denote $Z_{(N)}^0 = \{0\} \cup Z_{(N)}$. From the statement and proof of Theorem 1 in Mammen (1991), there exists a spline $\tilde{f}$ of degree k, with finite number of knots such that for all $z \in Z_{(N)}^0$

$$\begin{aligned} f(z) &= \tilde{f}(z), \\ f^{(j)}(z) &= \tilde{f}^{(j)}(z), \quad j \in [k-1], \\ f^{(k)}(z^+) &= \tilde{f}^{(k)}(z^+), \\ f^{(k)}(z^-) &= \tilde{f}^{(k)}(z^-). \end{aligned}$$

and importantly, $\mathrm{TV}(\tilde{f}^{(k)}) \le \mathrm{TV}(f^{(k)})$. As $0 \in Z_{(N)}^0$, we hence know that the boundary constraints (derivative conditions at 0) are met, and $\tilde{f} \in \mathcal{F}_k$.

Because $\tilde{f}$ is a spline with a given finite number of knot points, we know that it has an expansion in terms of truncated power functions. Write $t_0, t_1, \ldots, t_L$ for the knots of $\tilde{f}$, where $t_0 = 0$. Also denote $g_t = g_t^+$ when $t > 0$, and $g_t = g_t^-$ when $t < 0$. Then for some $\alpha_\ell \in \mathbb{R}$, $\ell \in \{0\} \cup [L]$, and a polynomial p of degree k, we have

$$\tilde{f} = p + \alpha_0 g_0^+ + \sum_{\ell=1}^{L} \alpha_\ell g_{t_\ell},$$

The boundary conditions on $\tilde{f}$, g_0^+, g_{t_ℓ}, $\ell \in [L]$ imply

$$p(0) = p^{(1)}(0) = \ldots = p^{(k-1)}(0) = 0,$$
$$(\alpha_0 g_{0^+} + p)^{(k)}(0^+) = 0 \;\text{ or }\; (\alpha_0 g_{0^+} + p)^{(k)}(0^-) = 0.$$

The second line above implies that

$$\alpha_0 + p^{(k)} = 0 \;\text{ or }\; p^{(k)} = 0.$$

In the second case, we have $p = 0$. In the first case, we have $p(x) = -\alpha_0 x^k/k!$, so $\alpha_0 g_0 + p = -(-1)^{k+1}\alpha_0 g_0^-$. Therefore, in all cases we can write

$$\tilde{f} = \sum_{\ell=0}^{L} \alpha_\ell g_{t_\ell},$$

with the new understanding that g_0 is either g_0^+ or g_0^-. This means that $\tilde{f}$ lies in the span of functions in $\mathcal{G}_k$. Furthermore, our last expression for $\tilde{f}$ implies

$$\|\alpha\|_1 = \sum_{\ell=0}^{L} |\alpha_\ell| = \mathrm{TV}(\tilde{f}^{(k)}) \le \mathrm{TV}(f^{(k)}) \le 1.$$

Finally, using the fact that f and $\tilde{f}$ agree on $Z^0_{(N)}$,

$$\begin{aligned}
|\mathbb{P}_m f - \mathbb{Q}_n f| &= |\mathbb{P}_m \tilde{f} - \mathbb{Q}_n \tilde{f}| \\
&= \left| \sum_{\ell=0}^{L} \alpha_\ell (\mathbb{P}_m g_{t_\ell} - \mathbb{Q}_n g_{t_\ell}) \right| \\
&\le \sum_{\ell=0}^{L} |\alpha_\ell| \cdot \sup_{g \in \mathcal{G}_k} |\mathbb{P}_m g - \mathbb{Q}_n g| \\
&\le \sup_{g \in \mathcal{G}_k} |\mathbb{P}_m g - \mathbb{Q}_n g|,
\end{aligned}$$

the last two lines following from Holder's inequality, and $\|\alpha\|_1 \le 1$. This completes the proof.

C.4 Proof of Proposition 4.3

From Shor (1998), Nesterov (2000), a polynomial of degree $2d$ is nonnegative on $\mathbb{R}$ if and only if it can be written as a sum of squares (SOS) of polynomials, each of degree d. Crucially, one can show that $p(x) = \sum_{i=0}^{2d} a_i x^i$ is SOS if and only if there is a positive semidefinite matrix $Q \in \mathbb{R}^{(d+1)\times(d+1)}$ such that

$$a_{i-1} = \sum_{j+k=i} Q_{jk}, \quad i \in [2d].$$

Finding such a matrix Q can be cast as a semidefinite program (SDP) (a feasibility program, to be precise), and therefore checking nonnegativity can be done by solving an SDP.

Furthermore, calculating the maximum of a polynomial p is equivalent to calculating the smallest γ such that $\gamma - p$ is nonnegative. This is therefore also an SDP.

Finally, a polynomial of degree k is nonnegative an interval $[a, b]$ if and only if it can be written as

$$p(x) = \begin{cases} s(x) + (x-a)(b-x)t(x) & k \text{ even} \\ (x-a)s(x) + (b-x)t(x) & k \text{ odd} \end{cases}, \tag{C.2}$$

where s, t are polynomials that are both SOS. Thus maximizing a polynomial over an interval is again equivalent to an SDP. For details, including a statement that such an SDP can be solved to ϵ-suboptimality in $c_k \log(1/\epsilon)$ iterations, where $c_k > 0$ is a constant that depends on k, see Nesterov (2000).

C.5 Proof of Lemma 4.2

Suppose t^* maximizes the criterion in (4.6). If $t^* = 0$, then $T^* = T$ and the result trivially holds. Assume without a loss of generality that $t^* > 0$, as the result for $t^* < 0$ will follow similarly.

If t^* is one of the sample points $Z_{(N)}$, then $T^* = T$ and the result trivially holds; if t^* is larger than all points in $Z_{(N)}$, then $T^* = T = 0$ and again the result trivially holds. Hence we can assume without a loss of generality that $t^* \in (a, b)$, where $a, b \in Z^0_{(N)}$. Define

$$\phi(t) = \frac{1}{k!} \sum_{i=1}^{N} c_i (z_i - t)_+^k, \quad t \in [a, b],$$

where $c_i = (\mathbb{1}_m/m - \mathbb{1}_n/n)_i$, $i \in [N]$, as before. Note that $T = \phi(t^*)$, and

$$|\phi'(t)| \leq \frac{1}{(k-1)!} \sum_{i=1}^{N} |c_i||z_i^{k-1}| = \frac{1}{(k-1)!} \left(\frac{1}{m} \sum_{i=1}^{m} |x_i|^{k-1} + \frac{1}{n} \sum_{i=1}^{n} |y_i|^{k-1} \right) := L.$$

Therefore

$$T - T^* \leq |f(t^*)| - |f(a)| \leq |f(t^*) - f(a)| \leq |t^* - a| L \leq \delta_N L,$$

as desired.

C.6 Proof of Lemma 4.3

Decompose $\mathcal{G}_k = \mathcal{G}_k^+ \cup \mathcal{G}_k^-$, where $\mathcal{G}_k^+ = \{g_t^+ : t \geq 0\}$, $\mathcal{G}_k^- = \{g_t^- : t \leq 0\}$. We will bound the bracketing number of $\mathcal{G}_k^+$, and the result for $\mathcal{G}_k^-$, and hence $\mathcal{G}_k$, follows similarly.

Our brackets for $\mathcal{G}_k^+$ will be of the form $[g_{t_i}, g_{t_{i+1}}]$, $i \in \{0\} \cup [R]$, where $0 = t_1 < t_2 < \cdots < t_{R+1} = \infty$ are to be specified, with the convention that $g_\infty = 0$. It is clear that such a set of brackets covers $\mathcal{G}_k^+$. Given $\epsilon > 0$, we need to choose the brackets such that

$$\|g_{t_i} - g_{t_{i+1}}\|_2 \leq \epsilon, \quad i \in \{0\} \cup [R], \tag{C.3}$$

and then show that the number of brackets R is small enough to satisfy the bound in the statement of the lemma.

For any $0 \le s < t$,

$$\begin{aligned} k!^2 \|g_s - g_t\|_2^2 &= \int_s^t (x-s)_+^{2k}\, dP(x) + \int_t^\infty \big((x-s)^k - (x-t)^k\big)^2\, dP(x) \\ &\le \int_s^\infty \big(k(x-s)^{k-1}(t-s)\big)^2\, dP(x) \\ &= k^2(t-s)^2 \int_s^\infty (x-s)^{2k-2}\, dP(x), \end{aligned}$$

where the second line follows from elementary algebra. Now in view of the moment bound assumption, we can bound the integral above using Holder's inequality with $p = (2k+\delta)/(2k-2)$ and $q = (2k+\delta)/(2+\delta)$ to get

$$\begin{aligned} k!^2 \|g_s - g_t\|_2^2 &\le k^2(t-s)^2 \bigg(\int_s^\infty (x-s)^{2k+\delta}\, dP(x) \bigg)^{1/p} \bigg(\int_s^\infty 1^q\,(x) dP \bigg)^{1/q} \\ &\le \frac{M^{1/p}}{(k-1)!^2}(t-s)^2, \end{aligned} \tag{C.4}$$

where recall the notation $M = \mathbb{E}[|X|^{2k+\delta}] < \infty$.

Also, for any $t > 0$, using Holder's inequality again, we have

$$\begin{aligned} k!^2 \|g_t - 0\|_2^2 &= \int_t^\infty (x-t)^{2k}\, dP(x) \\ &\le \bigg(\int_t^\infty (x-t)^{2k+\delta}\, dP(x) \bigg)^{2k/(2k+\delta)} \big(P(X \ge t)\big)^{\delta/(2k+\delta)} \\ &\le M^{2k/(2k+\delta)} \bigg(\frac{\mathbb{E}|X|^{2k+\delta)}}{t^{2k+\delta}} \bigg)^{\delta/(2k+\delta)} = \frac{M}{t^\delta}, \end{aligned} \tag{C.5}$$

where in the third line we used Markov's inequality.

Fix an $\epsilon > 0$. For parameters $\beta, R > 0$ to be determined, set $t_i = (i-1)\beta$ for $i \in [R]$ and $t_0 = 0$, $t_{R+1} = \infty$. Looking at (C.4), to meet (C.3), we see we can choose β such that

$$\frac{M^{1/p}}{(k-1)!^2}\beta^2 \le \epsilon^2.$$

Then for such a β, looking at (C.5), we see we can choose R such that

$$\frac{M}{k!^2((R-1)\beta)^\delta} \le \epsilon^2.$$

In other words, we can choose choose

$$\beta = \frac{(k-1)!}{M^{1/2p}}, \quad R = 1 + \left\lceil \frac{M^{1/2p+1/\delta}}{(k-1)!k!^{2/\delta}\epsilon^{2/\delta+1}} \right\rceil,$$

and (C.4), (C.5) imply that we have met (C.3). Therefore,

$$\log N_{[]}(\epsilon, \|\cdot\|, \mathcal{G}_k^+) \le \log R \le C \log \frac{M^{1+\frac{\delta(k-1)}{2k+\delta}}}{\epsilon^{2+\delta}},$$

where $C > 0$ depends only on k, δ.

C.7 Proof of Theorem 4.3

Once we have a finite bracketing integral for $\mathcal{G}_k$, we can simply apply Theorem 4.2 to get the result. Lemma 4.3 shows the log bracketing number of $\mathcal{G}_k$ to grow at the rate $\log(1/\epsilon)$, slow enough to imply a finite bracketing integral (the bracketing integral will be finite as long as the log bracketing number does not grow faster than $1/\epsilon^2$).

C.8 Proof of Corollaries 4.1 and 4.2

For the approximation from Proposition 4.3, observe

$$\sqrt{N}T_\epsilon = \sqrt{N}T + \sqrt{N}(T - T_\epsilon),$$

and $0 \leq \sqrt{N}(T - T_\epsilon) \leq \sqrt{N}\epsilon$, so for $\epsilon = o(1/\sqrt{N})$, we will have $\sqrt{N}T_\epsilon$ converging weakly to the same Gaussian process as $\sqrt{N}T$.

For the approximation in (4.9), the argument is similar, and we are simply invoking Lemma 5 in Wang et al. (2014) to bound the maximum gap δ_N in probability, under the density conditions.

C.9 Proof of Theorem 4.5

Let $W = \sqrt{m}\rho(P_m, P; \mathcal{G}_k)$. The bracketing integral of $\mathcal{G}_k$ is finite due to the slow growth of the log bracketing number from Lemma 4.3, at the rate $\log(1/\epsilon)$. Also, we can clearly take $F(x) = |x|^k/k!$ as an envelope function for $\mathcal{G}_k$. Thus, we can apply Theorem 4.5 to yield

$$\big(\mathbb{E}[\rho(P_m, P; \mathcal{G}_k)^p]\big)^{1/p} \leq \frac{C}{\sqrt{m}}$$

for a constant $C > 0$ depending only on k, p, and $\mathbb{E}|X|^p$. Combining this with Markov's inequality, for any a,

$$\mathbb{P}\big(\rho(P_m, P; \mathcal{G}_k) > a\big) \leq \left(\frac{C}{\sqrt{m}a}\right)^p,$$

thus for $a = C/(\sqrt{m}\alpha^{1/p})$, we have $\rho(P_m, P; \mathcal{G}_k) \leq a$ with probability at least $1 - \alpha$. The same argument applies to $W = \sqrt{n}\rho(Q_n, P; \mathcal{G}_k)$, and putting these together yields the result. The result when we additionally assume finite Orlicz norms is also similar.

C.10 Proof of Corollary 4.3

Let f maximize $|(\mathbb{P} - \mathbb{Q})f|$. Due to the moment conditions (see the proof of Proposition 4.1), we have $|\mathbb{P}f| < \infty$, $|\mathbb{Q}f| < \infty$. Assume without loss of generality that $(\mathbb{P} - \mathbb{Q})f > 0$. By the strong law of large numbers, we have $(\mathbb{P}_m - \mathbb{Q}_n)f \to (\mathbb{P} - \mathbb{Q})f$ as $m, n \to \infty$, almost surely. Also by the strong law, $\mathbb{P}_m|x|^{k-1} \to \mathbb{P}|x|^{k-1}$ as $m \to \infty$, almost surely, and $\mathbb{Q}_n|y|^{k-1} \to \mathbb{Q}|y|^{k-1}$ as $n \to \infty$, almost surely. For what follows, fix any samples $X_{(m)}, Y_{(n)}$ (i.e., take them to be nonrandom) such that the aforementioned convergences hold.

For each m, n, we know by the representer result in Theorem 4.1 that there exists $g_{mn} \in \mathcal{G}_k$ such that $(\mathbb{P}_m - \mathbb{Q}_n)f = |(\mathbb{P}_m - \mathbb{Q}_n)g_{mn}|$. (This is possible since the proof of

Theorem 4.1 does not rely on any randomness that is inherent to $X_{(m)}, Y_{(n)}$, and indeed it holds for any fixed sets of samples.) Assume again without a loss of generality that $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} > 0$. Denote by t_{mn} the knot of g_{mn} (i.e., $g_{mn} = g^+_{t_{mn}}$ if $t \geq 0$, and $g_{mn} = g^-_{t_{mn}}$ if $t \leq 0$). We now consider two cases.

If $|t_{mn}|$ is a bounded sequence, then by the Bolzano-Weierstrass theorem, it has a convergent subsequence, which converges say to $t \geq 0$. Passing to this subsequence (but keeping the notation unchanged, to avoid unnecessary clutter) we claim that $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} \to (\mathbb{P} - \mathbb{Q})g$ as $m, n \to \infty$, where $g = g^+_t$. To see this, assume $t_{mn} \geq t$ without a loss of generality (the arguments for $t_{mn} \leq t$ are similar), and note

$$g(x) - g_{mn}(x) = \begin{cases} 0 & x < t \\ (x-t)^k & t \leq x < t_{mn} \\ (t_{mn} - t)\sum_{i=0}^{k-1}(x-t)^i(x-t_{mn})^{k-1-i} & x \geq t_{mn} \end{cases},$$

where we have used the identity $a^k - b^k = (a-b)\sum_{i=0}^{k-1} a^i b^{k-1-i}$. Therefore, as $m, n \to \infty$,

$$|\mathbb{P}_m(g_{mn} - g)| \leq k|t_{mn} - t|\mathbb{P}_m|x|^{k-1} \to 0,$$

because $t_{mn} \to t$ by definition, and $\mathbb{P}_m|x|^{k-1} \to \mathbb{P}|x|^k$. Similarly, as $m, n \to \infty$, we have $|\mathbb{Q}_n(g_{mn} - g)| \to 0$, and therefore $|(\mathbb{P}_m - \mathbb{Q}_n)(g_{mn} - g)| \leq |\mathbb{P}_m(g_{mn} - g)| + |\mathbb{Q}_m(g_{mn} - g)| \to 0$, which proves the claim. But since $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} = (\mathbb{P}_m - \mathbb{Q}_n)f$ for each m, n, we must have $(\mathbb{P} - \mathbb{Q})g = (\mathbb{P} - \mathbb{Q})f$, i.e., there is a representer in $\mathcal{G}_k$, as desired.

If $|t_{mn}|$ is unbounded, then pass to a subsequence in which t_{mn} converges say to ∞ (the case for convergence to $-\infty$ is similar). In this case, we have $(\mathbb{P}_m - Q_n)g_{mn} \to 0$ as $m, n \to \infty$, and since $(\mathbb{P}_m - \mathbb{Q}_n)g_{mn} = (\mathbb{P}_m - \mathbb{Q}_n)f$ for each m, n, we have $(\mathbb{P} - \mathbb{Q})f = 0$. But we can achieve this with $(\mathbb{P} - \mathbb{Q})g^+_t$, by taking $\to \infty$, so again we have a representer in $\mathcal{G}_k$, as desired.

C.11 Proof of Corollary 4.4

When we reject as specified in the corollary, note that for $P = Q$, we have type I error at most α_N by Theorem 4.4, and as $\alpha_N = o(1)$, we have type I error converging to 0.

For $P \neq Q$, such that the moment conditions are met, we know by Corollary 4.3 that $\rho(P, Q; \mathcal{G}_k) \neq 0$. Recalling $1/\alpha_N = o(N^{p/2})$, we have as $N \to \infty$,

$$c(\alpha_N)\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) = \alpha^{-1/p}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) \to 0.$$

The concentration result from Theorem 4.5 shows that T will concentrate around $\rho(P, Q; \mathcal{G}_k) \neq 0$ with probability tending to 1, and thus we reject with probability tending to 1.

C.12 Additional Experiments

C.13 Local Density Differences Continued

Figure C.1 plots the densities used for the local density difference experiments, with the left panel corresponding to Figure 4.6, and the right panel to Figure 4.7.

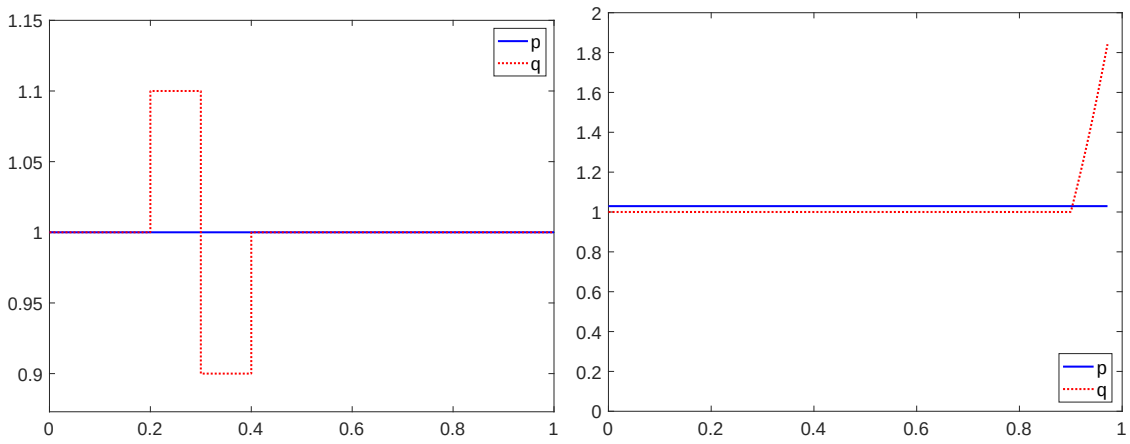

Figure C.1: *Densities for the local density difference experiments.*

C.14 Comparison to MMD with Polynomial Kernel

Now we compare the higher-order KS test to the MMD test with a polynomial kernel, as suggested by a referee of this paper. The MMD test with a polynomial kernel looks at moment differences up to some prespecified order $d \geq 1$, and its test statistic can be written as

$$\sum_{i=0}^{d} \binom{d}{i} (\mathbb{P}_n x^i - \mathbb{P}_m y^i)^2.$$

This looks at a weighted sum of *all* moments up to order d, whereas our higher-order KS test looks at truncated moments of a *single* order k. Therefore, to put the methods on more equal footing, we aggregated the higher-order KS test statistics up to order k, i.e., writing T_i to denote the ith order KS test statistic, $i \in [k]$, we considered

$$\sum_{i=0}^{k} \binom{k}{i} T_i^2,$$

borrowing the choice of weights from the MMD polynomial kernel test statistic.

Figure C.2 shows ROC curves from two experiments comparing the higher-order KS test and MMD polynomial kernel tests. We used distributions $P = N(0, 1)$, $Q = N(0.2, 1)$ in the left panel (as in Figure 4.4), and $P = N(0, 1)$, $Q = t(3)$ in the right panel (as in Figure 4.5). We can see that the (aggregated) higher-order KS tests and MMD polynomial kernel tests perform roughly similarly.

There is one important point to make clear: the population MMD test with a polynomial kernel is *not* a metric, i.e., there are distributions $P \neq Q$ for which the population-level test statistic is exactly 0. This is because it only considers moment differences up to order d, thus any pair of distributions P, Q that match in the first d moments but differ in (say) the $(d + 1)$st will lead to a population-level statistic that 0. In this sense, the MMD test with a polynomial kernel is not truly nonparametric, whereas the KS test, the higher-order KS tests the MMD test with a Gaussian kernel, the energy distance test, the Anderson-Darling test, etc., all are.

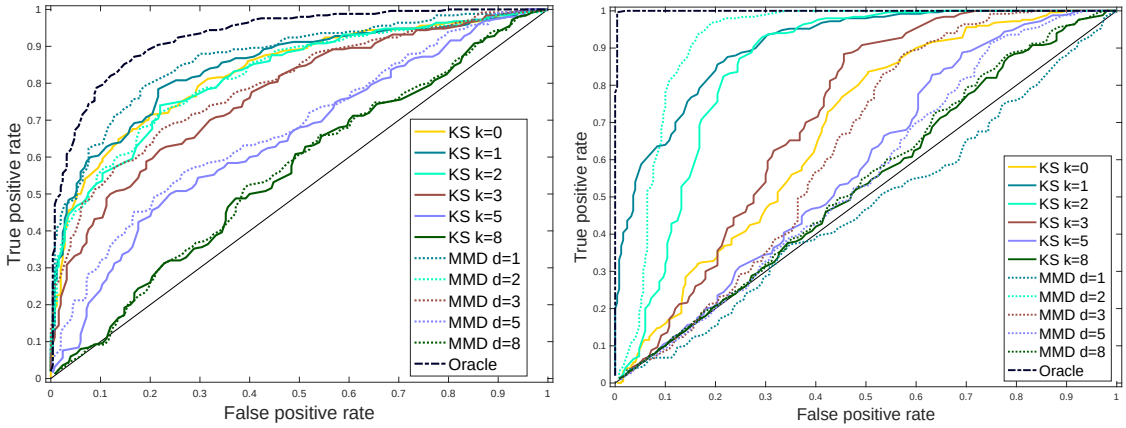

Figure C.2: *ROC curves for* $P = N(0,1)$, $Q = N(0.2,1)$ *(left), and* $P = N(0,1)$, $Q = t(3)$ *(right).*

C.15 Proof of Proposition 4.4

For $k \geq 1$, recall our definition of I^k the kth order integral operator,

$$(I^k f)(x) = \int_0^x \int_0^{t_k} \cdots \int_0^{t_2} f(t_1)\, dt_1 \cdots dt_k,$$

Further, for $k \geq 1$, denote by D^k the kth order derivative operator,

$$(D^k f)(x) = f^{(k)}(x),$$

Is it not hard to check that over all functions f with k weak derivatives, and that obey the boundary conditions $f(0) = f'(0) = \cdots = f^{(k-1)}(0) = 0$, these two operators act as inverses, in that

$$D^k I^k f = f, \quad \text{and} \quad I^k D^k f = f.$$

For a measure μ, denote $\langle f, d\mu \rangle = \int f(x)\, d\mu(x)$. (This is somewhat of an abuse of the notation for the usual L_2 inner product on square integrable functions, but it is convenient for what follows.) With this notation, we can write the kth order KS test statistic, at the population-level, as

$$\begin{aligned}
\sup_{f \in \mathcal{F}_k} |\mathbb{P}f - \mathbb{Q}f| &= \sup_{f \in \mathcal{F}_k} |\langle f, dP - dQ \rangle| \\
&= \sup_{f \in \mathcal{F}_k} |\langle I^k D^k f, dP - dQ \rangle| \\
&= \sup_{\substack{h:\mathrm{TV}(h) \leq 1, \\ h(0+)=0 \text{ or } h(0-)=0}} |\langle I^k h, dP - dQ \rangle| \\
&= \sup_{\substack{h:\mathrm{TV}(h) \leq 1, \\ h(0+)=0 \text{ or } h(0-)=0}} |\langle h, (I^k)^*(dP - dQ) \rangle| \\
&= \|(I^1)^* (I^k)^* (dP - dQ)\|_\infty. \qquad \text{(C.6)}
\end{aligned}$$

In the second line, we used the fact that I^k and D^k act as inverses over $f \in \mathcal{F}_k$ because these functions all satisfy the appropriate boundary conditions. In the third line, we simply reparametrized via $h = f^{(k)}$. In the fourth line, we introduced the adjoint operator $(I^k)^*$ of I^k (which will be described in detail shortly). In the fifth line, we leveraged the variational result for the KS test ($k = 0$ case), where $(I^1)^*$ denotes the adjoint of the integral operator I^1 (details below), and we note that the limit condition at 0 does do not affect the result here.

We will now study the adjoints corresponding to the integral operators. By definition $(I^1)^* g$ must satisfy for all functions f

$$\langle I^1 f, g\rangle = \langle f, (I^1)^* g\rangle.$$

We can rewrite this as

$$\int \int_0^x f(t)g(x)\,dt\,dx = \int f(t)((I^1)^* g)(t)\,dt,$$

and we can recognize by Fubini's theorem that therefore

$$((I^1)^* g)(t) = \begin{cases} \int_t^\infty g(x)\,dx & t \geq 0 \\ -\int_{-\infty}^t g(x)\,dx & t < 0. \end{cases}$$

For functions g that integrate to 0, this simplifies to

$$((I^1)^* g)(t) = \int_t^\infty g(x)dx, \quad t \in \mathbb{R}. \tag{C.7}$$

Returning to (C.6), because we can decompose $I^k = I^1 I^1 \cdots I^1$ (k times composition), it follows that $(I^k)^* = (I^1)^*(I^1)^* \cdots (I^1)^*$ (k times composition), so

$$\|(I^1)^*(I^k)^*(dP - dQ)\|_\infty = \|(I^k)^*(I^1)^*(dP - dQ)\|_\infty = \|(I^k)^*(F_P - F_Q)\|_\infty,$$

where in the last step we used (C.7), as $dP - dQ$ integrates to 0. This proves the first result in the proposition.

To prove the second result, we will show that

$$(I^k)^*(F_P - F_Q)(x) = \int_x^\infty \int_{t_k}^\infty \cdots \int_{t_2}^\infty (F_P - F_Q)(t_1)\,dt_1 \cdots dt_k,$$

when P, Q has nonnegative supports, or have k matching moments. In the first case, the above representation is clear from the definition of the adjoint. In the second case, we proceed by induction on k. For $k = 1$, note that $F_P - F_Q$ integrates to 0, which is true because

$$\langle 1, F_P - F_Q\rangle = \langle 1, (I^1)^*(dP - dQ)\rangle = \langle x, dP - dQ\rangle = 0,$$

the last step using the fact that P, Q have matching first moment. Thus, as $F_P - F_Q$ integrates to 0, we can use (C.7) to see that

$$(I^1)^*(F_P - F_Q)(x) = \int_x^\infty (F_P - F_Q)(t)\,dt.$$

Assume the result holds for $k-1$. We claim that $(I^{k-1})^*(F_P - F_Q)$ integrates to 0, which is true as

$$\langle 1, (I^{k-1})^*(F_P - F_Q)\rangle = \langle 1, (I^k)^*(dP - dQ)\rangle = \langle x^k/k!, dP - dQ\rangle = 0,$$

the last step using the fact that P, Q have matching kth moment. Hence, as $(I^{k-1})^*(F_P - F_Q)$ integrates to 0, we can use (C.7) and conclude that

$$\begin{aligned}(I^k)^*(F_P - F_Q)(x) &= (I^1)^*(I^{k-1})^*(F_P - F_Q)(x)\\ &= \int_x^\infty (I^{k-1})^*(F_P - F_Q)(t)\,dt\\ &= \int_x^\infty \int_{t_k}^\infty \cdots \int_{t_2}^\infty (F_P - F_Q)(t_1)\,dt_1 \cdots dt_k,\end{aligned}$$

where in the last step we used the inductive hypothesis. This completes the proof.