Additive Models with Trend Filtering

Veeranjaneyulu Sadhanala Ryan J. Tibshirani

Abstract

We consider additive models built with trend filtering, i.e., additive models whose components are each regularized by the (discrete) total variation of their $(k+1)$st (discrete) derivative, for a chosen integer $k \geq 0$. This results in kth degree piecewise polynomial components, (e.g., $k = 0$ gives piecewise constant components, $k = 1$ gives piecewise linear, $k = 2$ gives piecewise quadratic, etc.). In univariate nonparametric regression, the localized nature of the total variation regularizer used by trend filtering has been shown to produce estimates with superior local adaptivity to those from smoothing splines (and linear smoothers, more generally) (Tibshirani 2014). Further, the structured nature of this regularizer has been shown to lead to highly efficient computational routines for trend filtering (Kim et al. 2009, Ramdas & Tibshirani 2016). In this paper, we argue that both of these properties carry over to the additive models setting. We derive fast error rates for additive trend filtering estimates, and prove that these rates are minimax optimal when the underlying function is itself additive and has component functions whose derivatives are of bounded variation. We show that such rates are unattainable by additive smoothing splines (and by additive models built from linear smoothers, in general). We argue that backfitting provides an efficient algorithm for additive trend filtering, as it is built around the fast univariate trend filtering solvers; moreover, we describe a modified backfitting procedure whose iterations can be run in parallel. Finally, we conduct experiments to examine the empirical properties of additive trend filtering, and outline some possible extensions.

1 Introduction

Nonparametric regression becomes notoriously difficult as the dimension of the input space grows. In this work, we adopt the stance taken by many other authors, and consider an *additive model* for responses $Y^i \in \mathbb{R}$, $i = 1, \dots, n$ and input points $X^i = (X^i_1, \dots, X^i_d) \in \mathbb{R}^d$, $i = 1, \dots, n$, of the form

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

where $\mu \in \mathbb{R}$ is an overall mean parameter, each f_{0j} is a univariate function with $\sum_{i=1}^n f_{0j}(X^i_j) = 0$ 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^i_j denotes the jth component of the ith input point. (Exceptions will be made, e.g., for functions we will always use subscripts, but the role of the index should in general 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 are studied in Tibshirani (1983) and Hastie (1983), respectively. Some of the first asymptotic theory for additive models is 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) consider a formulation that uses fused lasso (i.e., total variation) penalization applied to the component functions.

The literature on additive models (and by now, sparse additive models) is vast and the above is far form 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.

1.1 Review: additive models and linear smoothers

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

$$\begin{aligned} (\hat{\theta}_1, \ldots, \hat{\theta}_d) = \underset{\theta_1,\ldots,\theta_d \in \mathbb{R}^n}{\operatorname{argmin}} \quad & \Big\| Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j \Big\|_2^2 + \lambda \sum_{j=1}^d \theta_j^T Q_j \theta_j \\ \text{subject to} \quad & \mathbb{1}^T \theta_j = 0, \quad j = 1, \ldots, d, \end{aligned} \tag{1}$$

where $Y = (Y^1, \ldots, Y^n) \in \mathbb{R}^n$ 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 $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^n$ the vector of all 1s. Each vector $\theta_j = (\theta_j^1, \ldots, \theta_j^n) \in \mathbb{R}^n$ represents the evaluations of the jth component function f_j in our model, i.e., tied together by the relationship

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

In the problem (1), $\lambda \geq 0$ is a regularization parameter and Q_j, $j = 1, \ldots, d$ 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 $j = 1, \ldots, d$. Under this choice, a backfitting (block coordinate descent) routine for (1) would repeatedly cycle through the updates

$$\theta_j = (I + \lambda Q_j)^{-1} \Big(Y - \bar{Y}\mathbb{1} - \sum_{\ell \neq j} \theta_\ell \Big), \quad j = 1, \ldots, d, \tag{2}$$

where the jth update fits a smoothing spline to the jth partial residual, over the jth dimension of the input points X_j. At convergence, we arrive at an additive smoothing spline estimate, which is the solution in (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`. As Buja et al. (1989) explain, the backfitting steps in (2) suggest that a more algorithmic approach to additive modeling can be taken. Instead of starting with a particular criterion in mind, as in (2), one can instead envision repeatedly cycling through updates

$$\theta_j = S_j \Big(Y - \bar{Y}\mathbb{1} - \sum_{\ell \neq j} \theta_\ell \Big), \quad j = 1, \ldots, d, \tag{3}$$

where each S_j is a particular (user-chosen) *linear smoother*, i.e., a linear operator that performs a univariate smoothing across the jth dimension of inputs X_j. The linear smoothers S_j, $j = 1, \dots, d$ could correspond to, e.g., 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. The convergence point of the iterations (3) solves a problem of the form (1) with $\lambda Q_j = S_j^+ - I$, where S_j^+ is the Moore-Penrose pseudoinverse of S_j, for $j = 1, \dots, d$.

The class of linear smoothers is broad enough to offer fairly flexible, interesting mechanisms for smoothing, and simple enough to understand precisely. Buja et al. (1989) provide a unified analysis of additive models with linear smoothers, $\lambda Q_j = S_j^+ - I$, for $j = 1, \dots, d$, in which they describe the effective degrees of freedom of the resulting estimators and a generalized cross-validation routine for tuning; they also study fundamental properties such as uniqueness of the component fits, and convergence of the backfitting steps.

Much of the work following Buja et al. (1989) remains in keeping with the idea of using linear smoothers in combination with additive models. Studying high-dimensional additive models, Lin & Zhang (2006), Ravikumar et al. (2009), Meier et al. (2009), Koltchinskii & Yuan (2010), Raskutti et al. (2012) all essentially build their methods off of linear smoothers, with modifications to induce sparsity in the component functions. Lin & Zhang (2006), Meier et al. (2009), Koltchinskii & Yuan (2010), Raskutti et al. (2012) write down sparsified versions of additive criteria that are similar to (1), while Ravikumar et al. (2009) introduce a sparsified version of the backfitting algorithm in (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 is constrained to lie in a function class

$$f_0 \in \mathcal{F}_k(C_0) = \Big\{ f : [0,1] \to \mathbb{R} : \mathrm{TV}(f^{(k)}) \le C_0 \Big\}, \tag{4}$$

for a constant $C_0 > 0$, where $\mathrm{TV}(\cdot)$ is the total variation operator, and $f^{(k)}$ the kth weak derivative of f. The class in (4) 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 $\mathcal{F}_k(C_0)$ 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., when $k = 0$ and f_0 is of bounded variation, this is a difference of $n^{-2/3}$ (optimal) versus $n^{-1/2}$ (optimal among linear smoothers).

It is important to emphasize that this shortcoming is not just a theoretical one; it is also clearly noticeable in basic practical examples. This does not bode well for additive models built from linear smoothers, when estimating component functions f_{0j}, $j = 1, \dots, d$ that display locally heterogeneous degrees of smoothness. 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 (or one) of the components f_{0j}, $j = 1, \dots, d$ 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 (4), 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 setting (especially for wavelets), but far fewer authors have considered using these locally adaptive estimators in the additive models context. Notable exceptions are Sardy & Tseng (2004), who study additive models built from wavelets, and Petersen et al. (2016), who study sparse additive models with components given by 0th order locally adaptive regression splines (equivalently, the components are regularized via fused lasso penalties, i.e., total variation penalties). The latter work is especially related to our focus in this paper.

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. Proposed independently by Steidl et al. (2006) and Kim et al. (2009), trend filtering is a relatively new approach to univariate nonparametric regression. As argued in Tibshirani (2014), it can be seen as a discrete-time analog of the locally adaptive regression spline estimator. Given responses $Y = (Y^1, \dots, Y^n)$ associated with sorted univariate inputs $X^1 < \dots < X^n \in \mathbb{R}$, the trend filtering estimate of order $k \geq 0$ is defined by the optimization problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \; \frac{1}{2}\|Y - \theta\|_2^2 + \lambda \|D^{(X,k+1)}\theta\|_1, \tag{5}$$

where $\lambda \geq 0$ is a regularization parameter, and $D^{(X,k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is a particular kth order discrete difference operator, constructed over the input points $X = (X^1, \dots, X^n)$. These operators are defined recursively, as in

$$D^{(X,1)} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}, \tag{6}$$

$$D^{(X,k+1)} = D^{(X,1)} \cdot \mathrm{diag}\left(\frac{k}{X^k - X^1}, \dots, \frac{k}{X^n - X^{n-k+1}}\right) \cdot D^{(X,k)} \in \mathbb{R}^{(n-k-1)\times n}, \quad k = 1, 2, 3, \dots. \tag{7}$$

(The leading matrix $D^{(X,1)}$ in (7) is the $(n-k-1) \times (n-k)$ version of the difference operator in (6).) Intuitively, the interpretation is that the problem (5) penalizes the sum of absolute $(k+1)$st order discrete derivatives of $\theta^1, \dots, \theta^n$ across the input points $X^1, \dots, X^n$. Thus, at optimality, the trend filtering solution has coordinates $\hat{\theta}^1, \dots, \hat{\theta}^n$ that obey a kth order piecewise polynomial form.

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 (4), 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 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 discrete difference operators in (6), (7), which means that specially implemented convex programming routines can solve (5) in an efficient manner.

It is this computational efficiency, along with its capacity for local adaptivity, that 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, defined as

$$\begin{aligned}(\hat{\theta}_1, \ldots, \hat{\theta}_d) = \underset{\theta_1,\ldots,\theta_d \in \mathbb{R}^n}{\mathrm{argmin}} \quad & \frac{1}{2}\Big\| Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j \Big\|_2^2 + \lambda \sum_{j=1}^d \big\| D^{(X_j,k+1)} S_j \theta_j \big\|_1 \\ \text{subject to} \quad & \mathbb{1}^T \theta_j = 0, \quad j = 1, \ldots, d.\end{aligned} \tag{8}$$

As before, $Y - \bar{Y}\mathbb{1}$ is the centered response vector, and $\lambda \geq 0$ is a regularization parameter. Not to be confused with the notation for linear smoothers from a previous subsection, $S_j \in \mathbb{R}^{n\times n}$ in (8) 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.,

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

The matrix $D^{(X_j,k+1)}$ in problem (8) is the $(k+1)$st order discrete difference operator, as in (6), (7), 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 the solution in (8) is still quite efficient, as we can leverage the efficient routines for univariate trend filtering.

Figure 1 gives a simple simulated example that compares the additive trend filtering estimates in (8) to the additive smoothing spline estimates in (1) (where the penalty matrices are chosen to be the Reinsch smoothing matrices along each dimension). Figure 2 shows the mean squared errors between the underlying regression function and the estimates from the two methods. Details are given in the figure captions. The takeaway message is that, for the regression function chosen in the simulation—whose components f_{01}, f_{02}, f_{03} display different levels of smoothness, and whose first component function f_{01} itself displays heterogeneous smoothness across its domain—the estimates from additive trend filtering are able to adapt better both visually and empirically (i.e., in terms of mean squared error) than those from additive smoothing splines.

1.4 Summary of contributions

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

- In Section 2, we develop some 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, we introduce two estimators related to additive trend filtering, based on splines, that facilitate theoretical analysis (and are perhaps of interest in their own right).
- In Section 4, we derive error bounds for additive trend filtering estimates. Assuming that the underlying regression function is itself additive, denoted $f_0 = \sum_{j=1}^d f_{0j}$, and that $\mathrm{TV}(f_{0j}^{(k)})$ is bounded, for each $j = 1, \ldots, d$, we prove that the kth order additive trend filtering estimate converges to f_0 at the rate $dn^{-(2k+2)/(2k+3)}$ (in both the squared empirical norm and squared L_2 norm). We also establish that this is the minimax rate over such a class of functions for f_0, and moreover, that additive smoothing spline estimates (more generally, additive models built from linear smoothers of any kind) achieve a rate of at best $dn^{-(2k+1)/(2k+2)}$ over this class.
- In Section 5, 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, provably convergent, parallelized backfitting algorithm for additive trend filtering.
- In Section 6, we present empirical experiments and comparisons, and in Section 7, we briefly describe extensions of the given additive trend filtering model and conclude with a discussion.

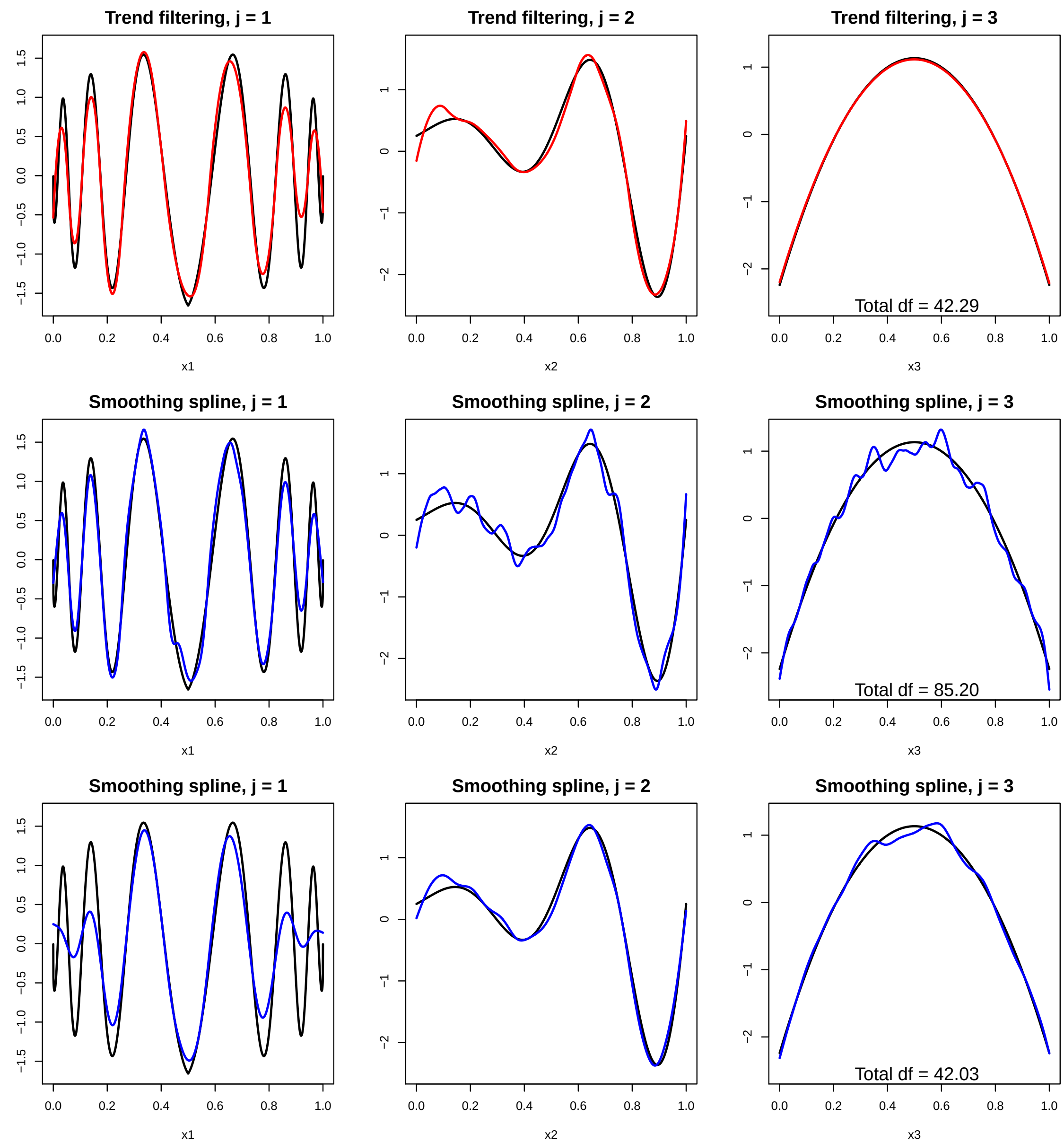

Figure 1: *A comparison of additive trend filtering and additive smoothing spline estimates, for a simulation with* $n = 3000$ *and* $d = 3$. *Three underlying components functions* f_{01}, f_{02}, f_{03} *were designed to have different levels of smoothness:* f_{03} *is the smoothest,* f_{02} *is less smooth, and* f_{01} *is the least smooth, and itself displays different amounts of smoothness at different spatial locations. The component functions are plotted in black, in each of the rows above. We drew input points* $X^i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1]^3$, $i = 1, \ldots, 3000$, *and we 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 make the signal-to-noise ratio about 1. We then computed estimates using additive trend filtering* (8) *(of quadratic order,* $k = 2$*) and additive smoothing splines* (1) *(of cubic order), each over their own sequences of tuning parameter values* λ. *The first row above shows the trend filtering components, at value of* λ *chosen to minimize mean squared error (MSE), computed over 20 repetitions (see Figure 2). The second row displays the smoothing spline components, again at a value of* λ *that minimizes 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 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. To put the methods on more equal footing, we also display in the third row above the smoothing spline components when the total estimate has about 42 df. Now the first component fit is oversmoothed, yet the third is still undersmoothed.*

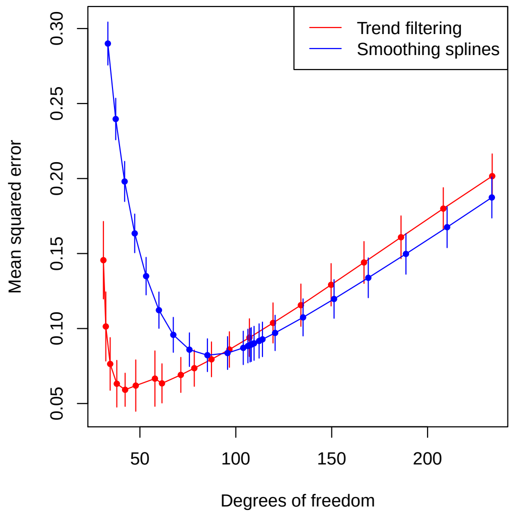

Figure 2: *MSE curves for additive trend filtering and additive smoothing splines, computed over 20 repetitions from the simulation setup of Figure 1. Vertical segments denote ± 1 standard deviations. The MSE curves are parametrized by effective degrees of freedom (df), estimated via standard Monte Carlo methods over the 20 repetitions. We see that trend filtering not only achieves a lower MSE, but its MSE curve is optimized at a lower df (i.e., less complex model) than that for smoothing splines.*

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).

2.1 Falling factorial representation

We may describe additive trend filtering in (8) as an estimation problem written in *analysis form*. The components are modeled directly by the parameters θ_j, $j = 1, \ldots, d$, and the desired structure is established by regularizing the (discrete) derivatives of these parameters, through the penalty terms $\|D^{(X_j,k+1)} S_j \theta_j\|_1$, $j = 1, \ldots, d$. Here, we present an alternative representation for (8) 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.

Before we derive the basis formulation that underlies additive trend filtering, we must first recall the *falling factorial basis* (Tibshirani 2014, Wang et al. 2014). 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

$$
\begin{aligned}
h_i(t) &= \prod_{\ell=1}^{i-1} (t - t^\ell), \quad i = 1, \ldots, k+1, \\
h_{i+k+1}(t) &= \prod_{\ell=1}^{k} (t - t^{i+\ell}) \cdot 1\{t > t^{i+k}\}, \quad i = 1, \ldots, n-k-1.
\end{aligned}
\tag{9}
$$

We denote $1\{t > a\} = 1$ when $t > a$, and 0 otherwise. (Also, 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 kth order truncated power basis functions. In fact, when $k = 0$ or $k = 1$, the two bases are exactly equivalent (meaning, they span the same space). 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 (9) are not splines, and when g (as defined above) has a knot at a particular location, it experiences a change in all of its lower order derivatives (i.e., all derivatives of orders $1, \ldots, k-1$).

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. As we show next, the analogous result holds for additive trend filtering.

Lemma 1 (**Falling factorial representation**). *For $j = 1, \ldots, d$, let $h_1^{(X_j)}, \ldots, h_n^{(X_j)}$ be the falling factorial basis in (9) with knots $(t^1, \ldots, t^n) = S_j X_j$, the jth dimension of the input points, properly sorted. Then the additive trend filtering problem in (8) is equivalent to the problem*

$$\begin{aligned} \hat{\alpha} = \underset{\alpha_1,\ldots,\alpha_d \in \mathbb{R}^n}{\operatorname{argmin}} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}(X_j^i) \Big)^2 + \lambda k! \sum_{j=1}^d \sum_{\ell=k+2}^n |\alpha_j^\ell| \\ \text{subject to} \quad & \sum_{i=1}^n \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}(X_j^i) = 0, \quad j = 1, \ldots, d, \end{aligned} \tag{10}$$

in that, at the solutions in (8), (10), we have

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

An alternative way of expressing problem (10) is

$$\begin{aligned} (\hat{f}_1, \ldots, \hat{f}_d) = \underset{f_j \in \mathcal{H}_j,\, j=1,\ldots,d}{\operatorname{argmin}} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \Big)^2 + \lambda \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d, \end{aligned} \tag{11}$$

where $\mathcal{H}_j = \operatorname{span}\{h_1^{(X_j)}, \ldots, h_n^{(X_j)}\}$ is the span of the falling factorial basis over the jth dimension, and $f_j^{(k)}$ is the kth weak derivative of f_j, for $j = 1, \ldots, d$. In this form, at the solutions in (8), (11),

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

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

$$H_{i\ell}^{(X_j,k)} = h_\ell^{(X_j)}(X_j^i), \quad i = 1, \ldots, n, \ \ell = 1, \ldots, n. \tag{12}$$

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.,

$$(H^{(X_j,k)})^{-1} = \begin{bmatrix} C^{(X_j,k+1)} \\ \frac{1}{k!} D^{(X_j,k+1)} \end{bmatrix} S_j. \tag{13}$$

Problem (10) is given by reparameterizing (8) according to $\theta_j = H^{(X_j,k)}\alpha_j$, for $j = 1,\dots,d$. As for (11), the equivalence between this and (10) 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 $j = 1,\dots,d$. □

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 $\hat{\theta} = (\hat{\theta}_1,\dots,\hat{\theta}_d)$ in (8), we can extend each component fit $\hat{\theta}_j$ to the real line, by forming an appropriate linear combination of falling factorial functions:

$$\hat{f}_j(x_j) = \sum_{\ell=1}^n \hat{\alpha}_j^\ell h_\ell^{(X_j)}(x_j), \quad x_j \in \mathbb{R}. \tag{14}$$

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 (13). 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

$$\hat{a}_j = C^{(X_j,k+1)} S_j \hat{\theta}_j, \tag{15}$$

$$\hat{b}_j = \frac{1}{k!} D^{(X_j,k+1)} S_j \hat{\theta}_j. \tag{16}$$

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 A.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 underlying fitted function $\hat{f}_j$ exhibits a knot at the $(\ell+k)$th sorted input point among $S_j X_j$, i.e., at $X_j^{(\ell+k)}$. See Figure 3 for an example.

We note that the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$ in (15), (16) can be computed in $O(n)$ operations and $O(1)$ memory. This makes extrapolation of the jth fitted function $\hat{f}_j$ in (14) highly efficient. Details are given in Appendix A.1.

2.2 Uniqueness of component fits

It is easy to see that, for the problem (8), 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 in (8), implies a unique additive fit $T\hat{\theta}$ at the solution $\hat{\theta} = (\hat{\theta}_1,\dots,\hat{\theta}_d) \in \mathbb{R}^{nd}$.

On the other hand, when $d > 1$, the criterion in (8) 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 points $X = [X_1 \cdots X_d] \in \mathbb{R}^{n\times d}$. 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 $\{\pm A_i : i \neq i_1,\dots,i_{\ell+1}\}$. Informally, if the columns of A are not in general position, then there must be some small subset of columns that are affinely dependent.

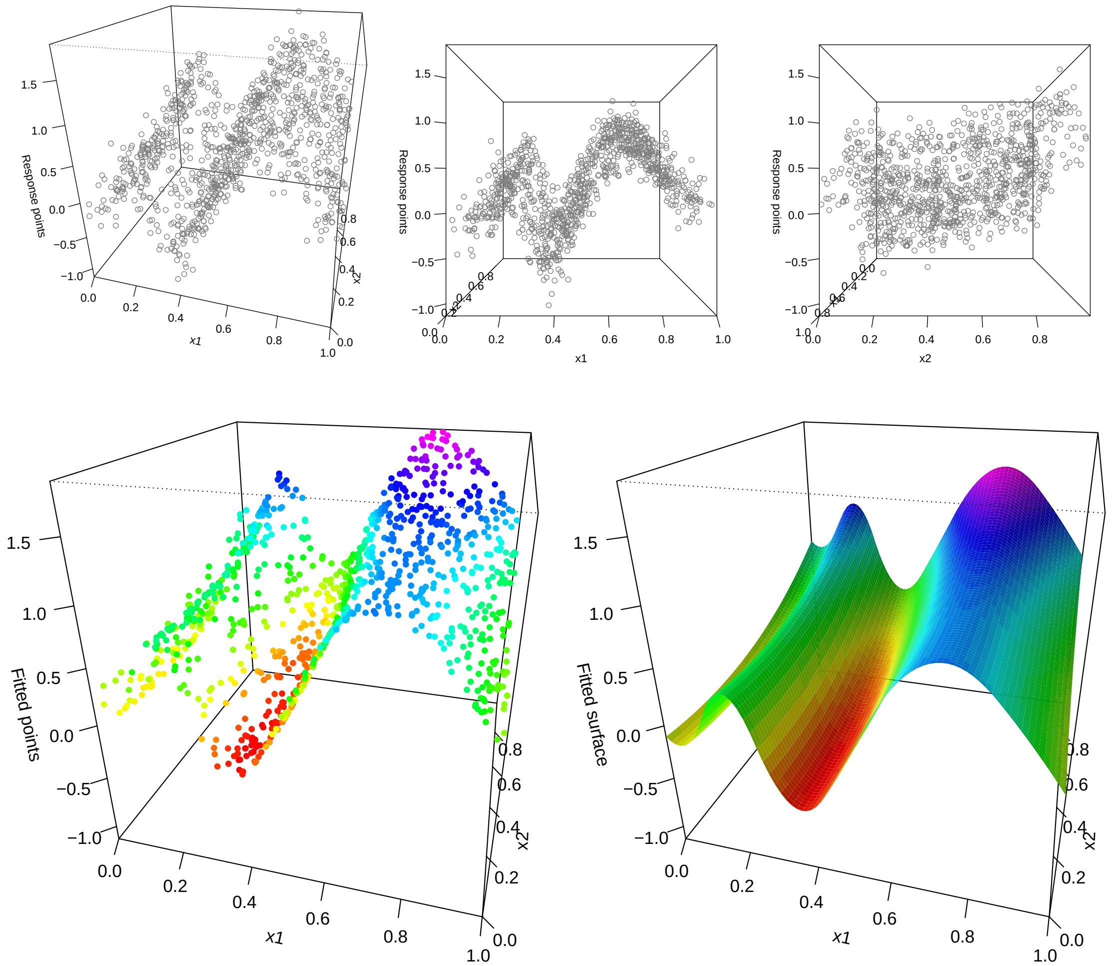

Figure 3: *An example of extrapolating the fitted additive trend filtering model, where* $n = 1000$ *and* $d = 2$. *We generated input points* $X^i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1]^2$, $i = 1, \ldots, 1000$, *and responses* $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^2 f_{0j}(X^i_j), \sigma^2)$, $i = 1, \ldots, 1000$, *where we* $f_{01}(x_1) = \sqrt{x_1}\sin(3\pi/(x_1 + 1/2))$ *and* $f_{02}(x_2) = x_2(x_2 - 1/3)$, *and* $\sigma = 0.36$. *The top row shows three perspectives of the data. The bottom left panel shows the fitted values from additive trend filtering* (8) *(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* (14).

Lemma 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 (12). 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*

$$\tilde{P} = M\big[\, \tilde{P}^{(X_1,k)} \quad \dots \quad \tilde{P}^{(X_d,k)} \,\big] \in \mathbb{R}^{n\times dk}, \tag{17}$$

the product of M and the columnwise concatenation of $\tilde{P}^{(X_j,k)}$, $j = 1, \dots, d$. 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

$$\tilde{K} = U^T M\big[\, K^{(X_1,k)} \quad \dots \quad K^{(X_d,k)} \,\big] \in \mathbb{R}^{(n-kd-1)\times(n-k-1)d}, \tag{18}$$

the product of $U^T M$ and the columnwise concatenation of $K^{(X_j,k)}$, $j = 1, \dots, d$. A sufficient condition for uniqueness of the additive trend filtering solution in (8) 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 $j = 1, \dots, d$, the projection of $\hat{\theta}_j$ onto the column space of $K^{(X_j,k)}$ is unique.*

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

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

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

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 $j = 1, \dots, d$. The lemma shows that the knot-producing parts $\hat{\theta}_j^{\text{knot}}$, $j = 1, \dots, d$ are uniquely determined when the columns of $\tilde{K}$ are in general position, and the polynomial parts $\hat{\theta}_j^{\text{knot}}$, $j = 1, \dots, d$ 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 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.

Remark 1 (**Relationship to concurvity**). *It is interesting to draw a connection to Buja et al. (1989). In the language used by these authors, when $\tilde{P}$ has linearly dependent columns, we say that the predictor variables display concurvity, i.e., linear dependence after nonlinear (here, polynomial) transformations are applied. Buja et al. (1989) establish that the components in the additive model (1), built with quadratic penalties, are unique provided there is no concurvity between variables. In comparison, Lemma 2 establishes uniqueness of the additive trend filtering components when there is no concurvity between variables, and additionally, the columns of $\tilde{K}$ are in general position. The latter two conditions together can be seen as requiring no generalized concurvity—if $\tilde{K}$ were to fail the general position assumption, then there would be a small subset of the variables that are linearly dependent after nonlinear (here, piecewise polynomial) transformations are applied.*

2.3 Dual problem

Let us abbreviate $D_j = D^{(X_j,k+1)}$, $j = 1,\dots,d$ for the penalty matrices in the additive trend filtering problem (8). Basic arguments in convex analysis, deferred to Appendix A.3, show that the dual of problem (8) can be expressed as:

$$\begin{aligned} \hat{u} = \operatorname*{argmin}_{u\in\mathbb{R}^n} \; \|Y - \bar{Y}\mathbb{1} - u\|_2^2 \;\; \text{subject to} \;\; u \in U = U_1 \cap \ldots \cap U_d, \\ \text{where} \quad U_j = \{S_j D_j^T v_j : \|v_j\|_\infty \le \lambda\}, \quad j = 1,\dots,d, \end{aligned} \tag{19}$$

and that the primal and dual solutions are related by:

$$\sum_{j=1}^d \hat{\theta}_j = Y - \bar{Y}\mathbb{1} - \hat{u}. \tag{20}$$

From the form of (19), 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 (20), 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 5, where we return to the dual perspective, and argue that the backfitting algorithm for the additive trend filtering problem (8) can be seen as a type of alternating projections algorithm for its dual problem (19).

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 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 estimators is well-understood (Tibshirani & Taylor 2011, 2012). The next result is an application of existing generalized lasso theory, proved in Appendix A.4.

Lemma 3 (**Degrees of freedom**). *Assume the conditions of Lemma 2, i.e., that the matrix $\tilde{P}$ in (17) has full column rank, and the matrix $\tilde{K}$ in (18) is in general position. Assume also that the response is Gaussian, $Y \sim N(\eta, \sigma^2 I)$, and treat the inputs $X^i = (X_1^i,\dots,X_d^i) \in \mathbb{R}^d$, $i = 1,\dots,n$ as fixed and arbitrary, as well as the tuning parameter value $\lambda \ge 0$. Then the additive trend filtering fit from (8) 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 2 (**Degrees of freedom without uniqueness**). *When the components $\hat{\theta}_j$, $j = 1,\dots,d$ of the additive trend filtering solution in (8) are not unique, it is still possible to derive an exact* *expression for degrees of freedom of the total fit (the generalized lasso results in Tibshirani & Taylor (2012) assume nothing about uniqueness); however, this expression is less transparent, so we assume uniqueness here for simplicity.*

Remark 3 (The effect of shrinkage). *Lemma 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 $j = 1, \ldots, d$, 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 (8) (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 general discussion of this phenomenon.*

Remark 4 (Easy unbiased degrees of freedom estimation). *It is worth emphasizing that an unbiased estimate from Lemma 3 for the degrees of freedom of the total fit in additive trend filtering is very easy to calculate: we scan the individual component fits and add up the number of knots that appear in each one. The same cannot be said for additive smoothing splines, or additive models built from univariate linear smoothers, in general. Although computing the fit itself is typically cheaper with additive linear smoothers than with additive trend filtering, computing the degrees of freedom is more challenging. For example, for the additive model in (1) built with quadratic penalties, we have*

$$\mathrm{df}\bigg(\sum_{j=1}^d \hat{\theta}_j\bigg) = \mathrm{tr}\Big(F^T F (F^T F + \lambda Q)^+\Big),$$

where $F \in \mathbb{R}^{n\times nd}$ has d copies of the centering matrix $M = I - \mathbb{1}\mathbb{1}^T/n \in \mathbb{R}^{n\times n}$ stacked across its columns, $Q \in \mathbb{R}^{nd\times nd}$ is a block diagonal matrix with blocks MQ_jM, $j = 1, \ldots, d$, and A^+ denotes the Moore-Penrose pseudoinverse of a matrix A. The above formula does not obviously decompose into a sum of quantities across components, and is nontrivial to compute post optimization of (1), specifically when a backfitting algorithm as in (2) has been used to compute the solution.

3 Two related additive spline estimators

Additive trend filtering is closely related to two other additive spline estimators, introduced here.

3.1 Additive locally adaptive splines

Consider the *additive locally adaptive regression spline* estimator, defined by

$$\begin{aligned}(\hat{f}_1, \ldots, \hat{f}_d) = \;\; \underset{f_1,\ldots,f_d}{\mathrm{argmin}} \quad & \frac{1}{2}\sum_{i=1}^n \bigg(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i)\bigg)^2 + \lambda \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d.\end{aligned} \tag{21}$$

Note the close analogy to the additive trend filtering problem, when written in the functional form of (11): the above minimization considers all functions $(f_1, \ldots f_d)$ (whose components have k weak derivatives), but the additive trend filtering problem just considers functions whose components are in the span of falling factorial functions, $f_j \in \mathcal{H}_j$, $j = 1, \ldots, d$.

Mammen & van de Geer (1997) study the above problem for $d = 1$; their results easily extend to (21), for a general d, and imply that the infinite-dimensional minimization problem in (21) has a solution $(\hat{f}_1, \ldots \hat{f}_d)$ such that each component function is a spline of degree k (justifying the choice of name). These results further imply that when $k = 0$ or $k = 1$, the knots of the spline $\hat{f}_j$ lie among the jth dimension of the input points $X_j^1, \ldots, X_j^n$, for $j = 1, \ldots, d$. But when $k \geq 2$, this is not necessarily true, and the spline component functions will generically have knots at locations other than the inputs; this makes computation of the solution in (21) very difficult.

3.2 Restricted additive locally adaptive splines

As a way of simplifying the computational issue mentioned above, we define the *restricted additive locally adaptive regression spline* estimator according to the problem

$$\begin{aligned} (\hat{f}_1, \ldots, \hat{f}_d) = \underset{f_j \in \mathcal{G}_j,\, j=1,\ldots,d}{\mathrm{argmin}} \quad & \frac{1}{2} \sum_{i=1}^n \Big(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \Big)^2 + \lambda \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}) \\ \text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d. \end{aligned} \tag{22}$$

Here, we have restricted the minimization domain, by requiring that $f_j \in \mathcal{G}_j$, where $\mathcal{G}_j$ is the space of spline functions of degree k with knots among the jth dimension of input points $X_j^1, \ldots, X_j^n$, for $j = 1, \ldots, d$. 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

$$T_j = \begin{cases} \{t_j^{k/2+2}, \ldots, t_j^{n-k/2}\} & \text{if } k \text{ is even}, \\ \{t_j^{(k+1)/2+1}, \ldots, t_j^{n-(k+1)/2}\} & \text{if } k \text{ is odd}, \end{cases} \tag{23}$$

i.e., defined by removing $k + 1$ inputs at the boundaries, for $j = 1, \ldots, d$. This makes (22) a finite-dimensional problem, just like (11). When $k = 0$ or $k = 1$, as remarked in Section 2.1 (and shown in Tibshirani (2014)), the falling factorial functions are simply splines, and this means that $\mathcal{H}_j = \mathcal{G}_j$ for $j = 1, \ldots, d$, hence (11) and (22) are equivalent problems. When $k \geq 2$, this is no longer true, and the solutions in (11) and (22) will differ; the former will be much easier to compute, since the latter does not admit a nice representation in terms of discrete derivatives, as in (8).

To summarize, when $k = 0$ or $k = 1$, all three total variation penalized additive estimators as defined in (21), (22), and (11) are equivalent. When $k \geq 2$, these estimators will generically differ, though our intuition tells us that their differences should not be too large: the first problem admits a solution that is a spline in each component; the second problem simply restricts this spline to have knots at the input points; the third problem swaps splines for falling factorial functions, which are highly similar in form. As splines—specifically, the truncated power basis—and the falling factorial basis have such similar structure, one would guess that the solutions to (21), (22), (11) should not be too different. The theory given next will confirm that this intuition holds true, in large samples.

4 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 component functions whose derivatives are of bounded variation. We first introduce some helpful notation, then we establish error bounds for additive model estimates built using roughness penalties, cast in a general form. Following this, we derive error bounds for additive trend filtering and additive locally adaptive splines. Lastly, we complement our derived error rates with matching minimax lower bounds.

4.1 Notation

For a distribution Q supported on $[0,1]^d$, we define the $L_2(Q)$ inner product $\langle \cdot, \cdot \rangle_{L_2(Q)}$, abbreviated as $\langle \cdot, \cdot \rangle_2$, for functions $m, r : [0,1]^d \to \mathbb{R}$, by

$$\langle m, r \rangle_2 = \int_{[0,1]^d} m(x) r(x) \, dQ(x),$$

Given independent draws $X^i = (X_1^i, \ldots, X_d^i)$, $i = 1, \ldots, n$ from Q, we also define the $L_2(Q_n)$ inner product $\langle \cdot, \cdot \rangle_{L_2(Q_n)}$, abbreviated as $\langle \cdot, \cdot \rangle_n$, by

$$\langle m, r \rangle_n = \frac{1}{n} \sum_{i=1}^n m(X^i) r(X^i),$$

Definitions for the norms $\| \cdot \|_{L_2(Q)}$ and $\| \cdot \|_{L_2(Q_n)}$, abbreviated as $\| \cdot \|_2$ and $\| \cdot \|_n$, respectively, arise naturally from these inner products, as in

$$\|m\|_2^2 = \langle m, m \rangle_2 = \int_{[0,1]^d} m(x)^2 \, dQ(x), \quad \text{and} \quad \|m\|_n^2 = \langle m, m \rangle_n = \frac{1}{n} \sum_{i=1}^n m(X^i)^2.$$

Of particular interest, of course, will be additive functions, of the form $m(x) = \sum_{j=1}^d m_j(x_j)$. In a slight abuse of notation, we will denote such functions by

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

(This is justified by viewing each m_j as a function over $[0,1]^d$, even though it only actually depends on the jth dimension, for each $j = 1, \ldots, d$.)

For (univariate) component functions, we can define analogous inner products and norms. Let Q_j denote the marginal distribution of Q along the jth dimension, for $j = 1, \ldots, d$. We define inner products $\langle \cdot, \cdot \rangle_{L_2(Q_j)}$ and $\langle \cdot, \cdot \rangle_{L_2(Q_{n,j})}$, which we abbreviate by $\langle \cdot, \cdot \rangle_{2,j}$ and $\langle \cdot, \cdot \rangle_{n,j}$, respectively, for $j = 1, \ldots, d$, over functions $g, h : [0,1] \to \mathbb{R}$ by

$$\langle g, h \rangle_{2,j} = \int_0^1 g(t) h(t) \, dQ_j(t), \quad \text{and} \quad \langle g, h \rangle_{n,j} = \frac{1}{n} \sum_{i=1}^n g(X_j^i) h(X_j^i), \quad j = 1, \ldots, d.$$

Similarly, the norms $\| \cdot \|_{L_2(Q_j)}$ and $\| \cdot \|_{L_2(Q_{n,j})}$, abbreviated by $\| \cdot \|_{2,j}$ and $\| \cdot \|_{n,j}$, for $j = 1, \ldots, d$, are defined by

$$\|g\|_{2,j} = \int_0^1 g(t)^2 \, dQ_j(t), \quad \text{and} \quad \|g\|_{n,j} = \frac{1}{n} \sum_{i=1}^n g(X_j^i)^2, \quad j = 1, \ldots, d.$$

We will refer to $\langle \cdot, \cdot \rangle_2$, $\| \cdot \|_2$, $\langle \cdot, \cdot \rangle_{2,j}$, $\| \cdot \|_{2,j}$, $j = 1, \ldots, d$ as the L_2 inner products and norms, and $\langle \cdot, \cdot \rangle_n$, $\| \cdot \|_n$, $\langle \cdot, \cdot \rangle_{n,j}$, $\| \cdot \|_{n,j}$, $j = 1, \ldots, d$ as the empirical inner products and norms.

A few more general definitions are in order. Given a domain D (e.g., $D = [0,1]^d$, or $D = [0,1]$), we denote the L^∞ norm (i.e., sup norm) of a function $f : D \to \mathbb{R}$ by $\|f\|_\infty = \operatorname{ess\,sup}_{z \in D} |f(z)|$. For a functional ν acting on functions $f : D \to \mathbb{R}$, we write $B_\nu(\delta)$ for the ν-ball of radius $\delta > 0$, i.e., $B_\nu(\delta) = \{f : D \to \mathbb{R} : \nu(f) \le \delta\}$. 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. In what follows, we will use these definitions fluidly, without 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.

4.2 Error bounds for roughness-penalized estimators

We study a generic roughness-penalized additive model, namely the estimator defined by

$$
\begin{aligned}
(\hat{f}_1, \ldots, \hat{f}_d) = \underset{f_j \in \mathcal{S}_j,\, j=1,\ldots,d}{\mathrm{argmin}} \quad & \frac{1}{2} \sum_{i=1}^n \bigg(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \bigg)^2 + \lambda \sum_{j=1}^d J(f_j) \\
\text{subject to} \quad & \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \ldots, d, \quad \bigg\| \Pi_{k,n}^\perp \bigg(\sum_{j=1}^d f_j \bigg) \bigg\|_\infty \le b_0.
\end{aligned} \tag{24}
$$

where $\mathcal{S}_j$, $j = 1, \ldots, d$ are univariate function spaces, and J is a regularizer that acts on univariate functions. Note that the problem (24) uses a sup norm bound as a constraint, $\Pi_{k,n}^\perp(\sum_{j=1}^d f_j) \le b_0$, where $b_0 \ge 1$ is a constant and $\Pi_{k,n}^\perp$ is the projection operator, with respect to the $\|\cdot\|_n$ norm, onto the orthocomplement of additive kth degree polynomial functions (i.e., functions $\sum_{j=1}^d p_j$ where p_j is a kth degree polynomial in the variable x_j, $j = 1, \ldots, d$). This is used as a technicality, to derive the sharpest error rate possible (in terms of the dependence on the input dimension d); it can be removed at the expense of a somewhat looser error rate (again in terms of the dependence on d); see Remark 9 following Theorem 1.

Before stating the theorem, we list and discuss our assumptions. First, we give our assumptions on the data generating distribution.

Assumption A1 (Random inputs, independent across dimensions). *The input points $X^i = (X_1^i, \ldots, X_d^i)$, $i = 1, \ldots, n$ are i.i.d. from a product distribution $Q = Q_1 \times \cdots \times Q_d$, supported on $[0,1]^d$, where $Q_1, \ldots, Q_d$ have densities uniformly bounded below and above by constants $b_1, b_2 > 0$, respectively, on $[0,1]$.*

Assumption A2 (Generic regression model, sub-Gaussian errors). *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. For a nonnegative integer k the function f_0 is sup norm bounded after removing all additive polynomials of degree k, more precisely $\|\Pi_{k,n}^\perp(f_0)\|_\infty \le b_0$ for a constant $b_0 \ge 1$. The errors ϵ^i, $i = 1, \ldots, n$ are uniformly sub-Gaussian and have mean zero, i.e.,

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

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

Assumption A2 is fairly standard. We do not place specific smoothness or additivity conditions on the underlying regression function f_0, since our error bound in Theorem 1 will involve the error of the closest additive function $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$ to f_0, with components lying in the designated function spaces, $\tilde{f}_j \in \mathcal{S}_j$, $j = 1, \ldots, d$. We assume $\Pi_{k,n}^\perp(f_0)$ is bounded in the sup norm, where recall $\Pi_{k,n}^\perp$ is the projection operator, with respect to the $\|\cdot\|_n$ norm, onto the orthocomplement of additive polynomials of degree k. Sup norm boundedness is a pretty common assumption in the analysis of additive models (with growing dimension), though some authors, such as Raskutti et al. (2012), avoid it. We feel that it is a mostly innocuous assumption in our case, as we allow f_0 itself to grow arbitrarily large, and we only require its nonparametric part $\Pi_{k,n}^\perp(f_0)$ to be bounded. Moreover, as noted before, this assumption can be removed at the expense of a somewhat weaker final error rate, see Remark 9 after the theorem.

Assumption A1 is more 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 an important decomposability property with respect to the L_2 norm defined over Q. In particular, for $m = \sum_{j=1}^d m_j$, such that its components that have L_2 mean zero, written as $\bar{m}_j = \int_0^1 m_j(x_j)\, dQ_j(x_j) = 0$, $j = 1, \dots, d$, we have

$$\Big\| \sum_{j=1}^d m_j \Big\|_2^2 = \sum_{j=1}^d \|m_j\|_{2,j}^2. \tag{25}$$

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 (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 (as the inner products between components do not vanish even if the empirical means are all zero). The property in (25) is important for deriving the sharp estimation error rate in Theorem 1, where the dependence on d is linear (see the result in (28)); but an incoherence bound as in (31) would also be sufficient; see Remark 8 after the theorem.

Our next assumption is on the dimension d of the input space, which we allow to grow with n, but not too quickly.

Assumption B1 (Dimension restriction). *The dimension d scales as $d = o(\sqrt{n})$.*

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

Assumption C1 (Seminorm regularizer, null space of polynomials). *The regularizer J is a seminorm, defined over (possibly a subset of) the set of k times weakly differentiable functions. Its null space contains all kth order polynomials, $J(g) = 0$ for all $g(t) = t^\ell$, $\ell = 0, \dots, k$.*

Assumption C2 (Relative boundedness of derivatives). *There is a constant $L > 0$ such that $\operatorname{ess\,sup}_{t\in[0,1]} g^{(k)}(t) - \operatorname{ess\,inf}_{t\in[0,1]} g^{(k)}(t) \le L$ for $g \in B_J(1)$ (with $g^{(k)}$ the kth weak derivative of g).*

Assumption C3 (Entropy bound). *There are constants $0 < w < 2$ and $K > 0$ such that*

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

These assumptions on the regularizer J are not strong, and are satisfied by various commonly used regularizers, e.g., $J(g) = \int_0^1 (g^{(k+1)}(t))^2\, dt$ or $J(g) = \mathrm{TV}(g^{(k)})$, the latter of which we will study in the next subsection.

We are now ready to state the main result, which is proved in Appendix A.5, A.6.

Theorem 1. *Assume A1, A2 on the data generating distribution, B1 on the dimension of the input space, and C1–C3 on the regularizer J. Then there exist constants $c_0, c_1, n_0 > 0$, that depend only on $b_0, b_1, b_2, \sigma, k, L, K, w$, such that for all $c \ge c_0$ and $n \ge n_0$, and all values of the tuning parameter $\lambda \ge 2c^2 n^{w/(2+w)}$, any solution in (24) satisfies the empirical norm bound*

$$\Big\| \sum_{j=1}^d \hat{f}_j - f_0 \Big\|_n^2 \le \Big\| \sum_{j=1}^d \tilde{f}_j - f_0 \Big\|_n^2 + \frac{5\lambda}{2n} \max\Big\{ d, \sum_{j=1}^d J(\tilde{f}_j) \Big\} \tag{26}$$

simultaneously over all functions $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$ that are feasible for the problem (24), with probability at least $1 - \exp(-c) - c_1 d/n$. In addition, we have the L_2 norm bound

$$\Big\| \sum_{j=1}^d \hat{f}_j - f_0 \Big\|_2^2 \le 3\Big\| \sum_{j=1}^d \tilde{f}_j - f_0 \Big\|_2^2 + 86\Big\| \sum_{j=1}^d \tilde{f}_j - f_0 \Big\|_n^2 + \frac{127\lambda}{n} \max\Big\{ d, \sum_{j=1}^d J(\tilde{f}_j) \Big\} \tag{27}$$

simultaneously over all functions $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$ that are feasible for the problem (24), with probability at least $1 - \exp(-c) - c_1 d/n$.

Remark 5 (Tightness of constants). *In the proof of Theorem 1 we are careful with the constants appearing in (26), or rather, we are careful enough to ensure that there is a factor of 1 multiplying the approximation error (first time on the right) in (26), making this result an oracle inequality for the estimator in (24). On the other hand, we were not careful with the constants in (27), and only sought a clean bound; these constants could likely be improved.*

Remark 6 (Error bound for additive, J-smooth f_0). *In the case $f_0 = \sum_{j=1}^d f_{0j}$, where $f_{0j} \in \mathcal{S}_j$ are such that $J(f_{0j}) \le C_0$, for $j = 1, \dots, d$, and a constant $C_0 > 0$, the approximation error terms in (26), (27) are zero when we set $\tilde{f} = f_0$. Hence for $c \ge c_0$, $n \ge n_0$, and $\lambda = 2c^2 n^{w/(2+w)}$,*

$$\max\left\{ \left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_n^2, \left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_2^2 \right\} \le c^2 d n^{-2/(2+w)}, \tag{28}$$

with probability at least $1 - \exp(-c) - c_1 d/n$. In Theorem 2, we will show that this rate $dn^{-2/(2+w)}$ is indeed minimax optimal over such a class of underlying functions f_0. When C_0 is growing with n, however, the minimax error rate is $C_0^{2w/(2+w)} dn^{-2/(2+w)}$, and as we can see from (26), (27), the additive estimate in (24) has an error rate of $C_0 dn^{-2/(2+w)}$ in either the squared empirical and L_2 norms, lacking the optimal dependence on C_0. This can be fixed by placing more assumptions on f_0 (namely, sup norm boundedness of its component functions), but we do not pursue this.

Further, we note that a truncation argument shows that the same rate $dn^{-2/(2+w)}$ also holds in expectation for the empirical norm error, i.e.,

$$\mathbb{E} \left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_n^2 \le c_2 d n^{-2/(2+w)}, \tag{29}$$

for all $n \ge n_1$, where $c_2, n_1 > 0$ are constants.

Remark 7 (Distance to best additive, J-smooth approximation of f_0). *The arguments used to prove the oracle inequality (26) also imply a result on the empirical norm distance between $\hat{f}$ and the best additive approximation of f_0. Precisely, let $\tilde{f}^{\text{best}} = \sum_{j=1}^d \tilde{f}_j^{\text{best}}$ minimize $\mathbb{E}\|f_0 - \tilde{f}\|_n^2$ over all additive functions $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$. Assuming that $J(\tilde{f}_j) \le C_0$, for all $j = 1, \dots, d$, and a constant $C_0 > 0$, a consequence of the proof of Theorem 1 (following from (78), and a truncation argument similar to that used for (29)) is the bound*

$$\mathbb{E} \left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d \tilde{f}_j^{\text{best}} \right\|_n^2 \le c_2 d n^{-2/(2+w)}, \tag{30}$$

for all $n \ge n_1$, where $c_2, n_1 > 0$ are constants. Notably, the right-hand side in (30) does not depend on the approximation error (of $\tilde{f}^{\text{best}}$ to f_0). This is analogous to classical results from Stone (1985) on fixed-dimensional additive models.

Remark 8 (Importance of L_2 decomposability). *The proof of Theorem 1 derives a basic inequality using empirical norms, and then eventually converts error terms to their L_2 norm analogs, so that the decomposability property in (25) can be applied. This property is a key to delivering the sharp (linear) dependence on d in the final estimation error rate (see also the result in (28)). But it is worth noting that all that is needed in the proof is in fact a lower bound of the form*

$$\left\| \sum_{j=1}^d m_j \right\|_2^2 \ge \phi_0 \sum_{j=1}^d \|m_j\|_{2,j}^2, \tag{31}$$

for a constant $\phi_0 > 0$, rather than exact an equality, as in (25). The condition (31) is an incoherence condition that can hold for nonproduct distributions Q, over an appropriate class of functions (additive functions with sufficiently smooth components), provided that the correlations between components of Q are under control. It is likely that Assumption A1 could be relaxed to allow for such a distribution Q with weak correlations, and the proof of the theorem could be modified to accommodate the nonproduct form of Q when drawing connections between the empirical and L_2 norms. We do not pursue such a relaxation, but we refer the reader to Meier et al. (2009), van de Geer (2014) for similar strategies.

Remark 9 **(The role of sup norm boundedness).** *The sup norm bound $\|\Pi^\perp_{k,n}(f_0)\|_\infty \le b_0$ in Assumption A2, as well as the sup norm constraint $\|\Pi^\perp_{k,n}(\sum_{j=1}^d f_j)\|_\infty \le b_0$ in the problem (24), are technicalities that are important for deriving the sharp estimation error rates in (26), (27), leading to the rate $dn^{-2/(2+w)}$ in (28) for additive f_0. Without these sup norm bounds, small modifications to the proof of the theorem show that the results (26), (27) still hold but with λ replaced by $\lambda d^{1-w/2}$, which would lead to an error rate of $d^{2-w/2}n^{-2/(2+w)}$ for additive f_0 in (28), which has a suboptimal dependence on the dimension d.*

Furthermore, still without sup norm bounds, we can alternatively analyze a constrained version of the estimator in (24), where the additive penalty $\sum_{j=1}^d J(f_j)$ is replaced by constraints on $J(f_j)$, $j = 1,\dots,d$, which is similar to the strategy taken in Raskutti et al. (2012). These authors introduce a clever technique to upper bound the L_2 norm by the empirical norm uniformly over functions from a (potentially) unbounded function class, under a fourth moment condition (see also Theorem 14.2 in Wainwright (2017)). Similar arguments can be used to establish essentially the same result as in Theorem 1 for the constrained estimator; however, we prefer to study the penalized estimator defined in (24), as this is in line with what is used in practice (the constrained version is truly a different estimator: constraints on each of $J(f_j)$, $j = 1,\dots,d$ are not dual to a single penalty $\sum_{j=1}^d J(f_j)$).

4.3 Error bounds for additive locally adaptive splines and additive trend filtering

We study the problem in (24) when the regularizer is $J(g) = \mathrm{TV}(g^{(k)})$, and derive the implications of Theorem 1 for additive locally adaptive splines and trend filtering estimates (which correspond to different choices for the function classes $\mathcal{S}_j$, $j = 1,\dots,d$ in (24)). The proof is given in Appendix A.7, A.8.

Corollary 1. *Assume A1, A2 and B1. Assume that the underlying regression function is additive, $f_0 = \sum_{j=1}^d f_{0j}$, where the components are k times weakly differentiable, and satisfy $\mathrm{TV}(f_{0j}^{(k)}) \le C_0$, $j = 1,\dots,d$, for a constant $C_0 > 0$. Then, for the seminorm $J(g) = \mathrm{TV}(g^{(k)})$, Assumptions C1–C3 hold with $L = 1$ and $w = 1/(k+1)$. Further, the following is true of the estimator in (24).*

(a) Let $\mathcal{S}_j$ be the space of all k times weakly differentiable functions, for $j = 1,\dots,d$. Then there are constants $c_0, c_1, n_0 > 0$, that depend only on $b_0, b_1, b_2, \sigma, k, C_0$, such that for all $c \ge c_0$ and $n \ge n_0$, any solution in the (sup norm bounded) additive locally adaptive spline problem (24) with regularization parameter value $\lambda = 2c^2 n^{1/(2k+3)}$ satisfies

$$\max\left\{\left\|\sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j}\right\|_n^2, \left\|\sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j}\right\|_2^2\right\} \le c^2 dn^{-(2k+2)/(2k+3)}, \tag{32}$$

with probability at least $1 - \exp(-c) - c_1 d/n$.

(b) Let $\mathcal{S}_j = \mathcal{G}_j$, the space of kth degree splines with knots in the set T_j in (23), for $j = 1,\dots,d$. There are constants $c_0, c_1, n_0 > 0$, that depend only on $b_0, b_1, b_2, \sigma, k, C_0$, such that for all $c \ge c_0$ and $n \ge n_0$, any solution in the (sup norm bounded) restricted additive locally adaptive spline

problem (24) *with* $\lambda = 2c^2 n^{1/(2k+3)}$, *satisfies the same error bounds as in* (32) *with probability at least* $1 - \exp(-c) - c_1 d/n$.

(c) *Let* $\mathcal{S}_j = \mathcal{H}_j$, *the space of* k*th degree falling factorial functions defined over the* j*th dimension of inputs* X_j, *for* $j = 1, \ldots, d$. *Then there exist constants* $c_0, c_1, n_0 > 0$, *that depend only on* $b_0, b_1, b_2, \sigma, k, C_0$, *such that for all* $c \geq c_0$ *and* $n \geq n_0$, *and all values* $\lambda \geq 2c^2 n^{1/(2k+3)}$, *any solution in the (sup norm bounded) additive trend filtering problem* (24) *with* $\lambda = 2c^2 n^{1/(2k+3)}$, *satisfies the same error bounds as in* (32) *with probability at least* $1 - \exp(-c) - c_1 d/n$.

Remark 10 (**Spline and falling factorial approximations**). *Part (a) of Corollary 1 is a direct application of Theorem 1, after we verify Assumptions C1–C3 for the regularizer* $J(g) = \mathrm{TV}(g^{(k)})$. *Parts (b) and (c) require control over the approximation error (the first term) in* (28), *because the underlying regression function* $f_0 = \sum_{j=1}^d f_{0j}$ *need not have components lying in the chosen spaces* $\mathcal{S}_j$, $j = 1, \ldots, d$, *for parts (b) and (c). (When* $k = 0$ *or* $k = 1$, *for the same reasons as discussed in Section 3.2, the approximation error is zero in both parts (b) and (c), so these approximation bounds are really only needed for* $k \geq 2$.*) For both parts (b) and (c), we control the approximation error by bounding the univariate approximation error and then using 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 use results from de Boor (1978)); for part (c), we develop a new falling factorial approximant that may be of independent interest.*

4.4 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. The input points will be assumed to follow a product distribution $Q = Q_1 \times \cdots \times Q_d$ as in Assumption A1. As for the responses, we will use the next assumption in place of A2.

Assumption A3 (Additive model, Gaussian errors). *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* $\int_{[0,1]^d} f_0(x)\, dQ(x) = 0$ *for identifiability, and 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 with the following assumption, on the log packing and log covering (entropy) numbers.

Assumption C4 (Matching packing and covering number bounds). *There exist constants* $0 < w < 2$ *and* $K_1, K_2 > 0$ *such that*

$$\log M\big(\delta, \|\cdot\|_{2,1}, B_J(1) \cap B_\infty(1)\big) \geq K_1 \delta^{-w},$$
$$\log N\big(\delta, \|\cdot\|_{2,1}, B_J(1) \cap B_\infty(1)\big) \leq K_2 \delta^{-w}.$$

Notice that the covering and packing numbers in Assumption C4 are defined using $\|\cdot\|_{2,1}$, the $L^2(Q_1)$ norm, but we could have equally well used $\|\cdot\|_{2,j}$, the $L^2(Q_j)$ norm, for any $j = 2, \ldots, d$ (or even the L^2 norm with respect to the univariate uniform measure), as the densities $Q_1, \ldots, Q_d$ are all within constant factors of each other under Assumption A1.

Now we state our minimax lower bound, for the function class

$$B_J^d(C_0) = \bigg\{ \sum_{j=1}^d f_j : J(f_j) \leq C_0,\ j = 1, \ldots, d \bigg\}. \tag{33}$$

Its proof is given in Appendix A.9, A.10.

Theorem 2. *Assume A1, A3 and C1, C2, C4. Then there exist constants $c_0, n_0 > 0$, that depend only on $b_1, b_2, \sigma, k, K_1, K_2, w$, such that for all $n \geq n_0$ and $C_0 \geq 1$,*

$$\inf_{\hat{f}} \sup_{f_0 \in B_J^d(C_0)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \geq c_0 C_0^{2w/(2+w)} dn^{-2/(2+w)}.$$

When we choose $J(g) = \mathrm{TV}(g^{(k)})$ as our regularizer, the additive function class in (33) becomes

$$\mathcal{F}_k^d(C_0) = \bigg\{ \sum_{j=1}^d f_j : \mathrm{TV}(f_j^{(k)}) \leq C_0,\ j = 1, \ldots, d \bigg\}, \tag{34}$$

and Theorem 2 implies the following.

Corollary 2. *Assume A1, A3, where 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 $n \geq n_0$ and $C_0 \geq 1$,*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_k^d(C_0)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \geq c_0 C_0^{2/(2k+3)} dn^{-(2k+2)/(2k+3)}. \tag{35}$$

Remark 11 (Optimality of additive trend filtering and additive locally adaptive regression splines). *Comparing the upper bound in (32) to the lower bound in (35), we see that the (sup norm bounded) additive locally adaptive spline, additive restricted locally adaptive spline, and additive trend filtering estimators are each minimax rate optimal for estimation over the class $\mathcal{F}_k^d(C_0)$, provided C_0 is a constant.*

Remark 12 (Suboptimality of additive smoothing splines and additive linear smoothers in general). *Seminal results from Donoho & Johnstone (1998) on minimax linear rates over Besov spaces imply that, under Assumption A3, but with inputs X^i, $i = 1, \ldots, n$ being now nonrandom and occurring over the regular lattice $\{1/N, 2/N, \ldots, 1\}^d \subseteq [0,1]^d$, with $N = n^{1/d}$, we have*

$$\inf_{\hat{f} \text{ additive linear}} \sup_{f_0 \in \mathcal{F}_k^d(C_0)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \geq c_0 C_0^{2/(2k+2)} dn^{-(2k+1)/(2k+2)}, \tag{36}$$

for a constant $c_0 > 0$ and sufficiently large n. On the left-hand side above, the infimum is taken over all additive linear smoothers, i.e., estimators $\hat{f} = \sum_{j=1}^d \hat{f}_j$ such that each $\hat{f}_j$ is a linear smoother, for $j = 1, \ldots, d$; also, the norm under consideration is $\|\cdot\|_2 = \|\cdot\|_{L_2(U)}$, the L_2 norm with respect to the uniform distribution U on $[0,1]^d$. When C_0 is a constant, we can compare (36) and (35) (and recall the tightness of the latter lower bound from (32)) to see that all additive linear smoothers—such as additive smoothing splines, additive kernel smoothing estimators, additive RKHS estimators, and so on—are suboptimal over the class $\mathcal{F}_k^d(C_0)$.

The argument to verify (36) is provided in Appendix A.12. We use the assumption of a regular lattice for the inputs for convenience, so that the additive problem considered here can be translated into a collection of appropriate univariate problems, and univariate results of Donoho & Johnstone (1998) can be applied. We do not see this as an unreasonable assumption; observing the input points on a regular lattice is intuitively an "easier" setup than observing random inputs, drawn from some distribution over $[0,1]^d$; thus a lower bound rate derived in the lattice setting suggests that the same should hold for random inputs. In fact, we conjecture that the techniques from Donoho et al. (1990), Donoho & Johnstone (1998) could be used to establish for random inputs the same lower bound rate as in (36), but do not pursue this.

5 Backfitting and the dual

We now examine computational approaches for the additive trend filtering problem in (8). This is a convex optimization problem, and many standard approaches can be applied. For its simplicity, and its history in additive modeling, we focus on the backfitting algorithm in particular.

Algorithm 1 Backfitting for additive trend filtering

Given responses $Y^i \in \mathbb{R}$ and input points $X^i = (X_1^i, \dots, X_d^i) \in \mathbb{R}^d$, $i = 1, \dots, n$.

1. Set $t = 0$ and initialize $\theta_j^{(t)} = 0$, $j = 1, \dots, d$.
2. For $t = 1, 2, 3, \dots$ (until convergence):

   a. For $j = 1, \dots, d$:

      (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) $\theta_j^{(t)} = \theta_j^{(t)} - \frac{1}{n}\mathbb{1}\mathbb{1}^T\theta_j^{(t)}$

3. Return $\hat{\theta}_j$, $j = 1, \dots, d$ (parameters $\theta_j^{(t)}$, $j = 1, \dots, d$ at convergence).

5.1 Backfitting

The backfitting approach for problem (8) 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 $r = (r^1, \dots, r^n) \in \mathbb{R}^n$ over input points $X_j = (X_j^1, \dots, X_j^n) \in \mathbb{R}^n$. 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, as the output of $\mathrm{TF}_\lambda(r, X_j)$ 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; in practice, to implement Algorithm 1 we can use fast interior 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 (8) over the coordinate blocks θ_j, $j = 1, \dots, d$. 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 (8), all limit points of the sequence of iterates produced by BCD are optimal solutions. More recent work from the optimization community offers refined convergence analyses for coordinate descent (or its variants) in particular settings. We do not pursue the implications of this work for BCD on problem (8); our interest here is primarily in developing a connection between BCD for problem (8) and alternating projections in its dual problem (19), which is the topic of the next subsection.

5.2 Dual alternating projections

Using the relationship between the additive trend filtering problem (8) and its dual (19), which are related by the transformation (20), we see that for any dimension $j = 1, \dots, d$, the univariate trend filtering fit to a response $r = (r^1, \dots, r^n)$ over input points $X_j = (X_j^1, \dots, X_j^n)$ may be written as

$$\mathrm{TF}_\lambda(r, X_j) = (\mathrm{Id} - \Pi_{U_j})(r), \tag{37}$$

where $U_j = \{S_j D_j^T v_j : \|u\|_\infty \le \lambda\}$ (and we abbreviate $D_j = D^{(X_j, k+1)}$). The backfitting or BCD approach in Algorithm 1 can be viewed (ignoring the optional centering step) as stepping through the updates, for $t = 1, 2, 3, \dots$,

$$\theta_j^{(t)} = (\mathrm{Id} - \Pi_{U_j})\Big(Y - \bar{Y}\mathbb{1} - \sum_{\ell<j} \theta_\ell^{(t)} - \sum_{\ell>j} \theta_\ell^{(t-1)}\Big), \quad j = 1, \dots, d, \tag{38}$$

or, reparametrized in terms of the primal-dual relationship $u = Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j$ in (20),

$$\begin{aligned} u_0^{(t)} &= Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j^{(t-1)}, \\ u_j^{(t)} &= \Pi_{U_j}\big(u_{j-1}^{(t)} + \theta_j^{(t-1)}\big), \quad j = 1,\ldots,d, \\ \theta_j^{(t)} &= \theta_j^{(t-1)} + u_{j-1}^{(t)} - u_j^{(t)}, \quad j = 1,\ldots,d. \end{aligned} \tag{39}$$

Thus the backfitting algorithm for (8), as expressed above in (39), is seen to be a particular type of *alternating projections* method applied to the dual problem (19), cycling through projections onto U_j, $j = 1,\ldots,d$. Interestingly, as opposed to the classical alternating projections approach, which would repeatedly project the current iterate $u_{j-1}^{(t)}$ onto U_j, $j = 1,\ldots,d$, the steps in (39) repeatedly project an "offset" version $u_{j-1}^{(t)} + \theta_j^{(t-1)}$ of the current iterate, for $j = 1,\ldots,d$ (this corresponds to running univariate trend filtering on the current residual, in the iterations (38)).

There is a considerable literature on alternating projections in optimization, see, e.g., Bauschke & Borwein (1996) for a review. Many alternating projections algorithms can be derived from the perspective of an operator splitting technique, e.g., the alternative direction method of multipliers (ADMM). In fact, the steps in (39) appear very similar to those from an ADMM algorithm applied to the dual (19), if we think of the "offset" variables θ_j, $j = 1\ldots,d$ in the iterations (39) as *dual* variables in the dual problem (19) (i.e., if we think of the primal variables θ_j, $j = 1,\ldots,d$ as dual variables in the dual problem (19)). This connection inspires a new parallel version of backfitting, presented in the next subsection.

5.3 Parallelized backfitting

We have seen that backfitting is a special type of alternating projections algorithm, applied to the dual problem (19). For classic intersection of sets problems (i.e., feasibility problems where we seek a point in the intersection of given closed, convex sets), with more than two sets, the optimization literature offers *parallel projections* methods (in contrast to alternating projections methods) that are provably convergent. One such method is derived using ADMM (e.g., see Section 5.1 of Boyd et al. (2011)). A similar construction may be used for the dual problem (19). We first rewrite this problem as

$$\begin{aligned} \min_{u_0,u_1,\ldots,u_d \in \mathbb{R}^n} \quad & \frac{1}{2}\|Y - \bar{Y}\mathbb{1} - u_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) \\ \text{subject to} \quad & u_0 = u_1,\ u_0 = u_2,\ \ldots\ u_0 = u_d, \end{aligned} \tag{40}$$

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

$$L_\rho(u_0, u_1, \ldots, u_d, \gamma_1, \ldots, \gamma_d) = \frac{1}{2}\|Y - \bar{Y}\mathbb{1} - u_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) + \frac{\rho}{2}\sum_{j=1}^d \|u_0 - u_j + \gamma_j\|_2^2, \tag{41}$$

where γ_j is a dual variable corresponding to the equality constraint $u_0 = u_j$ in (40), for $j = 1,\ldots d$. The ADMM steps for problem (40) are now given by repeating, for $t = 1, 2, 3, \ldots$,

$$\begin{aligned} u_0^{(t)} &= \frac{1}{d+1}\bigg(Y - \bar{Y}\mathbb{1} + \sum_{j=1}^d (u_j^{(t-1)} - \gamma_j^{(t-1)})\bigg) \\ u_j^{(t)} &= \Pi_{U_j}\big(u_0^{(t)} + \gamma_j^{(t-1)}\big), \quad j = 1,\ldots,d \\ \gamma_j^{(t)} &= \gamma_j^{(t-1)} + u_0^{(t)} - u_j^{(t)}, \quad j = 1,\ldots,d. \end{aligned} \tag{42}$$

Algorithm 2 Parallel backfitting for additive trend filtering

Given responses $Y^i \in \mathbb{R}$ and input points $X^i = (X^i_1, \dots, X^i_d) \in \mathbb{R}^d$, $i = 1, \dots, n$.

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. $u_0^{(t)} = \dfrac{1}{d+1}\left(Y - \bar{Y}\mathbb{1} - \displaystyle\sum_{j=1}^d \theta_j^{(t-1)}\right) + \dfrac{d}{d+1}\left(u_0^{(t-1)} + \dfrac{1}{d}\displaystyle\sum_{j=1}^d \big(\theta_j^{(t-2)} - \theta_j^{(t-1)}\big)\right)$

   b. For $j = 1, \dots, d$ (in parallel):

      (i) $\theta_j^{(t)} = \mathrm{TF}_\lambda\big(u_0^{(t)} + \theta_j^{(t-1)},\ X_j\big)$

      (ii) (Optional) $\theta_j^{(t)} = \theta_j^{(t)} - \frac{1}{n}\mathbb{1}^T\theta_j^{(t)}$

3. Return $\hat{\theta}_j$, $j = 1, \dots, d$ (parameters $\theta_j^{(t)}$, $j = 1, \dots, d$ at convergence).

(Here we have chosen $\rho = 1$ for the augmented Lagrangian parameter; in any case, due to the form of L_ρ in (41), only the u_0 update depends on ρ.) Now compare (42) to (39)—the key difference is that in (42), 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 (39).

Of course, the dual problem in (40) is only interesting insofar as it is connected to the additive trend filtering problem (8). Fortuitously, the parallel projections algorithm (42) maintains a very useful connection to the primal problem: the final parameters $\hat{\gamma}_j$, $j = 1, \dots, d$ (i.e., the iterates $\gamma_j^{(t)}$, $j = 1, \dots, d$ at convergence) precisely solve the additive trend filtering problem (8). This is simply because the dual of the dual problem (40) is indeed the additive trend filtering problem (8) (thus the parameters γ_j, $j = 1, \dots, d$, which are dual to the constraints in (40), are nothing more than the primal parameters θ_j, $j = 1, \dots d$ in (8)). For concreteness, we state this result as a theorem, and transcribe the iterations in (42) into an equivalent primal form, in Algorithm 2. For details, see Appendix A.13.

Theorem 3. *Initializing $u_j^{(0)}$, $\gamma_j^{(0)}$, $j = 1, \dots, d$ arbitrarily, the iterations (42) produce parameters $\hat{\gamma}_j$, $j = 1, \dots, d$ (i.e., $\gamma_j^{(t)}$, $j = 1, \dots, d$ at convergence) that solve the additive trend filtering problem (8), i.e., Algorithm 2 yields outputs $\hat{\theta}_j$, $j = 1, \dots, d$ that solve additive trend filtering (8).*

Written in primal form, we see that the 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. 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 simply 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 parallel updates to $\theta_j^{(t)}$, $j = 1, \dots, d$ that follow 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). An example of cyclic versus parallelized backfitting is given in Section 6.3.

6 Experiments

We provide empirical experiments to complement the algorithms and theory developed in the previous sections. We first compare additive trend filtering to additive smoothing splines on synthetic and real data, and then compare the cyclic and parallel backfitting methods. All experiments were performed in R. For the univariate trend filtering solver, we used the `trendfilter` function in the `glmgen` package, an implementation of the fast ADMM algorithm in Ramdas & Tibshirani (2016). For the univariate smoothing spline solver, we used the `smooth.spline` function in base R.

6.1 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 $\|f_{0j}\|_{n,j} = 1$, for $j = 1, \dots, d$. The responses were generated according to $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^2 f_{0j}(X_j^i), \sigma^2)$, $i = 1, \dots, 2500$. 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 4 shows a comparison of the MSE curves from additive trend filtering in (8) (of quadratic order $k = 2$) and additive smoothing splines in (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.

The right panel of Figure 4 shows the optimized MSE 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 1 to 16, 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 slightly 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, which is not surprising, since for high SNR it is able to better capture the heterogeneity in the component functions.

Lastly, in Appendix B, we present results from an experimental setup that mimics that in this subsection, except with the component functions f_{0j}, $j = 1, \dots, d$ having homogeneous smoothness throughout. The results show that additive trend filtering and additive smoothing splines perform nearly exactly the same.

The high-level message conveyed by these simulations is as follows: additive trend filtering can outperform additive linear smoothers like additive smoothing splines when there is heterogeneity in smoothness (either within or between component functions); when heterogeneity is not present, its performance does not seem to degrade compared to additive linear smoothers.

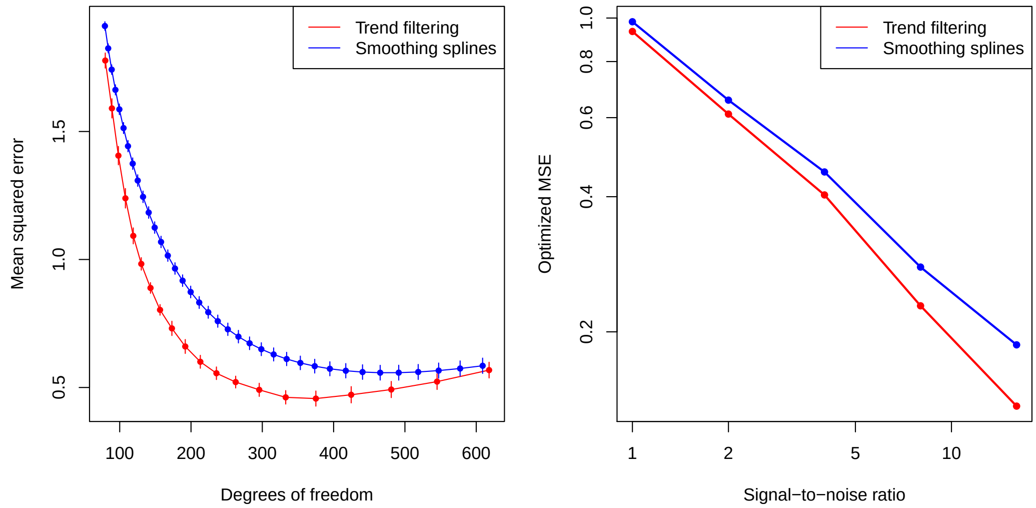

Figure 4: *The left panel displays the MSE curves for additive trend filtering and additive smoothing splines computed over 10 repetitions from the heterogeneous smoothness simulation setup described in Section 6.1, with the SNR set to 4. Vertical segments denote* ± 1 *standard deviations. The right panel shows the optimized MSE (minimum MSE over their respective regularization paths) for the two methods as the SNR varies from 1 to 16, in equally spaced values on the log scale.*

6.2 Electricity load data

We now consider a data set, over a 3 year period (2004–2007), of $n = 18792$ hourly total electricity load readings from ISO New England, a regional transmission company in the United States (this data is available through the standard MATLAB distribution). We used $d = 4$ variables to predict the electricity load: dry bulb temperature, dew point temperature, time of year, and hour of day, each of which was standardized to lie in $[0, 1]$. We partitioned the data into 10 folds, and instead of the standard cross-validation scheme, we trained on one fold, tested on the remaining 9, and so on. (This was done to highlight the differences in predictive accuracy between additive trend filtering and additive smoothing splines in a more data-constrained setting. Given the large data set size, when standard cross-validation was run, both additive trend filtering and additive smoothing splines achieved minimal cross-validated MSEs at nearly saturated models, i.e., at models with very little regularization, and we saw this data-rich case as a less interesting comparison.)

Figure 5 displays the 10-fold cross-validated MSEs for quadratic order additive trend filtering (8) and cubic order additive smoothing splines (1), calculated over a range of tuning parameters for each method. On the x-axis is *relative optimism*, defined as $(\mathrm{MSE}_{\mathrm{test}} - \mathrm{MSE}_{\mathrm{train}})/\mathrm{MSE}_{\mathrm{train}}$, where $\mathrm{MSE}_{\mathrm{test}}$ is the MSE on the test set, and $\mathrm{MSE}_{\mathrm{train}}$ is the MSE on the training set (averaged over the 10 folds). This metric allows us to put the estimates across different tuning parameter values from the two methods on roughly equal footing (in terms of their complexity). It acts as an alternative to effective degrees of freedom, and is generally easier to compute.[1] We see additive trend filtering achieves a lower minimum cross-validated MSE than additive smoothing splines, indicating that it is able to better adapt to possible heterogeneity within or across the components (the contributions of dry bulb temperature, dew point temperature, etc., to the prediction of electricity load).

[1] In fact, effective degrees of freedom is a highly related metric; when the test and training sets are each of size n_0, and other conditions hold, the effective degrees of freedom equals the expectation of $n_0(\mathrm{MSE}_{\mathrm{test}} - \mathrm{MSE}_{\mathrm{train}})/(2\sigma^2)$.

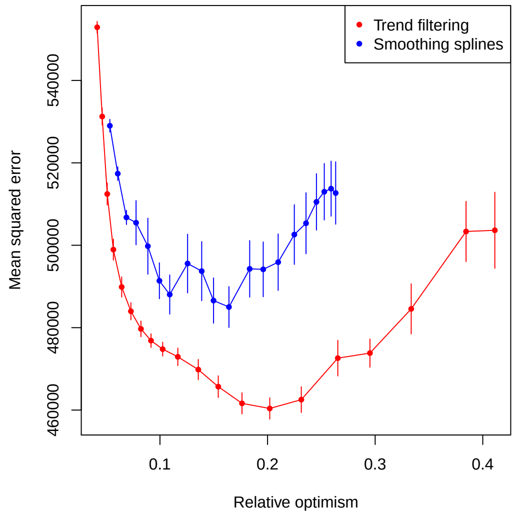

Figure 5: *Cross-validated MSE curves for additive trend filtering and additive smoothing splines, computed over 10 folds of the electricity load data set. Vertical segments denote* ± 1 *standard deviations. The MSE curves are parametrized by relative optimism (as defined in the text).*

6.3 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 6.1, except with $n = 2000$ and $d = 24$. We computed the additive trend filtering estimate (8) (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).

Figure 6 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, compared to the cyclic method. We should expect these computational gains to be even larger as the number of components d grows.

7 Discussion and extensions

We have studied additive models built around the univariate trend filtering estimator, i.e., defined by penalizing according to the sums of ℓ_1 norms of discrete derivatives of the component functions. We examined basic properties of these additive models, like extrapolation of the fitted values to a d-dimensional surface, uniqueness of the component fits, and characterization of effective degrees of freedom of the total additive fit. We derived an upper bound, on the order of $dn^{-(2k+2)/(2k+3)}$, for the estimation error of kth order additive trend filtering, in a problem setting where the underlying regression function is additive and has components whose kth derivatives are of bounded variation. We showed that this bound is sharp in rate, by establishing a matching minimax lower bound, and showed that additive linear smoothers (e.g., additive smoothing splines) can only achieve a rate of $dn^{-(2k+1)/(2k+2)}$ over the same class of functions. In terms of computation, we devised a provably

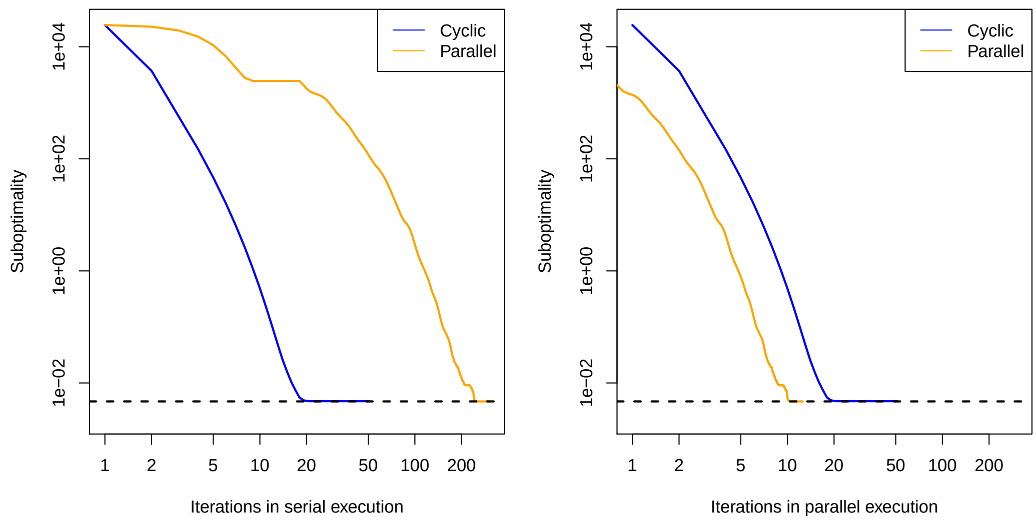

Figure 6: *Suboptimality in criterion value versus iteration number for the cyclic (Algorithm 1) and parallel (Algorithm 2) backfitting methods, on a synthetic data set with $n = 2000$ and $d = 24$. On the left, iterations for the parallel method are counted as if "ideal" parallelization is used, where 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).*

convergent parallel backfitting algorithm, and showed that for additive trend filtering, it can lead to large savings in computation time over the usual cyclic backfitting scheme if we make full use of the parallelism allowed by the algorithm.

We finish by describing some natural extensions of our work.

7.1 Mixed polynomial degrees

A natural extension of the additive trend filtering formulation in (8) would be to allow each component j to have to have its own polynomial degree $k_j \geq 0$, i.e., to replace the penalty term by $\lambda \sum_{j=1}^d \|D^{(X_j,k_j+1)} S_j \theta_j\|_1$. Computationally, this mixed-degree additive trend filtering model does not pose much more difficulty, as we can still use either of the backfitting algorithms (Algorithms 1 or 2) with only minor modifications (essentially, we just modify the univariate trend filtering solver appropriately, for each $j = 1, \ldots, d$).

Theoretically, however, the analysis of such a mixed-degree additive model, over say the mixed-degree function class $\{\sum_{j=1}^d f_j : \mathrm{TV}(f_j^{(k_j)}) \leq C_0,\ j = 1, \ldots, d\}$, seems like it would be substantially more challenging. It seems intuitively reasonable that (when $C_0 > 0$ is taken to be a constant), the attained error rate over this class should be $\sum_{j=1}^d n^{-(2k_j+2)/(2k_j+3)}$. van de Geer & Muro (2015) consider an additive model with two components that have different degrees of smoothness (i.e., lie in different function classes), and show an analogous result for this case; but, as noted by these authors, the extension of this work to the case of a (growing) number components d seems difficult.

7.2 Exponential families

Another natural extension is to consider generalized additive models with trend filtering penalties. In this setting, the conditional distribution of $Y^i|X^i$ is assumed to be in exponential family form, with link $g(\mathbb{E}(Y^i|X^i) = \mu + \sum_{j=1}^d f_{0j}(X_j^i)$, for $i = 1, \ldots, n$. The problem (8), a special case of this setup where the exponential family distribution is simply Gaussian, now becomes more generally

$$
\begin{aligned}
(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_d) = \underset{\theta_0 \in \mathbb{R},\, \theta_1,\ldots,\theta_d \in \mathbb{R}^n}{\mathrm{argmin}} \quad & -\sum_{i=1}^n \big(Y^i z^i + b(z^i)\big) + \lambda \sum_{j=1}^d \big\|D^{(X_j, k+1)} S_j \theta_j\big\|_1 \\
\text{subject to} \quad & z = \theta_0 + \sum_{j=1}^d \theta_j, \quad \mathbb{1}^T \theta_j = 0, \quad j = 1, \ldots, d.
\end{aligned} \tag{43}
$$

In the above, b is a convex function, e.g., $b(t) = t^2/2$ for the Gaussian family, $b(t) = \log(1 + e^t)$ for Bernoulli, and $b(t) = e^t$ for Poisson, and thus (43) is a convex program. Block coordinate descent (backfitting) can still be applied, and now each block update can be done using a proximal Newton expansion of the criterion into a simple weighted quadratic with diagonal Hessian, i.e., the inner loop here amounts to solving a weighted Gaussian univariate trend filtering problem.

7.3 High dimensions

In a high-dimensional setting, where d is comparable to or possibly even much larger than n, it is imperative to apply more regularization on top of the componentwise smoothness enforced in (8). One strategy is to estimate a *sparse additive model*, where regularization is applied to zero out all but a reasonably small number of the component functions. Lin & Zhang (2006), Ravikumar et al. (2009), Meier et al. (2009), Koltchinskii & Yuan (2010), Raskutti et al. (2012) all proposed such sparse high-dimensional estimators, but did not consider locally adaptive penalties like the ones we considered here. As an extension of (8) to high-dimensional inputs, it is natural to define the *sparse additive trend filtering* estimator as

$$
\begin{aligned}
(\hat{\theta}_1, \ldots, \hat{\theta}_d) = \underset{\theta_1,\ldots,\theta_d \in \mathbb{R}^n}{\mathrm{argmin}} \quad & \frac{1}{2}\Big\|Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j\Big\|_2^2 + \lambda_1 \sum_{j=1}^d \big\|D^{(X_j, k+1)} S_j \theta_j\big\|_1 + \lambda_2 \sum_{j=1}^d \|\theta_j\|_2 \\
\text{subject to} \quad & \mathbb{1}^T \theta_j = 0, \quad j = 1, \ldots, d,
\end{aligned} \tag{44}
$$

where $\lambda_1, \lambda_2 \geq 0$ are tuning parameters, and a group lasso penalty on the components is used in order to enforce sparsity (one could also use the ℓ_∞ norm in place of the ℓ_2 norm here). Petersen et al. (2016) studied the estimator in (44) when $k = 0$, i.e., the sparse additive fused lasso. They showed that backfitting for this estimator is especially simple and efficient, since the proximal operator for the sum of a fused lasso penalty and a group lasso penalty can be decomposed into a composition of their individual proximal operators; fortunately, their computational result is actually more general and covers the trend filtering penalty of an arbitrary order k. Theoretical analysis for the estimator in (44) is still an open problem.

Acknowledgements

We are very thankful to Garvesh Raskutti for his generous help and insight on various subtle issues to do with entropy numbers and Rademacher complexities, as well as boundedness assumptions. We also thank Martin Wainwright for generously sharing his unpublished book with us.

A Appendix

A.1 Fast extrapolation

We discuss extrapolation using the fitted functions $\hat{f}_j$, $j = 1, \ldots, d$ from additive trend filtering (11), as in (14). We must compute the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$, whose block form is given in (15), (16). Clearly, the computation of $\hat{b}_j$ in (16) 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 (15), it can be seen from the structure of $C^{(X_j,k+1)}$ as described in Wang et al. (2014) that

$$(\hat{a}_j)_1 = (S_j \hat{\theta}_j)_1,$$
$$(\hat{a}_j)_\ell = \frac{1}{(\ell-1)!}\left[\operatorname{diag}\left(\frac{1}{X_j^\ell - X_j^1}, \ldots, \frac{1}{X_j^n - X_j^{n-\ell+1}}\right) D^{(X_j,\ell-1)} S_j \hat{\theta}_j\right]_1, \quad \ell = 2, \ldots, k+1,$$

which takes only $O(1)$ operations (again treating k as constant, and now using the bandedness of each $D_j^{(X_j,\ell-1)}$, $\ell = 2, \ldots, k+1$). In total then, computing the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$ requires $O(n)$ operations, and computing $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \alpha_d)$ requires $O(nd)$ operations.

After having computed $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \alpha_d)$, which only needs to be done once, a prediction at a new point $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ 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.

A.2 Proof of Lemma 2

We begin by eliminating the constraint in the additive trend filtering problem (8), rewriting it as

$$\min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \; \frac{1}{2}\left\| MY - \sum_{j=1}^d M\theta_j \right\|_2^2 + \lambda \sum_{j=1}^d \left\| D^{(X_j,k+1)} S_j M \theta_j \right\|_1,$$

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 1, yields the problem

$$\min_{a \in \mathbb{R}^{(k+1)d},\, b \in \mathbb{R}^{(n-k-1)d}} \quad \frac{1}{2}\left\| MY - M\sum_{j=1}^d P_j a_j - M \sum_{j=1}^d K_j b_j \right\|_2^2 + \lambda k! \sum_{j=1}^d \|b_j\|_1,$$

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

$$\min_{a \in \mathbb{R}^{kd},\, b \in \mathbb{R}^{(n-k-1)d}} \quad \frac{1}{2}\left\| MY - M\sum_{j=1}^d \tilde{P}_j a_j - M \sum_{j=1}^d K_j b_j \right\|_2^2 + \lambda k! \sum_{j=1}^d \|b_j\|_1,$$

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 (8) 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 $\hat{a} = (\hat{a}_1, \dots \hat{a}_d)$ solves

$$\min_{a \in \mathbb{R}^{kd}} \; \frac{1}{2} \left\| \left(MY - M \sum_{j=1}^d K_j \hat{b}_j \right) - \tilde{P} a \right\|_2^2, \tag{45}$$

where $\tilde{P}$ is as defined in (17), and $\hat{b} = (\hat{b}_1, \dots, \hat{b}_d)$ solves

$$\min_{b \in \mathbb{R}^{(n-k-1)d}} \; \frac{1}{2} \left\| UU^T MY - UU^T M \sum_{j=1}^d K_j b_j \right\|_2^2 + \lambda k! \|b\|_1,$$

where UU^T is the projection orthogonal to the column space of $\tilde{P}$, i.e., it solves

$$\min_{b \in \mathbb{R}^{(n-k-1)d}} \; \frac{1}{2} \big\| U^T MY - \tilde{K} b \big\|_2^2 + \lambda k! \|b\|_1, \tag{46}$$

where $\tilde{K}$ is as in (18). Since problem (46) 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 (45) 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.

A.3 Derivation of additive trend filtering dual

As in the proof of Lemma 2, we begin by rewriting the problem (8) as

$$\min_{\theta_1, \dots, \theta_d \in \mathbb{R}^n} \; \frac{1}{2} \left\| MY - \sum_{j=1}^d M\theta_j \right\|_2^2 + \lambda \sum_{j=1}^d \|D_j S_j M \theta_j\|_1,$$

where $M = I - \mathbb{1}\mathbb{1}^T/n$. Then, we reparametrize the above problem,

$$\begin{aligned} \min_{\substack{\theta_1, \dots, \theta_d \in \mathbb{R}^n \\ w \in \mathbb{R}^n, \; z \in \mathbb{R}^{md}}} \quad & \frac{1}{2} \|MY - w\|_2^2 + \lambda \sum_{j=1}^d \|z_j\|_1 \\ \text{subject to} \quad & w = \sum_{j=1}^d M\theta_j, \quad z_j = D_j S_j M \theta_j, \quad j = 1, \dots, d, \end{aligned}$$

and form the Lagrangian

$$L(\theta, w, z, u, v) = \frac{1}{2} \|MY - w\|_2^2 + \lambda \sum_{j=1}^d \|z_j\|_1 + u^T \left(w - \sum_{j=1}^d M\theta_j \right) + \sum_{j=1}^d v_j^T (D_j S_j M \theta_j - z_j).$$

Minimizing the Lagrangian L over all θ, z yields the dual problem

$$\begin{aligned} \max_{\substack{u \in \mathbb{R}^n \\ v_1, \dots, v_d \in \mathbb{R}^m}} \quad & \frac{1}{2} \|MY\|_2^2 - \frac{1}{2} \|MY - u\|_2^2 \\ \text{subject to} \quad & u = S_j D_j^T v_j, \quad \|v_j\|_\infty \le \lambda, \quad j = 1, \dots, d. \end{aligned}$$

The claimed dual problem (19) is just the above, rewritten in an equivalent form.

A.4 Proof of Lemma 3

We first eliminate the equality constraint in (8), rewriting this problem, as was done in the proof of Lemma 2, as

$$\min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^d} \ \frac{1}{2}\bigg\| MY - \sum_{j=1}^d M\theta_j \bigg\|_2^2 + \lambda \sum_{j=1}^d \|D_j S_j \theta_j\|_1,$$

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

$$\dim\big(T\mathrm{null}(D_{-A})\big) = \bigg(\sum_{j=1}^d |A_j|\bigg) + kd,$$

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 1 and 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.

A.5 Preliminaries for the proof of Theorem 1

Before giving the proof, we introduce helpful notation on projection operators, and we also collect important preliminary results. Over functions $m : [0,1]^d \to \mathbb{R}$, we define projection operators Π_k and $\Pi_{k,n}$ onto the space $\mathcal{P}_k$ of additive kth degree polynomials, with respect to the L_2 norm $\|\cdot\|_2$ and empirical norm $\|\cdot\|_n$, respectively. To be more precise,

$$\mathcal{P}_k = \bigg\{ \sum_{j=1}^d p_j : p_j(x_j) \text{ is a } k\text{th degree polynomial in } x_j,\ j = 1,\ldots,d \bigg\}, \tag{47}$$

and the projection operators are defined as

$$\Pi_k(m) = \mathop{\mathrm{argmin}}_{p \in \mathcal{P}_k} \|m - p\|_2, \quad \text{and} \quad \Pi_{k,n}(m) = \mathop{\mathrm{argmin}}_{p \in \mathcal{P}_k} \|m - p\|_n.$$

The orthocomplement projection operators are defined in the usual way, by

$$\Pi_k^\perp(m) = m - \Pi_k(m), \quad \text{and} \quad \Pi_{k,n}^\perp(m) = m - \Pi_{k,n}(m).$$

We note that the empirical projection operators $\Pi_{k,n}, \Pi_{k,n}^\perp$ are well-defined if and only if $\tilde{P}$ in (17) has full column rank. When $n > kd$, this happens almost surely with respect to the distribution of the inputs X^i, $i = 1,\ldots,n$, under Assumption A1. The explanation is as follows[2]: the determinant of any square subblock of $\tilde{P}$ is a polynomial function of the elements X_j^i, $i = 1,\ldots,n$, $j = 1,\ldots,d$, and by Lemma 1 of Okamoto (1973), the roots of any polynomial (that is not identically zero) form a set of Lebesgue measure zero. In all usages of $\Pi_{k,n}, \Pi_{k,n}^\perp$, throughout, we implicitly assume that the inputs lie outside of (the inverse image of) this null set, to ensure that the empirical projection operators are well-defined.

[2]We thank Mathias Drton for pointing this out.

It is not hard to verify that the decomposability property of the L_2 norm, in (25), implies a certain decomposition of the L_2 projection operators $\Pi_k, \Pi_k^\perp$ over additive functions:

$$\Pi_k\bigg(\sum_{j=1}^d m_j\bigg) = \sum_{j=1}^d \widetilde{\Pi}_{k,j}(m_j), \quad \Pi_k^\perp\bigg(\sum_{j=1}^d m_j\bigg) = \sum_{j=1}^d \widetilde{\Pi}_{k,j}^\perp(m_j), \tag{48}$$

where $\widetilde{\Pi}_{k,j}$ is the projection operator onto the degree k univariate polynomials with respect to the univariate L_2 norm $\|\cdot\|_{2,j}$ (based on the measure Q_j), and $\widetilde{\Pi}_{k,j}^\perp$ is its orthocomplement projector, for $j = 1, \ldots, d$. Furthermore, note that (25) and (48) together imply

$$\|\Pi_k^\perp(\hat{\Delta})\|_2^2 = \sum_{j=1}^d \|\widetilde{\Pi}_{k,j}^\perp(\hat{\Delta}_j)\|_{2,j}^2. \tag{49}$$

Other useful properties of the projection operators are given next; we will rely on several of these properties in what follows (though not all of them, but we list them anyway for completeness).

Lemma 4. *For any function $m : [0,1]^d \to \mathbb{R}$, the projection operators satisfy the properties:*

(i) $\Pi_{k,n}(\Pi_k(m)) = \Pi_k(m)$;

(ii) $\Pi_k(\Pi_{k,n}(m)) = \Pi_{k,n}(m)$;

(iii) $\Pi_{k,n}^\perp(\Pi_k^\perp(m)) = \Pi_{k,n}^\perp(m)$;

(iv) $\Pi_k^\perp(\Pi_{k,n}^\perp(m)) = \Pi_k^\perp(m)$;

(v) $(\Pi_{k,n} - \Pi_k)(m) = (\Pi_k^\perp - \Pi_{k,n}^\perp)(m) = \Pi_{k,n}(\Pi_k^\perp(m))$;

(vi) $(\Pi_k - \Pi_{k,n})(m) = (\Pi_{k,n}^\perp - \Pi_k^\perp)(m) = \Pi_k(\Pi_{k,n}^\perp(m))$;

(vii) $\|\Pi_{k,n}^\perp(m)\|_2^2 = \|\Pi_k^\perp(m)\|_2^2 + \|(\Pi_{k,n} - \Pi_k)(m)\|_2^2$;

(viii) $\|\Pi_k^\perp(m)\|_n^2 = \|\Pi_{k,n}^\perp(m)\|_n^2 + \|(\Pi_k - \Pi_{k,n})(m)\|_n^2$;

(ix) $\|\Pi_k^\perp(m)\|_2 \le \|\Pi_{k,n}^\perp(m)\|_2$;

(x) $\|\Pi_{k,n}^\perp(m)\|_n \le \|\Pi_k^\perp(m)\|_n$;

(xi) $\|(\Pi_{k,n} - \Pi_k)(m)\|_2 \le \|\Pi_{k,n}^\perp(m)\|_2$;

(xii) $\|(\Pi_k - \Pi_{k,n})(m)\|_n \le \|\Pi_k^\perp(m)\|_n$;

Proof. For property (i), note that $\Pi_k(m)$ is already in $\mathcal{P}_k$, and $\Pi_{k,n}$ acts as the identity map on $\mathcal{P}_k$. Property (ii) follows similarly. For property (iii), we write $\Pi_{k,n}^\perp(\Pi_k^\perp(m)) = \Pi_{k,n}^\perp(m) - \Pi_{k,n}^\perp(\Pi_k(m))$ by linearity of the projection operator, and notice that the second term here is zero by property (i). Property (iv) follows similarly.

In property (v), the first equality is immediate by definition of the orthocomplement projectors. The second equality follows by expanding $\Pi_k^\perp(m) = m - \Pi_k(m)$, using linearity, and then property (i). Property (vi) follows similarly. For property (vii), observe that

$$\Pi_{k,n}^\perp(m) = \Pi_k^\perp(m) + (\Pi_k - \Pi_{k,n})(m),$$

and the two terms on the right are orthogonal with respect to the L_2 inner product, so taking the squared L_2 norm of both sides gives the result. Property (viii) follows similarly. Properties (ix) and (x) follow from (vii) and (viii), respectively, and the same with properties (xi) and (xii). $\square$

Now we present a series of lemmas that will be useful for the proof of Theorem 1. Here are two simple results, on shifting exponents around sums and products.

Lemma 5. *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 6. *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 $ab^{1-w/2} \leq a^{2/(1+w/2)}$. ☐

In the remaining lemmas of this subsection (as well as the proof of Theorem 1), the univariate function classes

$$\mathcal{M}_j(\rho) = \widetilde{\Pi}_{k,j}^{\perp}(B_J(\rho)) = \big\{\widetilde{\Pi}_{k,j}^{\perp}(g) : J(g) \leq \rho\big\}, \quad j = 1, \ldots, d \tag{50}$$

will play an important role, as will the multivariate function classes

$$\mathcal{M}(\rho) = \left\{\Pi_k^{\perp}\left(\sum_{i=1}^d m_j\right) : \sum_{j=1}^d J(m_j) \leq \rho\right\}, \tag{51}$$

$$\overline{\mathcal{M}}(\rho) = \left\{\sum_{i=1}^d m_j : \sum_{j=1}^d J(m_j) \leq \rho\right\} \cap B_\infty(1). \tag{52}$$

Also, we will implicitly assume (in the lemmas that follow) A1, A2, B1, C1–C3, as in Theorem 1.

We start with a result on orthonormal polynomials[3] in preparation for considering the sup norm boundedness of the function classes $\mathcal{M}_j(1)$, $j = 1, \ldots, d$.

Lemma 7. *Given an integer $\kappa \geq 0$. For $j = 1, \ldots, d$, let $\phi_{j0}, \ldots, \phi_{j\kappa}$ be an orthonormal basis for the space of univariate polynomials on $[0,1]$ of degree κ, formed by running the Gram-Schmidt procedure on the polynomials $1, t, \ldots, t^\kappa$, with respect to the $L_2(Q_j)$ inner product. Hence, for $j = 1, \ldots, d$ and $\ell = 0, \ldots, \kappa$, $\phi_{j\ell}$ is an ℓth degree polynomial with leading coefficient $a_{j\ell} > 0$, orthogonal (in $L_2(Q_j)$) to all polynomials of degree less than ℓ.*

Now define for $j = 1, \ldots, d$ and $t \in [0,1]$,

$$\Phi_{j,\kappa,0}(t) = \phi_{j\kappa}(t)q_j(t),$$
$$\Phi_{j,\kappa,\ell+1}(t) = \int_0^t \Phi_{j,\kappa,\ell}(u)\,du, \quad \ell = 0, \ldots, \kappa,$$

where q_j denotes the density of Q_j. Then the following two relations hold, for $j = 1, \ldots, d$:

$$\Phi_{j,\kappa,\ell}(1) = \begin{cases} 0 & \text{for } \ell = 1, \ldots, \kappa, \\ \frac{(-1)^\kappa}{a_{j\kappa}\kappa!} & \text{for } \ell = \kappa + 1, \end{cases} \tag{53}$$

and

$$a_{j\kappa}\kappa!|\Phi_{j,\kappa,\kappa}(t)| \leq \binom{2\kappa}{\kappa}\sqrt{\frac{b_2}{b_1}}, \quad t \in [0,1]. \tag{54}$$

[3] We thank Dejan Slepcev for his help in conceiving the result in Lemma 7, and its role in establishing the desired boundedness result in Lemma 8.

Remark 13 (Special case: uniform measure and Rodrigues' formula). *If Q_j is the Lebesgue measure on $[0,1]$, then we can just take $\phi_{j0},\dots,\phi_{j\kappa}$ to be the Legendre polynomials, shifted to $[0,1]$ and normalized appropriately. Invoking Rodrigues' formula to express these functions,*

$$\phi_{j\ell}(t) = \frac{\sqrt{2\ell+1}}{\ell!}\frac{d^\ell}{dt^\ell}(t^2-t)^\ell, \quad \ell = 0,\dots\kappa,$$

the results in Lemma 7 can be easily directly verified.

Proof. Fix an arbitrary $j = 1,\dots,d$. First, we use induction to show that for $t \in [0,1]$,

$$\Phi_{j,\kappa,\ell}(t) = \int_0^t \phi_{j\kappa}(u)\frac{(t-u)^{\ell-1}}{(\ell-1)!}q_j(u)\,du, \quad \ell = 1,\dots,\kappa+1. \tag{55}$$

This statement holds for $\ell = 1$ by definition of $\Phi_{j,\kappa,0}, \Phi_{j,\kappa,1}$. Assume it holds at some $\ell > 1$. Then

$$\begin{aligned}
\Phi_{j,\kappa,\ell+1}(t) &= \int_0^t \int_0^u \phi_{j\kappa}(v)\frac{(u-v)^{\ell-1}}{(\ell-1)!}q_j(v)\,dv\,du \\
&= \int_0^t \phi_{j\kappa}(v)\bigg(\int_v^t \frac{(u-v)^{\ell-1}}{(\ell-1)!}\,du\bigg)q_j(v)\,dv \\
&= \int_0^t \phi_{j\kappa}(v)\frac{(t-v)^\ell}{\ell!}q_j(v)\,dv,
\end{aligned}$$

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 (55) shows that $\Phi_{j,\kappa,\ell}(1)$ is the $L_2(Q_j)$ inner product of $\phi_{j\kappa}$ and an $(\ell-1)$st degree polynomial, for $\ell = 1,\dots,\kappa$. As $\phi_{j\kappa}$ is orthogonal to all polynomials of degree less than κ, we have $\Phi_{j,\kappa,\ell}(1) = 0$, $\ell = 1,\dots,\kappa$. For $\ell = \kappa+1$, note that this same orthogonality along with (55) also shows

$$\Phi_{j,\kappa,\kappa+1}(1) = \bigg\langle \phi_{j\kappa}, \frac{(-1)^\kappa}{a_{j\kappa}\kappa!}\phi_{j\kappa}\bigg\rangle_{2,j} = \frac{(-1)^\kappa}{a_{j\kappa}\kappa!}.$$

This establishes the statement in (53).

As for (54), note that if $\kappa = 0$ then the claim holds, as the left-hand side is 1 and the right-hand side is always larger than 1. Hence consider $\kappa \ge 1$. From (55), we have, for any $t \in [0,1]$,

$$\begin{aligned}
|\Phi_{j,\kappa,\kappa}(t)| &\le \int_0^t |\phi_{j\kappa}(u)|\frac{(t-u)^{\kappa-1}}{(\kappa-1)!}q_j(u)\,du \\
&\le \bigg(\int_0^t \phi_{j\kappa}^2(u)q_j(u)\,du\bigg)^{1/2}\bigg(\int_0^t \frac{(t-u)^{2\kappa-2}}{(\kappa-1)!^2}q_j(u)\,du\bigg)^{1/2} \\
&\le \frac{\sqrt{b_2}}{(\kappa-1)!\sqrt{2\kappa-1}},
\end{aligned} \tag{56}$$

where in the second line we used the Cauchy-Schwartz inequality, and in the third line we used the fact that $\phi_{j\kappa}$ has unit norm, and the upper bound b_2 on the density q_j, from Assumption A1. Next we bound $a_{j\kappa}$. Let p be the projection of x^κ onto the polynomials of degree less than κ, with respect to the $L_2(Q_j)$ inner product. Then we have $\phi_{j\kappa} = (x^\kappa - p)/\|x^\kappa - p\|_{2,j}$, so its leading coefficient is $a_{j\kappa} = 1/\|x^\kappa - p\|_{2,j}$. Consider

$$\begin{aligned}
\|x^\kappa - p\|_{2,j} &\ge \sqrt{b_1}\bigg(\int_0^1 (x^\kappa - p)^2(t)\,dt\bigg)^{1/2} \\
&\ge \sqrt{b_1}\bigg(\int_0^1 P_\kappa^2(t)\,dt\bigg)^{1/2} = \frac{\sqrt{b_1}}{\sqrt{2\kappa+1}\binom{2\kappa}{\kappa}}.
\end{aligned} \tag{57}$$

In the first line we used the lower bound b_1 on q_j from Assumption A1. In the second line we used the fact the the Legendre polynomial P_κ of degree κ, shifted to $[0,1]$ but unnormalized, is the result from projecting out $1, t, \dots, t^{\kappa-1}$ 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 (56) and (57) gives the result (54). $\square$

We use Lemma 7 to construct a uniform sup norm bound on the univariate functions in $\mathcal{M}_j(1)$, $j = 1, \dots, d$.

Lemma 8. *There exists a constant $R \geq 1$, that depends only on k, b_1, b_2, L, such that*

$$\|g\|_\infty \leq R, \quad \textit{for all } g \in \bigcup_{j=1}^d \mathcal{M}_j(1),$$

where $\mathcal{M}_j(1)$, $j = 1, \dots, d$ are as defined in (50).

Proof. Fix an arbitrary $j = 1, \dots, d$ and $g \in \mathcal{M}_j(1)$. By integration by parts and repeated application of Lemma 7, we have

$$0 = a_{j\ell}\ell! \cdot \langle g, \phi_{j\ell} \rangle_{2,j} = \int_0^1 g^{(\ell)}(t) w_{j\ell}(t)\, dt, \quad \ell = 0, \dots, k, \tag{58}$$

where $w_{j\ell}(t) = (-1)^\ell a_{j\ell}\ell! \Phi_{j,\ell,\ell}(t)$, $\ell = 0, \dots, k$, and by properties (53), (54) of Lemma 7,

$$\int_0^1 w_{j\ell}(t)\, dt = 1, \quad \int_0^1 |w_{j\ell}(t)|\, dt \leq \binom{2\ell}{\ell} \sqrt{\frac{b_2}{b_1}}, \quad \ell = 0, \dots, k. \tag{59}$$

Now we will prove the following by induction:

$$\|g^{(\ell)}\|_\infty \leq L \left(\frac{b_2}{b_1}\right)^{(k-\ell+1)/2} \prod_{i=\ell}^k \binom{2i}{i}, \quad \ell = 0, \dots, k. \tag{60}$$

Starting at $\ell = k$, the statement holds because, using (58), for almost every $t \in [0,1]$,

$$\begin{aligned}
|g^{(k)}(t)| &= \left| g^{(k)}(t) - \int_0^1 g^{(k)}(u) w_{jk}(u)\, du \right| \\
&= \left| \int_0^1 \Big(g^{(k)}(t) - g^{(k)}(u)\Big) w_{jk}(u)\, du \right| \\
&\leq L \binom{2k}{k} \sqrt{\frac{b_2}{b_1}},
\end{aligned}$$

where in the second line we used the fact that the weight function integrates to 1 from (59), and in the third we used Assumption C2 and the upper bound on the integrated absolute weights from (59). Assume the statement holds at some $\ell < k$. Then again by (58), (59), for almost every $t \in [0,1]$,

$$\begin{aligned}
|g^{(\ell-1)}(t)| &= \left| \int_0^1 \Big(g^{(\ell-1)}(t) - g^{(\ell-1)}(u)\Big) w_{j,\ell-1}(u)\, du \right| \\
&\leq \left(\operatorname*{ess\,sup}_{0 \leq u < v \leq 1} |g^{(\ell-1)}(v) - g^{(\ell-1)}(u)| \right) \binom{2\ell-2}{\ell-1} \sqrt{\frac{b_2}{b_1}} \\
&= \left(\operatorname*{ess\,sup}_{0 \leq u < v \leq 1} \left| \int_u^v g^{(\ell)}(s)\, ds \right| \right) \binom{2\ell-2}{\ell-1} \sqrt{\frac{b_2}{b_1}} \\
&\leq L \left(\frac{b_2}{b_1}\right)^{(k-\ell+2)/2} \prod_{i=\ell-1}^k \binom{2i}{i},
\end{aligned}$$

the last line using $\operatorname{ess\,sup}_{0\le u<v\le 1} |\int_u^v g^{(\ell)}(s)\,ds| \le \|g^{(\ell)}\|_\infty$ and the inductive hypothesis. This verifies (60). Taking $\ell = 0$ in (60) and defining $R = L(b_2/b_1)^{(k+1)/2} \prod_{i=0}^k \binom{2i}{i}$ proves the lemma. $\square$

The next lemma is from Wainwright (2017), transcribed here for convenience. (Older results of similar flavor can be found in van de Geer (2000), Bartlett et al. (2005), Raskutti et al. (2012).)

Lemma 9 (Theorem 14.1 of Wainwright 2017)**.** *Let $\mathcal{F}$ be a class of functions, whose elements $f \in \mathcal{F}$ are of the form $f : D \to \mathbb{R}$. Let Q be a distribution over D, abbreviate $\|\cdot\|_2$ for the $L_2(Q)$ norm, and $\|\cdot\|_n$ for the $L_2(Q_n)$ norm, based on an i.i.d. sample of size n from Q. Assume that the class $\mathcal{F}$ is star-shaped, meaning that $f \in \mathcal{F} \Rightarrow \alpha f \in \mathcal{F}$ for all $0 \le \alpha \le 1$. Also, assume that there exists $b > 0$ such that $\|f\|_\infty \le b$ for all $f \in \mathcal{F}$. Define the local Rademacher complexity*

$$\mathcal{R}(\mathcal{F} \cap B_2(t)) = \mathbb{E}_{\xi,\sigma}\left[\sup_{\substack{f\in\mathcal{F},\\ \|f\|_2\le t}} \frac{1}{n}\left|\sum_{i=1}^n \sigma^i f(\xi^i)\right|\right], \tag{61}$$

where the expectation is taken over i.i.d. draws ξ^i, $i = 1,\dots,n$ from Q, and over i.i.d. Rademacher variables σ^i, $i = 1,\dots,n$. Define the critical radius

$$\tau_n = \inf\left\{t \ge 0 : \frac{\mathcal{R}(\mathcal{F} \cap B_2(t))}{t} \le \frac{t}{b}\right\}. \tag{62}$$

Then there exist universal constants $c_1, c_2 > 0$ such that, for any $t \ge \tau_n$,

$$\big|\|f\|_n^2 - \|f\|_2^2\big| \le \frac{1}{2}\|f\|_2^2 + \frac{t^2}{2}, \quad \text{for all } f \in \mathcal{F},$$

with probability at least $1 - c_1 \exp(-c_2 n t^2)$.

We use Lemma 9 to upper bound the empirical norm of a (univariate) function by its L_2 norm, uniformly over $\mathcal{M}_j(1)$, and uniformly over $j = 1,\dots,d$.

Lemma 10. *There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on k, K, w, R, where $R \ge 1$ is as in Lemma 8, such that for any $t \ge c_1 n^{-1/(2+w)}$ and $n \ge n_0$,*

$$\|g\|_{n,j} \le \sqrt{\frac{3}{2}}\|g\|_{2,j} + \frac{t}{\sqrt{2}}, \quad \text{for all } g \in \mathcal{M}_j(1), \text{ and } j = 1,\dots,d,$$

with probability at least $1 - c_2 d \exp(-c_3 n t^2)$, where $\mathcal{M}_j(1)$, $j = 1,\dots d$ are as in (50).

Proof. It is easy to see that each $\mathcal{M}_j(1)$, $j = 1,\dots,d$ is star-shaped (by the positive homogeneity of the seminorm J), and these classes are uniformly bounded by $R \ge 1$, from Lemma 8. Now fixing an arbitrary $j = 1,\dots,d$, we must analyze the local Rademacher complexity

$$\mathcal{R}(\mathcal{M}_j(1) \cap B_2(t)) = \mathbb{E}_{X_j,\sigma}\left[\sup_{\substack{g\in\mathcal{M}_j(1),\\ \|g\|_{2,j}\le t}} \frac{1}{n}\left|\sum_{i=1}^n \sigma^i g(X_j^i)\right|\right],$$

the expectation being taken over i.i.d. draws X_j^i, $i = 1,\dots,n$ from Q_j, as well as i.i.d. Rademacher variables σ^i, $i = 1,\dots,n$. Define the critical radius

$$\tau_{nj} = \inf\left\{t \ge 0 : \frac{\mathcal{R}(\mathcal{M}_j(1) \cap B_2(t))}{t} \le \frac{t}{R}\right\}.$$

We will prove that $\tau_{nj} \le c_1 n^{-1/(2+w)}$, for $j = 1, \dots, d$ and a constant $c_1 > 0$. We can then apply Lemma 9 (Theorem 14.1 of Wainwright (2017)) along with a union bound over $j = 1, \dots, d$, to give the desired result.

Consider the empirical local Rademacher complexity

$$\mathcal{R}_n(\mathcal{M}_j(1) \cap B_2(t)) = \mathbb{E}_\sigma \left[\sup_{\substack{g \in \mathcal{M}_j(1), \\ \|g\|_{2,j} \le t}} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i g(X_j^i) \right| \right],$$

the expectation being over σ^i, $i = 1, \dots, n$ alone. As we are considering $t \ge \tau_{nj}$, by Corollary 2.2 of Bartlett et al. (2005), we have

$$\{g \in \mathcal{M}_j(1) : \|g\|_{2,j} \le t\} \subseteq \{g \in \mathcal{M}_j(1) : \|g\|_{n,j} \le \sqrt{2}t\},$$

with probability at least $1 - 1/n$. Denote by $\mathcal{E}$ the event that this occurs. Then

$$\begin{aligned}
\mathcal{R}_n(\mathcal{M}_j(1) \cap B_2(t)) = \mathbb{E}_\sigma \left[\sup_{\substack{g \in \mathcal{M}_j(1), \\ \|g\|_{2,j} \le t}} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i g(X_j^i) \right| \right] &\le \mathbb{E}_\sigma \left[\sup_{\substack{g \in \mathcal{M}_j(1), \\ \|g\|_{n,j} \le \sqrt{2}t}} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i g(X_j^i) \right| \right] \\
&\le \frac{c}{\sqrt{n}} \int_0^{\sqrt{2}t} \sqrt{\log N(\delta, \|\cdot\|_{n,j}, \mathcal{M}_j(1))}\, d\delta,
\end{aligned}$$

on $\mathcal{E}$, where in the second line we used Dudley's entropy integral (Dudley 1967), with $c > 0$ being a universal constant. Observe that

$$\begin{aligned}
\log N(\delta, \|\cdot\|_{n,j}, \mathcal{M}_j(1)) &\le \log N\big(\delta, \|\cdot\|_{n,j}, B_J(1) \cap B_\infty(R)\big) \\
&\le \log N\big(\delta/(R), \|\cdot\|_{n,j}, B_J(1) \cap B_\infty(1)\big) \\
&\le K R^w \delta^{-w},
\end{aligned}$$

where in the first line we the sup norm boundedness from Lemma 8, in the second line we used the fact that $R \ge 1$, and in the last line we invoked Assumption C3, where $0 < w < 2$ and $K > 0$ are universal constants. We can now use this to bound the entropy integral, so that

$$\mathcal{R}_n(\mathcal{M}_j(1) \cap B_2(t)) \le \frac{K R^{w/2} c}{\sqrt{n}} \int_0^{\sqrt{2}t} \delta^{-w/2}\, d\delta = \frac{c}{\sqrt{n}} t^{1-w/2},$$

on $\mathcal{E}$, where in the second inequality, we redefined the constant $c > 0$, but importantly, it does not depend on the input points or on j. On $\mathcal{E}^c$, we have the trivial bound $\mathcal{R}_n(\mathcal{M}_j(1) \cap B_2(t)) \le R$, and so we can upper bound the local Rademacher complexity, splitting the expectation over $\mathcal{E}$ and $\mathcal{E}^c$,

$$\mathcal{R}(\mathcal{M}_j(1) \cap B_2(t)) = \mathbb{E}_{X_j} \mathcal{R}_n(\mathcal{M}_j(1) \cap B_2(t)) \le \frac{c t^{1-w/2}}{\sqrt{n}} + \frac{R}{n} \le \frac{c t^{1-w/2}}{\sqrt{n}},$$

where the second inequality holds when n is large enough, for another redefined constant $c > 0$ that still does not depend on j, as we may assume $t \ge n^{-1/2}$ without a loss of generality. We see that an upper bound on the critical radius τ_{nj} is given by the solution of

$$\frac{c t^{-w/2}}{\sqrt{n}} = \frac{t}{R},$$

which is $t = c n^{-1/(2+w)}$, for another redefined constant $c > 0$, that again does not depend on j. This completes the proof. $\square$

We again use Lemma 9, now to bound the L_2 norm of a (multivariate) function by its empirical norm, uniformly over $\overline{\mathcal{M}}(\rho_n)$, for a large enough scaling of the radius ρ_n.

Lemma 11. *There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on k, K, w, R, where $R \geq 1$ is as in Lemma 8, such that for any $t \geq c_1 \max\{\sqrt{d}, \sqrt{\rho_n}\} n^{-1/(2+w)}$ and $n \geq n_0$,*

$$\|f\|_2 \leq \sqrt{2}\|f\|_n + t, \quad \text{for all } f \in \overline{\mathcal{M}}(\rho_n),$$

with probability at least $1 - c_2 \exp(-c_3 n t^2)$, where $\overline{\mathcal{M}}(\rho_n)$ is as defined in (52).

Proof. It is easy to check that the class $\overline{\mathcal{M}}(\rho_n)$ is star-shaped (by the positive homogeneity of the seminorm J), and it is uniformly bounded in the sup norm by 1 by definition. In what follows, we will bound the local Rademacher complexity

$$\mathcal{R}(\overline{\mathcal{M}}(\rho_n) \cap B_2(t)) = \mathbb{E}_{X,\sigma} \left[\sup_{\substack{f \in \overline{\mathcal{M}}(\rho_n), \\ \|f\|_2 \leq t}} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i f(X^i) \right| \right],$$

where the expectation is taken over i.i.d. samples X^i, $i = 1, \dots, n$ from Q, and i.i.d. Rademacher variables σ^i, $i = 1, \dots, n$. We also will analyze the critical radius

$$\tau_n = \inf \left\{ t \geq 0 : \frac{\mathcal{R}(\overline{\mathcal{M}}(\rho_n) \cap B_2(t))}{t} \leq t \right\},$$

and we will show that $\tau_n \leq c_1 \max\{\sqrt{d}, \sqrt{\rho_n}\} n^{-1/(2+w)}$ for a constant $c_1 > 0$. We can then apply Lemma 9 (Theorem 14.1 of Wainwright (2017)) to establish the desired result.

Notice that we can decompose each $f \in \overline{\mathcal{M}}(\rho_n)$ as $f = m + p$, where $m = \Pi_k^\perp(f) \in \mathcal{M}(\rho_n)$, the function space defined in (51), and $p = \Pi_k(f) \in \mathcal{P}_k$, the space of additive kth degree polynomials in (47). By orthogonality of m, p, we see that $\|f\|_2 \leq t$ implies both $\|m\|_2 \leq t$ and $\|p\|_2 \leq t$. Therefore

$$\mathcal{R}(\overline{\mathcal{M}}(\rho_n) \cap B_2(t)) \leq \mathcal{R}(\mathcal{M}(\rho_n) \cap B_2(t)) + \mathcal{R}(\mathcal{P}_k \cap B_2(t)), \tag{63}$$

and we will bound each local Rademacher complexity on the right-hand side above individually.

First, we consider the local Rademacher complexity of the class $\mathcal{M}(\rho_n)$. We can write, recalling the L_2 orthogonality of the components of any $m = \sum_{j=1}^d m_j \in \mathcal{M}(\rho_n)$,

$$\begin{aligned} \sup_{m \in \mathcal{M}(\rho_n) \cap B_2(t)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i m(X^i) \right| &= \sup_{\substack{\|\alpha\|_1 \leq \rho_n, \\ \|\beta\|_2 \leq t}} \sup_{\substack{m_j \in \mathcal{M}_j(|\alpha_j|) \cap B_2(|\beta_j|), \\ j=1,\dots,d}} \left| \sum_{i=1}^n \sigma^i \sum_{j=1}^d m_j(X_j^i) \right| \\ &\leq \sup_{\substack{\|\alpha\|_1 \leq \rho_n, \\ \|\beta\|_2 \leq t}} \sum_{j=1}^d \sup_{m_j \in \mathcal{M}_j(|\alpha_j|) \cap B_2(|\beta_j|)} \frac{1}{n} \left| \sum_{i=1}^n \sigma^i m_j(X_j^i) \right|. \end{aligned}$$

We now bound the inner supremum above, for an arbitrary $j = 1, \dots, d$. We will denote by τ_{nj} the critical radius of $\mathcal{M}_j(1)$, and by $\mathcal{R}_n$ the empirical Rademacher complexity (where the expectation is taken over the Rademacher variables only). We use $a \vee b = \max\{a, b\}$, and use $c > 0$ for a constant whose value may change from line to line. Observe that

$$\begin{aligned}
\sup_{m_j\in\mathcal{M}_j(1)\cap B_2(t)} \frac{1}{n}\bigg|\sum_{i=1}^n \sigma^i m_j(X_j^i)\bigg| &\le c\bigg(\mathcal{R}_n(\mathcal{M}_j(1)\cap B_2(t)) + \sqrt{\frac{\log n}{n}}\sup_{m_j\in\mathcal{M}_j(1)\cap B_2(t)} \|m_j\|_n\bigg) \\
&\le c\bigg(\mathcal{R}(\mathcal{M}_j(1)\cap B_2(t)) + \frac{R\log n}{n} + \sqrt{\frac{\log n}{n}}\sup_{m_j\in\mathcal{M}_j(1)\cap B_2(t)} \|m_j\|_n\bigg) \\
&\le c\bigg(\frac{(t\vee\tau_{nj})^{1-w/2}}{\sqrt{n}} + \frac{R\log n}{n} + \sqrt{\frac{\log n}{n}}\sqrt{2}(t\vee\tau_{nj})\bigg) \\
&\le c\bigg(\frac{(t\vee r_n)^{1-w/2}}{\sqrt{n}} + \frac{(t\vee r_n)\sqrt{\log n}}{\sqrt{n}} + \frac{R\log n}{n}\bigg).
\end{aligned}$$

The first three inequalities in the above display hold with probability at least $1-1/3n^2$ each. The first one is by Theorem 3.6 in Wainwright (2017), and the second and third are by Lemma A.4 and Lemma 3.6 in Bartlett et al. (2005), respectively. The last inequality just uses the upper bound on the critical radius $\tau_{nj}\le c_2 r_n$, for a constant $c_2>0$ and $r_n=n^{-1/(2+w)}$, as established in the proof of Lemma 10. Hence the final result of the above display holds with probability at least $1-1/n^2$, and by a union bound, it holds with probability at least $1-d/n^2$ simultaneously over $j=1,\dots,d$. Let us call this event $\mathcal{E}$. Then on $\mathcal{E}$, by rescaling the result in the last display,

$$\begin{aligned}
&\sup_{m_j\in\mathcal{M}_j(|\alpha_j|)\cap B_2(|\beta_j|)} \frac{1}{n}\bigg|\sum_{i=1}^n \sigma^i m_j(X_j^i)\bigg| \\
&\quad\le c\bigg(\frac{|\alpha_j|^{w/2}(|\beta_j|+|\alpha_j|r_n)^{1-w/2}}{\sqrt{n}} + \frac{(|\beta_j|+|\alpha_j|r_n)\sqrt{\log n}}{\sqrt{n}} + \frac{R|\alpha_j|\log n}{n}\bigg), \quad j=1,\dots,d,
\end{aligned}$$

which implies that, on $\mathcal{E}$,

$$\begin{aligned}
&\sup_{m\in\mathcal{M}(\rho_n)\cap B_2(t)} \frac{1}{n}\bigg|\sum_{i=1}^n \sigma^i m(X^i)\bigg| \\
&\quad\le c\sup_{\substack{\|\alpha\|_1\le\rho_n,\\ \|\beta\|_2\le t}} \sum_{j=1}^d \bigg(\frac{|\alpha_j|^{w/2}(|\beta_j|+|\alpha_j|r_n)^{1-w/2}}{\sqrt{n}} + \frac{(|\beta_j|+|\alpha_j|r_n)\sqrt{\log n}}{\sqrt{n}} + \frac{R|\alpha_j|\log n}{n}\bigg) \\
&\quad\le \frac{c}{\sqrt{n}}\sup_{\substack{\|\alpha\|_1\le\rho_n,\\ \|\beta\|_2\le t}} \sum_{j=1}^d |\alpha_j|^{w/2}|\beta_j|^{1-w/2} + c\bigg(\frac{\rho_n r_n^{1-w/2}}{\sqrt{n}} + \frac{(\sqrt{d}t+\rho_n r_n)\sqrt{\log n}}{\sqrt{n}} + \frac{R\rho_n\log n}{n}\bigg) \\
&\quad\le \frac{c}{\sqrt{n}}\sup_{\|\alpha\|_1\le\rho_n} \bigg(\sum_{j=1}^d |\alpha_j|^{2w/(2+w)}\bigg)^{(2+w)/4} t^{1-w/2} + c\bigg(\frac{\sqrt{d}t\sqrt{\log n}}{\sqrt{n}} + \rho_n r_n^2\bigg) \\
&\quad\le c\bigg(\frac{\rho_n^{w/2}d^{(2-w)/4}t^{1-w/2}}{\sqrt{n}} + \frac{\sqrt{d}t\sqrt{\log n}}{\sqrt{n}} + \rho_n r_n^2\bigg).
\end{aligned}$$

In the third line here, we used the simple inequality from Lemma 5, below; in the fourth, we used Holder's inequality $a^Tb\le\|a\|_p\|b\|_q$, with $p=4/(2+w)$ and $q=4/(2-w)$, and also the fact that $\sqrt{\log n/n}\le r_n$ and $R\log n/n\le r_n^2$ for n large enough; in the fifth, we again used Holder's inequality, with $p=(2+w)/(2w)$ and $q=(2+w)/(2-w)$;

Meanwhile, on $\mathcal{E}^c$, we can use the simple bound $\|m\|_\infty\le\sum_{j=1}^d\|m_j\|_\infty\le R\sum_{j=1}^d J(m_j)\le R\rho_n$, where $R\ge1$ is the constant from Lemma 8, so that

$$\sup_{m\in\mathcal{M}(\rho_n)\cap B_2(t)} \frac{1}{n}\bigg|\sum_{i=1}^n \sigma^i m(X^i)\bigg| \le R\rho_n.$$

Hence, splitting the expectation that defines the local Rademacher complexity over $\mathcal{E}$ and $\mathcal{E}^c$,

$$
\begin{aligned}
\mathcal{R}(\mathcal{M}(\rho_n) \cap B_2(t)) &= \mathbb{E}_{X,\sigma}\left[\sup_{m \in \mathcal{M}(\rho_n)\cap B_2(t)} \frac{1}{n}\left|\sum_{i=1}^n \sigma^i m(X^i)\right|\right] \\
&\leq c\left(\frac{\rho_n^{w/2} d^{(2-w)/4} t^{1-w/2}}{\sqrt{n}} + \frac{\sqrt{d} t \sqrt{\log n}}{\sqrt{n}} + \rho_n r_n^2\right) + \frac{R\rho_n d}{n^2} \\
&\leq c\left(\frac{\rho_n^{w/2} d^{(2-w)/4} t^{1-w/2}}{\sqrt{n}} + \frac{\sqrt{d} t \sqrt{\log n}}{\sqrt{n}} + \rho_n r_n^2\right),
\end{aligned} \tag{64}
$$

where in the second inequality we used the fact that $R\rho_n d/n^2 \leq c\rho_n r_n^2$ (as $d/n \leq 1$ and $r_n^2 \leq n$) for n large enough.

Next, we study the local Rademacher complexity of $\mathcal{P}_k$. This is an easier calculation, following say Example 14.1 of Wainwright (2017). Let $v_1, \dots, v_{kd+1}$ be an orthonormal basis for the additive polynomials $\mathcal{P}_k$ with respect to the L_2 norm. Expanding any $p \in \mathcal{P}_k$ as $p = \sum_{\ell=1}^{kd+1} \alpha_\ell v_\ell$, note that $\|p\|_2 = 1 \iff \|\alpha\|_2 = 1$. Thus, denoting by $V \in \mathbb{R}^{n \times (kd+1)}$ the matrix whose columns contain the evaluations of $v_1, \dots, v_{kd+1}$ at the input points,

$$
\begin{aligned}
\mathcal{R}(\mathcal{P}_k \cap B_2(t)) &= \mathbb{E}_{X,\sigma}\left[\sup_{\|\alpha\|_2 \leq t} \frac{1}{n}\sigma^T V\alpha\right] \\
&\leq \frac{t}{n}\mathbb{E}_{X,\sigma}\|V^T\sigma\|_2 \\
&\leq \frac{t}{n}\sqrt{\mathbb{E}_{X,\sigma}\|V^T\sigma\|_2^2} \\
&= \frac{t}{n}\sqrt{\mathbb{E}_X \operatorname{tr}(VV^T)} = \frac{t\sqrt{kd+1}}{n},
\end{aligned} \tag{65}
$$

where we used the Cauchy-Schwartz inequality, and Jensen's inequality.

Finally, putting together (63), (64), (65), we have

$$
\begin{aligned}
\mathcal{R}(\overline{\mathcal{M}}(\rho_n) \cap B_2(t)) &\leq c\left(\frac{\rho_n^{w/2} d^{(2-w)/4} t^{1-w/2}}{\sqrt{n}} + \frac{\sqrt{d} t \sqrt{\log n}}{\sqrt{n}} + \rho_n r_n^2\right) \\
&\leq c\left(\rho_n^{w/(2+w)} d^{(2-w)/(4+2w)} r_n t + \rho_n^{2w/(2+w)} d^{(2-w)/(2+w)} r_n^2 + \sqrt{d} t r_n + \rho_n r_n^2\right),
\end{aligned}
$$

where in the second inequality we used Lemma 6 (with $a = \rho_n^{w/2} d^{(2-w)/4} n^{-1/2}$ and $b = t$), and also the fact that $\sqrt{\log n/n} \leq r_n$ for n sufficiently large. Therefore an upper bound on the critical radius τ_n is given by solving

$$
c\left(\rho_n^{w/(2+w)} d^{(2-w)/(4+2w)} r_n t + \rho_n^{2w/(2+w)} d^{(2-w)/(2+w)} r_n^2 + \sqrt{d} t r_n + \rho_n r_n^2\right) = t^2,
$$

which is a quadratic in t. Noting that $t^2 = at + b$ for $a, b > 0$ implies $t \leq a + \sqrt{b}$, we obtain

$$
\tau_n \leq c\left(\rho_n^{w/(2+w)} d^{(2-w)/(4+2w)} + \sqrt{d} + \sqrt{\rho_n}\right) r_n.
$$

This completes the proof as $\rho_n^{w/(2+w)} d^{(2-w)/(4+2w)} + \sqrt{d} + \sqrt{\rho_n} \leq 3\max\{\sqrt{d}, \sqrt{\rho_n}\}$. $\square$

We derive a uniform sup norm bound on the additive polynomials in $\mathcal{P}_k$ that have L_2 norm at most 1.

Lemma 12. *It holds that*

$$\|p\|_\infty \le (k+1)\sqrt{\frac{d}{b_1}}, \quad \textit{for all } p \in \mathcal{P}_k,\ \|p\|_2 \le 1,$$

where $\mathcal{P}_k$ is as defined in (47).

Proof. Fix any $p \in \mathcal{P}_k$ with $\|p\|_2 \le 1$, and let $1, v_{j1}, \ldots, v_{jk}$ be an orthonormal basis for the space of degree k univariate polynomials on $[0,1]$ with respect to the measure Q_j, for $j = 1, \ldots, d$. Then, an orthonormal basis for $\mathcal{P}_k$ with respect to Q is given by 1 and the functions $v_{j\ell}, j = 1, \ldots, d,\ \ell = 1, \ldots, k$. For any $j = 1, \ldots, d$, it is well-known that we can choose polynomial basis functions such that $\|v_{j\ell}\|_\infty \le \sqrt{(\ell+1)/b_1}$, for $\ell = 1, \ldots, k$ (e.g., see Lemma 0.1 in Aptekarev et al. (2016)). Thus writing

$$p = \alpha_0 + \sum_{j=1}^{d} \sum_{\ell=1}^{k} \alpha_{j\ell} v_{j\ell},$$

and noting $\|\alpha\|_2 = \|p\|_2 = 1$, we have, by Cauchy-Schwartz inequality,

$$\|p\|_\infty \le \|\alpha\|_1 \max_{\substack{\ell=1,\ldots,k \\ j=1,\ldots,d}} \|v_{j\ell}\|_\infty \le \sqrt{kd+1}\sqrt{(k+1)/b_1} \le (k+1)\sqrt{\frac{d}{b_1}}.$$

□

Using Lemma 12, and classical empirical process theory for finite-dimensional function classes, we derive bounds between the L_2 and empirical norm of an additive polynomial function, uniformly over $\mathcal{P}_k$.

Lemma 13. *There exist constants $c_0, n_0 > 0$, that depend only on k, b_1, such that for all $n \ge n_0$,*

$$\frac{1}{2}\|p\|_n^2 \le \|p\|_2^2 \le 2\|p\|_n^2, \quad \textit{for all } p \in \mathcal{P}_k,$$

with probability at least $1 - c_0 d/n$, where $\mathcal{P}_k$ is as defined in (47).

Proof. Lemma 12 shows that there is a constant $R_0 \ge 1$ such that $\sup_{p \in \mathcal{P}_k} \|p\|_\infty / \|p\|_2 \le R_0\sqrt{d}$. As $\mathcal{P}_k$ has dimension $kd+1$, we have

$$\frac{(R_0\sqrt{d})^2(kd+1)}{n} \to 0 \quad \text{as } n \to \infty,$$

by Assumption B1. Then somewhat standard chaining arguments, e.g., Lemma 10 in Huang (1998), give the result as stated the lemma. □

A.6 Proof of Theorem 1

In this proof, $c \ge 1$ is a constant that will multiply the final estimation error rate, and also control the probability with which our final result holds. The value of c may change from line to line (i.e., we will redefine it as needed) to keep notation simple. Some steps will only hold for n sufficiently large, though we do not always make this explicit. Also, without loss of generality, we conduct our analysis on the Q-almost sure event on which $\Pi_{k,n}, \Pi_{k,n}^\perp$ are well-defined.

Denote by $\hat{f} = \sum_{j=1}^d \hat{f}_j$ the total additive fit in (24). By orthogonality,

$$\|\hat{f} - f_0\|_n^2 = \|\Pi_{k,n}(\hat{f} - f_0)\|_n^2 + \|\Pi_{k,n}^\perp(\hat{f} - f_0)\|_n^2. \tag{66}$$

so we can separately analyze the error in the directions given by $\Pi_{k,n}$ and $\Pi^{\perp}_{k,n}$. For the first term in (66), note by orthogonality, and the fact that J is zero over the space $\mathcal{P}_k$ of additive polynomials in (47), we can view $\Pi_{k,n}(\hat{f})(X^i)$, $i = 1, \dots, n$, as the fitted values from the least squares regression of $Z^i = Y^i - \bar{Y} - \Pi^{\perp}_{k,n}(f_0)(X^i) = \Pi_{k,n}(f_0)(X^i) + \epsilon^i - \bar{\epsilon}$, $i = 1, \dots, n$ onto the linear subspace

$$L_k = \left\{ (p(X^1), \dots, p(X^n)) \in \mathbb{R}^n : p \in \mathcal{P}_k \right\}, \tag{67}$$

of dimension $kd + 1$. Hence

$$\|\Pi_{k,n}(\hat{f} - f_0)\|_n^2 = \frac{1}{n} \|VV^T(\epsilon - \bar{\epsilon}\mathbb{1})\|_2^2 = \frac{1}{n} \|V_{-1}^T \epsilon\|_2^2, \tag{68}$$

where $V \in \mathbb{R}^{n \times (kd+1)}$ has orthonormal columns that span L_k, having first column $\mathbb{1}/\sqrt{n}$, and V_{-1} denotes V with its first column removed. By a somewhat standard sub-Gaussian tail bound (e.g., Theorem 1 in Hsu et al. (2012)), we have, for all $c \geq 2$,

$$\|\Pi_{k,n}(\hat{f} - f_0)\|_n^2 \leq c\sigma^2 \frac{kd}{n}, \tag{69}$$

on an event Ω_1 with probability at least $1 - \exp(-c/16)$.

Now, we focus on the error in the direction of $\Pi^{\perp}_{k,n}$, i.e., the second term in (66). Our analysis will also consider the error in the L_2 norm, after projecting out polynomials in the appropriate L_2 sense. For the rest of the proof, we break down the exposition into mini sections for readability.

Deriving a basic inequality. By orthogonality of $\Pi_{k,n}, \Pi^{\perp}_{k,n}$, and the fact that J is zero over the space $\mathcal{P}_k$ in (47), we see that $\Pi^{\perp}_{k,n}(\hat{f})$ must be optimal for

$$\underset{f \in \mathcal{S}}{\operatorname{argmin}} \; \frac{1}{2} \|W - f\|_n^2 + \lambda_n J_d(f),$$

where $W^i = Y^i - \bar{Y} - \Pi_{k,n}(\hat{f}) = \Pi^{\perp}_{k,n}(f_0)(X^i) + \Pi^{\perp}_{k,n}(\epsilon)(X^i)$, $i = 1, \dots, n$, we denote by $\mathcal{S}$ the feasible set in (24), and we abbreviate $\lambda_n = \lambda/n$ and $J_d(f) = \sum_{j=1}^d J(f_j)$. A note on notation: here and throughout, we will abuse our notation for the empirical norms, inner products, and projectors by applying them over vector arguments, to be interpreted in the appropriate sense.

Standard arguments (from first-order optimality) imply

$$\langle W - \Pi^{\perp}_{k,n}(\hat{f}), \Pi^{\perp}_{k,n}(\tilde{f}) - \Pi^{\perp}_{k,n}(\hat{f}) \rangle_n \leq \lambda_n (J_d(\tilde{f}) - \lambda_n J_d(\hat{f})),$$

for any feasible $\tilde{f} \in \mathcal{S}$. Expanding the definition of W on both sides and rearranging gives

$$\langle \Pi^{\perp}_{k,n}(\hat{f}) - \Pi^{\perp}_{k,n}(f_0), \Pi^{\perp}_{k,n}(\hat{f}) - \Pi^{\perp}_{k,n}(\tilde{f}) \rangle_n \leq \langle \epsilon, \Pi^{\perp}_{k,n}(\hat{f}) - \Pi^{\perp}_{k,n}(\tilde{f}) \rangle_n + \lambda_n (J_d(\tilde{f}) - \lambda_n J_d(\hat{f})).$$

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 $\| \cdot \|$,

$$\begin{aligned} \|\Pi^{\perp}_{k,n}(\hat{f}) - \Pi^{\perp}_{k,n}(f_0)\|_n^2 + \|\Pi^{\perp}_{k,n}(\hat{f}) - \Pi^{\perp}_{k,n}(\tilde{f})\|_n^2 \leq \langle 2\epsilon, \Pi^{\perp}_{k,n}(\hat{f}) - \Pi^{\perp}_{k,n}(\tilde{f}) \rangle_n + \\ 2\lambda_n (J_d(\tilde{f}) - \lambda_n J_d(\hat{f})) + \|\Pi^{\perp}_{k,n}(\tilde{f}) - \Pi^{\perp}_{k,n}(f_0)\|_n^2. \end{aligned}$$

Abbreviating $\hat{\Delta} = \hat{f} - \tilde{f}$, $\hat{J} = J_d(\hat{f})$, and $\tilde{J} = J_d(\tilde{f})$, this becomes

$$\|\Pi^{\perp}_{k,n}(\hat{f} - f_0)\|_n^2 + \|\Pi^{\perp}_{k,n}(\hat{\Delta})\|_n^2 \leq 2\langle \epsilon, \Pi^{\perp}_{k,n}(\hat{\Delta}) \rangle_n + 2\lambda_n (\tilde{J} - \hat{J}) + \|\Pi^{\perp}_{k,n}(\tilde{f} - f_0)\|_n^2, \tag{70}$$

which is our basic inequality.

Bounding the sub-Gaussian complexity. We focus on bounding the first term in (70), i.e., the sub-Gaussian complexity. Recall the relationship between the empirical and L_2 projection operators

$$\Pi_{k,n}^{\perp}(\hat{\Delta}) = \Pi_k^{\perp}(\hat{\Delta}) + (\Pi_k - \Pi_{k,n})(\hat{\Delta}),$$

as noted in the proof of Lemma 4. Hence we have

$$\langle \epsilon, \Pi_{k,n}^{\perp}(\hat{\Delta}) \rangle_n = \underbrace{\langle \epsilon, \Pi_k^{\perp}(\hat{\Delta})}_{a} + \underbrace{\langle \epsilon, (\Pi_k - \Pi_{k,n})(\hat{\Delta}) \rangle_n}_{b}. \tag{71}$$

We examine the two terms a, b in (71) separately. For the first term, by the decomposability of $\Pi_k^{\perp}$ in (48), we can express

$$\begin{aligned} a &= \sum_{j=1}^{d} \langle \epsilon, \widetilde{\Pi}_{k,j}^{\perp}(\hat{\Delta}_j) \rangle_{n,j} \\ &= 2 \sum_{j=1}^{d} (\hat{J}_j + \tilde{J}_j) \left\langle \epsilon, \frac{\widetilde{\Pi}_{k,j}^{\perp}(\hat{\Delta}_j)}{\hat{J}_j + \tilde{J}_j} \right\rangle_{n,j} \\ &\leq \sum_{j=1}^{d} (\hat{J}_j + \tilde{J}_j)^{w/2} \|\widetilde{\Pi}_{k,j}^{\perp}(\hat{\Delta}_j)\|_{n,j}^{1-w/2} \left(\sup_{m_j \in \mathcal{M}_j(1)} \frac{\langle \epsilon, m_j \rangle_{n,j}}{\|m_j\|_{n,j}^{1-w/2}} \right), \end{aligned}$$

where we let $\hat{J}_j = \widetilde{\Pi}_{k,j}^{\perp}(\hat{f}_j)$ and $\tilde{J}_j = \widetilde{\Pi}_{k,j}^{\perp}(\tilde{f}_j)$, $j = 1, \ldots, d$, and we recall that $\mathcal{M}_j(1)$, $j = 1, \ldots, d$ are as defined in (50). For simultaneous control over dimensions $j = 1, \ldots, d$, we use

$$\max_{j=1,\ldots,d} \sup_{m_j \in \mathcal{M}_j(1)} \frac{\frac{1}{n}\sum_{i=1}^{n} \epsilon^i m_j(X_j^i)}{\|m_j\|_{n,j}^{1-w/2}} \leq \sup_{Z_n = \{z^1, \ldots, z^n\} \subseteq [0,1]} \sup_{g \in B_J(1) \cap B_\infty(R)} \frac{\frac{1}{n}\sum_{i=1}^{n} \epsilon^i g(z^i)}{\|g\|_{Z_n}^{1-w/2}},$$

where we have used the upper bound $R \geq 1$ from Lemma 8. Now consider

$$\log N\big(\delta, \|\cdot\|_{Z_n}, B_J(1) \cap B_\infty(R)\big) \leq \log N\big(\delta/R, \|\cdot\|_{Z_n}, B_J(1) \cap B_\infty(1)\big) \leq K R^w \delta^{-w}.$$

The first inequality uses $R \geq 1$ and the second invokes Assumption C3; in fact, by this assumption, the same entropy bound holds uniformly over all choices of $Z_n = \{z^1, \ldots, z^n\} \subseteq [0,1]$. Therefore, appealing to Lemma 3.5 in van de Geer (1990) (also Lemma 8.4 in van de Geer (2000)), there exist constants $c_0, n_0 > 0$, depending only on σ, k, R, K, such that for all $c \geq c_0$ and $n \geq n_0$,

$$\sup_{g \in B_J(1) \cap B_\infty(R)} \frac{\frac{1}{n}\sum_{i=1}^{n} \epsilon^i g(z^i)}{\|g\|_{Z_n}^{1-w/2}} \leq \frac{c}{\sqrt{n}}$$

on an event Ω_2 with probability at least $1 - \exp(-c^2/c_0^2)$. (We note that Lemma 3.5 of van de Geer (1990) studies a local type of entropy, where the function class of interest is intersected with a ball of constant empirical norm; however, it applies directly to our case since the class $B_J(1) \cap B_\infty(R)$ is clearly bounded in the empirical norm, by R). This implies that, for $c \geq c_0$,

$$a \leq \frac{c}{\sqrt{n}} \sum_{j=1}^{d} (\hat{J}_j + \tilde{J}_j)^{w/2} \|\widetilde{\Pi}_{k,j}^{\perp}(\hat{\Delta}_j)\|_{n,j}^{1-w/2}, \tag{72}$$

on Ω_2. For the second term in (71), observe simply that

$$\begin{aligned} b &= \langle VV^T \epsilon, (\Pi_k - \Pi_{k,n})(\hat{\Delta}) \rangle_n \\ &\leq \frac{1}{\sqrt{n}} \|V^T \epsilon\|_2 \|(\Pi_k - \Pi_{k,n})(\hat{\Delta})\|_n \\ &= \frac{1}{\sqrt{n}} \|V^T \epsilon\|_2 \|\Pi_k(\Pi_{k,n}^{\perp}(\hat{\Delta}))\|_n, \end{aligned}$$

where we recall $V \in \mathbb{R}^{n\times(kd+1)}$ is the matrix whose orthonormal columns span L_k in (67). The first line above uses the fact that the restriction of $(\Pi_k - \Pi_{k,n})(\hat{\Delta})$ to the input points already lies in L_k, the second uses the Cauchy-Schwartz inequality and the fact that V has orthogonal columns, and the third uses property (vi) from Lemma 4. From the same sub-Gaussian tail bound as in (68), we have $\|V^T\epsilon\|_2 \le c\sigma\sqrt{kd}$ on Ω_1, and hence

$$b \le c\sigma\sqrt{\frac{kd}{n}}\|\Pi_k(\Pi^\perp_{k,n}(\hat{\Delta}))\|_n, \tag{73}$$

on Ω_1. Combining (71), (72), (73), we have the sub-Gaussian complexity bound

$$\langle \epsilon, \Pi^\perp_{k,n}(\hat{\Delta})\rangle_n \le \frac{c}{\sqrt{n}}\sum_{j=1}^d (\hat{J}_j+\tilde{J}_j)^{w/2}\|\widetilde{\Pi}^\perp_{k,j}(\hat{\Delta}_j)\|_{n,j}^{1-w/2} + c\sigma\sqrt{\frac{kd}{n}}\|\Pi_k(\Pi^\perp_{k,n}(\hat{\Delta}))\|_n,$$

on $\Omega_1\cap\Omega_2$. We can plug this into the basic inequality (70), yielding

$$\begin{aligned}\|\Pi^\perp_{k,n}(\hat{f}-f_0)\|_n^2 + \|\Pi^\perp_{k,n}(\hat{\Delta})\|_n^2 \le \frac{c}{\sqrt{n}}\sum_{j=1}^d (\hat{J}_j+\tilde{J}_j)^{w/2}\|\widetilde{\Pi}^\perp_{k,j}(\hat{\Delta}_j)\|_{n,j}^{1-w/2} + \\ c\sqrt{\frac{d}{n}}\|\Pi_k(\Pi^\perp_{k,n}(\hat{\Delta}))\|_n + 2\lambda_n(\tilde{J}-\hat{J}) + \|\Pi^\perp_{k,n}(\tilde{f}-f_0)\|_n^2,\end{aligned} \tag{74}$$

on $\Omega_1\cap\Omega_2$.

Bringing in the L_2 norm. We bound the empirical norm error terms $\|\widetilde{\Pi}^\perp_{k,j}(\hat{\Delta})\|_{n,j}$, $j=1,\dots,d$ in (74) with their L_2 norm analogs. By Lemma 10, there exist constants $c_1,c_2,c_3,n_1>0$, such that for $n\ge n_1$,

$$\|\widetilde{\Pi}^\perp_{k,j}(\hat{\Delta}_j)\|_{n,j} \le \sqrt{\frac{3}{2}}\|\widetilde{\Pi}^\perp_{k,j}(\hat{\Delta}_j)\|_{2,j} + c_1(\hat{J}_j+\tilde{J}_j)r_n, \quad j=1,\dots,d,$$

on an event Ω_3 with probability at least $1-c_2 d\exp(-c_3 n r_n^2)$, where we abbreviate $r_n = n^{-1/(2+w)}$. Plugging this into (74), and using the simple inequality from Lemma 5, gives

$$\begin{aligned}\|\Pi^\perp_{k,n}(\hat{f}-f_0)\|_n^2 + \|\Pi^\perp_{k,n}(\hat{\Delta})\|_n^2 \le \frac{c}{\sqrt{n}}\sum_{j=1}^d (\hat{J}_j+\tilde{J}_j)^{w/2}\|\widetilde{\Pi}^\perp_{k,j}(\hat{\Delta}_j)\|_{2,j}^{1-w/2} + c(\hat{J}+\tilde{J})r_n^2 + \\ c\sqrt{\frac{d}{n}}\|\Pi_k(\Pi^\perp_{k,n}(\hat{\Delta}))\|_n + 2\lambda_n(\tilde{J}-\hat{J}) + \|\Pi^\perp_{k,n}(\tilde{f}-f_0)\|_n^2,\end{aligned}$$

on $\Omega_1\cap\Omega_2\cap\Omega_3$, where we have used $\sum_{j=1}^d(\hat{J}_j+\tilde{J}_j) = \hat{J}+\tilde{J}$, as well as $n^{-1/2}r_n^{1-w/2} = r_n^2$. Using Holder's inequality $a^Tb \le \|a\|_p\|b\|_q$ on the first term on the right-hand side above, with $p=4/(2+w)$ and $q=4/(2-w)$, we have

$$\begin{aligned}\|\Pi^\perp_{k,n}(\hat{f}-f_0)\|_n^2 + \|\Pi^\perp_{k,n}(\hat{\Delta})\|_n^2 \le \frac{c}{\sqrt{n}}\bigg(\sum_{j=1}^d (\hat{J}_j+\tilde{J}_j)^{2w/(2+w)}\bigg)^{(2+w)/4}\|\Pi^\perp_k(\hat{\Delta})\|_2^{1-w/2} + \\ c(\hat{J}+\tilde{J})r_n^2 + c\sqrt{\frac{d}{n}}\|\Pi_k(\Pi^\perp_{k,n}(\hat{\Delta}))\|_n + 2\lambda_n(\tilde{J}-\hat{J}) + \|\Pi^\perp_{k,n}(\tilde{f}-f_0)\|_n^2,\end{aligned}$$

on $\Omega_1\cap\Omega_2\cap\Omega_3$, where we have used the key decomposability property in (49). Applying Holder's inequality once more on the first term on the right-hand side above, now with $p=(2+w)/(2w)$ and

$q = (2+w)/(2-w)$, we get

$$\|\Pi_{k,n}^\perp(\hat f - f_0)\|_n^2 + \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^2 \le \frac{c}{\sqrt{n}}(\hat J + \tilde J)^{w/2} d^{(2-w)/4} \|\Pi_k^\perp(\hat\Delta)\|_2^{1-w/2} + c(\hat J + \tilde J) r_n^2 + c\sqrt{\frac{d}{n}}\|\Pi_k(\Pi_{k,n}^\perp(\hat\Delta))\|_n + 2\lambda_n(\tilde J - \hat J) + \|\Pi_{k,n}^\perp(\tilde f - f_0)\|_n^2, \quad (75)$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3$.

Next, we bound the term $\|(\Pi_k(\Pi_{k,n}^\perp(\hat\Delta))\|_n$ in (75) by its L_2 norm analog. By Lemma 13, there are constants $c_4, n_2 > 0$, such that for $n \ge n_2$,

$$\|\Pi_k(\Pi_{k,n}^\perp(\hat\Delta))\|_n \le \sqrt{2}\|\Pi_k(\Pi_{k,n}^\perp(\hat\Delta))\|_2$$

on an event Ω_4 with probability at least $1 - c_4 d/n$. Then by nonexpansiveness of the L_2 projection operator with respect to the $\|\cdot\|_2$ norm,

$$\|\Pi_k(\Pi_{k,n}^\perp(\hat\Delta))\|_n \le \sqrt{2}\|\Pi_{k,n}^\perp(\hat\Delta)\|_2,$$

on Ω_4. Thus from (75),

$$\|\Pi_{k,n}^\perp(\hat f - f_0)\|_n^2 + \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^2 \le \frac{c}{\sqrt{n}}(\hat J + \tilde J)^{w/2} d^{(2-w)/4} \|\Pi_k^\perp(\hat\Delta)\|_2^{1-w/2} + c(\hat J + \tilde J) r_n^2 + c\sqrt{\frac{d}{n}}\|\Pi_{k,n}^\perp(\hat\Delta)\|_2 + 2\lambda_n(\tilde J - \hat J) + \|\Pi_{k,n}^\perp(\tilde f - f_0)\|_n^2, \quad (76)$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$.

Back to the empirical norm. Property (ix) of Lemma 4 implies that $\|\Pi_k^\perp(\hat\Delta)\|_2 \le \|\Pi_{k,n}^\perp(\hat\Delta)\|_2$, so from (76), we have

$$\|\Pi_{k,n}^\perp(\hat f - f_0)\|_n^2 + \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^2 \le \frac{c}{\sqrt{n}}(\hat J + \tilde J)^{w/2} d^{(2-w)/4} \|\Pi_{k,n}^\perp(\hat\Delta)\|_2^{1-w/2} + c(\hat J + \tilde J) r_n^2 + c\sqrt{\frac{d}{n}}\|\Pi_{k,n}^\perp(\hat\Delta)\|_2 + 2\lambda_n(\tilde J - \hat J) + \|\Pi_{k,n}^\perp(\tilde f - f_0)\|_n^2, \quad (77)$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. We now bound the L_2 norm error term $\|\Pi_{k,n}^\perp(\hat\Delta)\|_2$ appearing above with its empirical norm analog. By Lemma 11, there exist constants $c_5, c_6, n_3 > 0$, such that for $n \ge n_3$,

$$\|\Pi_{k,n}^\perp(\hat\Delta)\|_2 \le \sqrt{2}\|\Pi_{k,n}^\perp(\hat\Delta)\|_n + c_5(2b_0)(d \vee (\hat J + \tilde J))^{1/2} r_n,$$

on an event Ω_5 with probability at least $1 - \exp(-c_6 d n r_n^2)$. Here we have used the sup norm bound $\|\Pi_{k,n}^\perp(\hat\Delta)\|_\infty \le \|\Pi_{k,n}^\perp(\hat f)\|_\infty + \|\Pi_{k,n}^\perp(\tilde f)\|_\infty \le 2b_0$, by feasibility of $\hat f, \tilde f$, and we let $a \vee b = \max\{a, b\}$. Plugging the bound in the last display into (77), and using Lemma 5 again for the first appearance of $\|\Pi_{k,n}^\perp(\hat\Delta)\|_2$, we have

$$\|\Pi_{k,n}^\perp(\hat f - f_0)\|_n^2 + \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^2 \le \frac{c}{\sqrt{n}}(\hat J + \tilde J)^{(2+w)/4} \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^{1-w/2} + c(d \vee (\hat J + \tilde J)) r_n^2 + c\sqrt{\frac{d}{n}}\|\Pi_{k,n}^\perp(\hat\Delta)\|_n + 2\lambda_n(\tilde J - \hat J) + \|\Pi_{k,n}^\perp(\tilde f - f_0)\|_n^2,$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4 \cap \Omega_5$, where we used $(\hat J + \tilde J)^{w/2} d^{(2-w)/4} (d \vee (\hat J + \tilde J))^{(2-w)/4} \le d \vee (\hat J + \tilde J)$. By the simple inequality in Lemma 6 applied to the first term on the right-hand side above (with $a = (\hat J + \tilde J)^{(2+w)/4} n^{-1/2}$ and $b = \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^{1-w/2}$),

$$\|\Pi_{k,n}^\perp(\hat f - f_0)\|_n^2 + \|\Pi_{k,n}^\perp(\hat\Delta)\|_n^2 \le c(\hat J + \tilde J)^{1/2} r_n \|\Pi_{k,n}^\perp(\hat\Delta)\|_n + c(d \vee (\hat J + \tilde J)) r_n^2 + 2\lambda_n(\tilde J - \hat J) + \|\Pi_{k,n}^\perp(\tilde f - f_0)\|_n^2, \quad (78)$$

on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4 \cap \Omega_5$.

Deriving a final error bound. Consider the first term on the right-hand side in (78). Using the simple inequality $2ab \le a^2 + b^2$, and canceling out the common term of $\|\Pi^\perp_{k,n}(\hat\Delta)\|_n^2$ on both sides, we have

$$\|\Pi^\perp_{k,n}(\hat f - f_0)\|_n^2 \le c(d \vee (\hat J + \tilde J)) r_n^2 + 2\lambda_n(\tilde J - \hat J) + \|\Pi^\perp_{k,n}(\tilde f - f_0)\|_n^2,$$

on $\Omega = \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4 \cap \Omega_5$. Choosing $\lambda_n \ge 2cr_n^2$, we have

$$\|\Pi^\perp_{k,n}(\hat f - f_0)\|_n^2 \le \frac{5}{2}\lambda_n(d \vee \tilde J) + \|\tilde f - f_0\|_n^2, \tag{79}$$

on Ω, where we used the nonexpansiveness of the empirical projection operator with respect to the $\|\cdot\|_n$ norm to simplify the upper bound. Adding (79) to the polynomial part bound in (69), we have derived the empirical norm error bound in (26), on Ω. Further, we can bound $\hat J$ on Ω as follows. If $\hat J \ge 2\tilde J$, then from the second to last display, we have

$$0 \le c(d \vee (\hat J + \tilde J)) r_n^2 - \lambda_n \hat J + \|\tilde f - f_0\|_n^2,$$

on Ω, and after rearranging and using $\lambda_n \ge 2cr_n^2$, this gives $\hat J \le d \vee \tilde J + \|\tilde f - f_0\|_n^2/(cr_n)^2$ on Ω. As this was derived in the case $\hat J \ge 2\tilde J$, we conclude that $\hat J \le 3(d \vee \tilde J) + \|\tilde f - f_0\|_n^2/(cr_n)^2$ on Ω.

Finally, for the L_2 norm the error bound in (27), we can study the error in the directions given by Π_k and $\Pi_k^\perp$ separately, by orthogonality. For the part in the direction of $\Pi_k^\perp$, from what we have already shown, it holds on Ω that

$$\begin{aligned}
\|\Pi_k^\perp(\hat\Delta)\|_2 &\le \|\Pi^\perp_{k,n}(\hat\Delta)\|_2 \\
&\le \sqrt{2}\|\Pi^\perp_{k,n}(\hat\Delta)\|_n + c_5(2b_0)(d \vee (\hat J + \tilde J))^{1/2} r_n \\
&\le \sqrt{2}\|\Pi^\perp_{k,n}(\hat\Delta)\|_n + c(d \vee \tilde J)^{1/2} r_n + \|\tilde f - f_0\|_n \\
&\le \sqrt{2}\|\Pi^\perp_{k,n}(\hat f - f_0)\|_n + c(d \vee \tilde J)^{1/2} r_n + (1+\sqrt{2})\|\tilde f - f_0\|_n \\
&\le (1+\sqrt{5})(\lambda_n(d \vee \tilde J))^{1/2} + (1+2\sqrt{2})\|\tilde f - f_0\|_n,
\end{aligned} \tag{80}$$

In the above, the first line holds by property (ix) of Lemma 4; the second holds by definition of Ω_5; the third holds by our upper bound on $\hat J$ (and by adjusting c to be larger, if needed); the fourth holds by the triangle inequality; and the fifth holds by our bound in (79) and our choice of λ_n. By the triangle inequality, and the simple inequality $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$,

$$\|\Pi_k^\perp(\hat f - f_0)\|_2^2 \le 32\lambda_n(d \vee \tilde J) + 44\|\tilde f - f_0\|_n^2 + 3\|\Pi_k^\perp(\tilde f - f_0)\|_2^2, \tag{81}$$

on Ω. For the part in the direction of Π_k, note that on Ω,

$$\begin{aligned}
\|\Pi_k(\hat\Delta)\|_2 &\le \|\Pi_{k,n}(\hat\Delta)\|_2 + \|\Pi^\perp_{k,n}(\hat\Delta)\|_2 \\
&\le \sqrt{2}\|\Pi_{k,n}(\hat\Delta)\|_n + \|\Pi^\perp_{k,n}(\hat\Delta)\|_2 \\
&\le \sqrt{2}c\sigma\sqrt{\frac{kd}{n}} + \sqrt{2}\|\Pi_{k,n}(\tilde f - f_0)\|_n + \|\Pi^\perp_{k,n}(\hat\Delta)\|_2 \\
&\le (2+\sqrt{5})(\lambda_n(d \vee \tilde J))^{1/2} + (1+3\sqrt{2})\|\tilde f - f_0\|_n.
\end{aligned}$$

Here, the first line holds by property (xi) of Lemma 4 and the triangle inequality; the second holds by construction of the event Ω_4; the third holds from the triangle inequality and our polynomial part bound in (69); and the fourth holds by what we have just shown about in (80). Applying the triangle inequality once again, along with $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$, we get

$$\|\Pi_k(\hat f - f_0)\|_2^2 \le 54\lambda_n(d \vee \tilde J) + 83\|\tilde f - f_0\|_n^2 + 3\|\Pi_k(\tilde f - f_0)\|_2^2, \tag{82}$$

on Ω. Adding (82) to (81) gives the L_2 norm error bound in (27), on Ω. This completes the proof.

A.7 Preliminaries for the proof of Corollary 1

The following two lemmas will be helpful for the proof of Corollary 1.

Lemma 14. *Given $f = \sum_{j=1}^d f_j$, whose component functions are each k times weakly differentiable, there exists an additive spline approximant $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, where $\tilde{f}_j \in \mathcal{G}_j$, the set of kth order splines with knots in the set T_j defined in (23), for $j = 1, \dots, d$, such that*

(i) $\sum_{j=1}^d \mathrm{TV}(\tilde{f}_j^{(k)}) \le a_k \sum_{j=1}^d \mathrm{TV}(f_j^{(k)})$; *and*

(ii) $\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_j \|_\infty \le a_k W_{\max}^k \sum_{j=1}^d \mathrm{TV}(f_j^{(k)})$.

Above, $a_k \ge 1$ is a constant depending only on k, and we define $W_{\max} = \max_{j=1,\dots,d} W_j$, where

$$W_j = \max_{i=1,\dots,n-1} |X_j^{(i)} - X_j^{(i+1)}|, \quad j = 1, \dots, d.$$

When the input points are drawn from a distribution Q satisfying Assumption A1, there are universal constants $c_0, n_0 > 0$ such that all for $n \ge n_0$, it holds that $W_{\max} \le (c_0/b_1) \log n/n$ with probability at least $1 - 2b_1 d n^{-10}$, and so the bound in (ii) becomes

$$\left\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_j \right\|_\infty \le \frac{c_0 a_k}{b_1} \left(\frac{\log n}{n} \right)^k \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}), \tag{83}$$

with probability at least $1 - 2b_1 d n^{-10}$.

Proof. From Proposition 7 in Mammen & van de Geer (1997), we know that for each $j = 1, \dots, d$, there is a kth degree spline function $\tilde{f}_j$ whose knots lie in T_j in (23), with $\mathrm{TV}(\tilde{f}_j^{(k)}) \le a_k \mathrm{TV}(f_j^{(k)})$ and

$$\|\tilde{f}_j - f_j\|_\infty \le a_k W_j^k \mathrm{TV}(f_j^{(k)}),$$

where $a_k \ge 1$ is a constant depending only on k, and W_j defined as in the lemma. This is derived using strong quasi-interpolating properties of spline functions, from de Boor (1978). Thus we have, by the triangle inequality,

$$\left\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_j \right\|_\infty \le \sum_{j=1}^d \|\tilde{f}_j - f_j\|_\infty \le a_k W_{\max}^k \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}).$$

This verifies parts (i) and (ii) of the lemma.

When we consider random input points drawn from a distribution Q satisfying Assumption A1, the boundedness condition on the marginal densities Q_j implies that there are universal constants $c_0, n_0 > 0$ such that for $n \ge n_0$, we have $W_j \le (c_0/b_1) \log n/n$ with probability at least $1 - 2b_1 n^{-10}$ (see, e.g., Lemma 5 in Wang et al. (2014)), for each $j = 1, \dots, d$. Applying a union bound gives the result for $W_{\max}$. □

Lemma 15. *Given $f = \sum_{j=1}^d f_j$, whose component functions are each k times weakly differentiable, there is an additive falling factorial approximant $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, where $f_j \in \mathcal{H}_j$, the set of kth order falling factorial functions defined over $X_j^1, \dots, X_j^n$, for each $j = 1, \dots, d$, such that*

(i) $\sum_{j=1}^d \mathrm{TV}(\tilde{f}_j^{(k)}) \le a_k \sum_{j=1}^d \mathrm{TV}(f_j^{(k)})$; *and*

(ii) $\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_j \|_\infty \le a_k (W_{\max}^k + 2k^2 W_{\max}) \sum_{j=1}^d \mathrm{TV}(f_j^{(k)})$.

Again, $a_k \geq 1$ is a constant depending only on k, and $W_{\max}$ is as defined in Lemma 14. When the inputs are drawn from a distribution Q satisfying Assumption A1, the bound in (ii) becomes

$$\Big\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_j \Big\|_\infty \leq \frac{c_0 a_k}{b_1} \left(\left(\frac{\log n}{n} \right)^k + 2k^2 \frac{\log n}{n} \right) \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}), \tag{84}$$

with probability at least $1 - 2b_1 d n^{-10}$.

Proof. First we apply Lemma 14 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 (23), 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 $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, our falling factorial approximant, to have component functions

$$\tilde{f}_j = \sum_{\ell=1}^{k+1} \alpha_j^\ell g_{j\ell} + \sum_{\ell=k+2}^n \alpha_j^\ell h_{j\ell}, \quad j = 1, \ldots, d.$$

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 in each $\tilde{f}_j$, we have preserved the polynomial part of f_j^* exactly, for $j = 1, \ldots, d$.) Again, it is straightforward to check that $\mathrm{TV}(\tilde{f}_j^{(k)}) = \sum_{\ell=k+2}^n |\alpha_j^\ell|$, for $j = 1, \ldots, d$, and so

$$\sum_{j=1}^d \mathrm{TV}(\tilde{f}_j^{(k)}) = \sum_{j=1}^d \mathrm{TV}\big((f_j^*)^{(k)}\big) \leq a_k \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}),$$

the inequality coming from part (i) of of Lemma 14. This verifies part (i) of the current lemma. As for part (ii), we note that Lemma 4 of Wang et al. (2014) shows that

$$|h_{j\ell}(X_j^i) - g_{j\ell}(X_j^i)| \leq k^2 W_j, \quad \text{for all } \ell = k+2, \ldots, n,\ i = 1, \ldots, n,\ j = 1, \ldots, d,$$

where recall W_j is the maximum gap between sorted input points along the jth dimension, $j = 1, \ldots, d$, as defined in Lemma 14. In fact, a straightforward modification of their proof can be used to strengthen this result to

$$\|h_{j\ell} - g_{j\ell}\|_\infty \leq 2k^2 W_j, \quad \text{for all } \ell = k+2, \ldots, n,\ j = 1, \ldots, d,$$

which means that by Holder's inequality,

$$\|\tilde{f}_j - f_j^*\|_\infty \leq 2k^2 W_j \sum_{\ell=k+2}^n |\alpha_j^\ell| \leq 2k^2 a_k W_j \mathrm{TV}(f_j^{(k)}) \quad \text{for all } j = 1, \ldots, d.$$

Then, by two applications of the triangle inequality,

$$\Big\| \sum_{j=1}^d \tilde{f}_j - \sum_{j=1}^d f_i \Big\|_\infty \leq \sum_{j=1}^d \Big(\|\tilde{f}_j - f_j^*\|_\infty + \|f_j^* - f_j\|_\infty \Big) \leq a_k \Big(W_{\max}^k + 2k^2 W_{\max} \Big) \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}),$$

where we have used part (ii) of Lemma 14. This verifies part (ii) of the current lemma.

Lastly, for random input points drawn from a distribution Q that satisfies Assumption A1, the proof of (84) follows exactly as in the proof of (83). ∎

A.8 Proof of Corollary 1

We consider first the statement in part (a). We must check that Assumptions C1–C3 hold for our choice of regularizer $J(g) = \mathrm{TV}(g^{(k)})$, and then we can apply Theorem 1. Assumptions C1, C2 are immediate. As for Assumption C3, consider the univariate function class

$$\mathcal{W}_{k+1} = \left\{ f : \int_0^1 |f^{(k+1)}(t)|\, dt \le 1,\ \|f\|_\infty \le 1 \right\}.$$

The results in Birman & Solomyak (1967) imply that for any set $Z_n = \{z^1, \ldots, z^n\} \subseteq [0,1]$,

$$\log N(\delta, \|\cdot\|_{Z_n}, \mathcal{W}_{k+1}) \le K\delta^{-1/(k+1)},$$

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 1, with $\tilde{f} = f_0$, gives the result in (32).

For the statement in part (b), note first that we can consider $k \ge 2$ without a loss of generality, as pointed out in Remark 10 following the corollary. Lemma 14 shows we can choose an additive kth degree spline $\tilde{f}$, feasible for the problem (24) when $\mathcal{S}_j = \mathcal{G}_j$, $j = 1, \ldots, d$, such that (83) holds for $f = f_0$. Applying Theorem 1 and noting the approximation error, i.e., the square of the right-hand side in (83) is of (much) smaller order than $dn^{-(2k+2)/(2k+3)}$, proves the result in (32) for restricted additive locally adaptive splines.

For the statement in part (c), note that again we can consider $k \ge 2$ without a loss of generality. Lemma 15 shows we can choose an additive kth falling factorial function $\tilde{f}$, feasible for (24) with $\mathcal{S}_j = \mathcal{H}_j$, $j = 1, \ldots, d$, such that (84) holds when $f = f_0$. Applying Theorem 1 and noting that the approximation error, i.e., the square of the right-hand side in (84) is of (much) smaller order than $dn^{-(2k+2)/(2k+3)}$, proves the result in (32) for additive trend filtering.

A.9 Preliminaries for the proof of Theorem 2

The first two lemmas in this subsection are helper lemmas for the third, which will be used in the proof of Theorem 2.

Lemma 16. *There exist constants $\tilde{K}_1, \tilde{\delta}_1 > 0$, that depend only on k, K_1, w, R, where $R \ge 1$ is the constant from Lemma 8, such that for all $\delta \le \tilde{\delta}_1$,*

$$\log M\big(\delta, \|\cdot\|_{2,j}, \widetilde{\Pi}_{k,j}^{\perp}(B_J(1))\big) \ge \tilde{K}_1 \delta^{-w}, \quad j = 1, \ldots, d.$$

Proof. Fix an arbitrary $j = 1, \ldots, d$. Assumption A1 implies $\|\cdot\|_{2,j}^2 \ge (b_1/b_2)\|\cdot\|_{2,1}^2$, and thus by Assumption C4,

$$\log M\big(\delta, \|\cdot\|_{2,j}, B_J(1) \cap B_\infty(1)\big) \ge K_1' \delta^{-w}, \tag{85}$$

where $K_1' = (b_1/b_2)^{w/2} K_1$. Denote by $\widetilde{\mathcal{P}}_k$ the space of univariate kth degree polynomials, and recall the sup norm bound $R \ge 1$ given by Lemma 8. Note that we can decompose

$$B_J(1) \cap B_\infty(R) = \widetilde{\Pi}_{k,j}^{\perp}(B_J(1)) + (\widetilde{\mathcal{P}}_k \cap B_\infty(R)), \tag{86}$$

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 (86),

$$\log M\big(\delta, \|\cdot\|_{2,j}, \widetilde{\Pi}^{\perp}_{k,j}(B_J(1))\big) \geq \log \frac{M(4\delta, \|\cdot\|_{2,j}, B_J(1) \cap B_\infty(R))}{N(\delta, \|\cdot\|_{2,j}, \widetilde{\mathcal{P}}_k \cap B_\infty(R))} \geq K_1' 4^{-w} \delta^{-w} - A(k+1)\log(1/\delta),$$

where in the second inequality we have used (85) and a well-known entropy bound for a bounded, finite-dimensional ball (e.g., Mammen (1991)), where $A > 0$ is a constant that depends only on R. For small enough δ, the right-hand side is of the desired order, which completes the proof. □

Lemma 17. *Let d, M be positive integers, and $I = \{1, \dots, M\}$. Write $H(u,v) = \sum_{j=1}^d 1\{u_j \neq v_j\}$ for the Hamming distance between $u, v \in I^d$. Then there exists a subset $U \subseteq I^d$ with $|U| \geq (M/4)^{d/2}$ such that $u, v \in U \Rightarrow H(u,v) \geq d/2$.*

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 $U = \{u_0, \dots, u_E\}$. For $0 \leq i, j \leq E$, by construction, $H(u_i, u_j) > a$. For $j = 0, \dots, E$,

$$\begin{aligned} n_j = |\Omega_j - \Omega_{j+1}| &= |\{u \in \Omega_j : H(u, u_j) \leq a\}| \\ &\leq |\{u \in I^d : H(u, u_j) \leq a\}| \\ &= \binom{d}{d-a} M^a \end{aligned}$$

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

$$E + 1 \geq \frac{M^{d-a}}{\binom{d}{d-a}} \geq \frac{M^{d-a}}{2^d} \geq (M/4)^{d/2}.$$

□

Lemma 18. *There exist 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 16, such that for all $\delta \leq \bar{\delta}_1$,*

$$\log M\big(\delta, \|\cdot\|_2, \Pi^{\perp}_k(B^d_J(1))\big) \geq \bar{K}_1 d^{1+w/2} \delta^{-w}.$$

Proof. Note that, by decomposability of the L^2 projection operator, as in (48),

$$\Pi^{\perp}_k(B^d_J(1)) = \Bigg\{ \sum_{j=1}^d f_j : f_j \in \widetilde{\Pi}^{\perp}_{k,j}(B_J(1)),\ j = 1, \dots, d \Bigg\}.$$

Abbreviate $M = \min_{j=1,\dots,d} M(\delta/\sqrt{d/2}, \|\cdot\|_{2,j}, \widetilde{\Pi}^{\perp}_{k,j}(B_J(1)))$. By Assumption C4, we have

$$\log M \geq 2^{-w/2} K_1 d^{w/2} \delta^{-w}.$$

Let $f_j^1, \dots, f_j^M$ be a $(\delta/\sqrt{d/2})$-packing of $\widetilde{\Pi}^{\perp}_{k,j}(B_J(1))$, for each $j = 1, \dots, d$. Let $I = \{1, \dots, M\}$, and for $u \in I^d$, define $f^u \in \Pi^{\perp}_k(B^d_J(1))$ by

$$f^u = \sum_{j=1}^d f_j^{u_j}.$$

If the Hamming distance between indices u, v satisfies $H(u,v) \geq d/2$, then

$$\|f^u - f^v\|_2^2 = \sum_{j=1}^{d} \|f^{u_j} - f^{v_j}\|_2^2 \geq H(u,v)\frac{\delta^2}{d/2} \geq \delta^2.$$

Thus, it is sufficient to find a subset U of I^d such that $u, v \in U \Rightarrow H(u,v) \geq d/2$. By Lemma 17, we can choose such a U with $|U| \geq (M/4)^{d/2}$. For small enough δ, such that $M \geq 16$, this gives the desired result because

$$\log|U| \geq \frac{d}{2}\log\frac{M}{4} \geq \frac{d}{4}\log M \geq 2^{-w/2-2}K_1 d^{1+w/2}\delta^{-w}.$$

$\square$

A.10 Proof of Theorem 2

Clearly, by orthogonality, for any functions $\hat{f}, f_0$,

$$\|\hat{f} - f_0\|_2^2 = \|\Pi_k(\hat{f}) - \Pi_k(f_0)\|_2^2 + \|\Pi_k^{\perp}(\hat{f}) - \Pi_k^{\perp}(f_0)\|_2^2 \geq \|\Pi_k^{\perp}(\hat{f}) - \Pi_k^{\perp}(f_0)\|_2^2,$$

so it suffices to consider the minimax error over $\Pi_k^{\perp}(B_J^d(C_0))$.

First, we lower bound the packing number and upper bound the covering number of the class $\Pi_k^{\perp}(B_J^d(C_0))$. The upper bound is straightforward: by decomposability of the L^2 norm for additive functions whose components have L^2 mean zero, as in (25), we have

$$\log N\big(\epsilon, \|\cdot\|_2, \Pi_k^{\perp}(B_J^d(C_0))\big) \leq \log\prod_{j=1}^{d} N\big(\epsilon/\sqrt{d}, \|\cdot\|_{2,j}, \Pi_{k,j}^{\perp}(B_J(C_0))\big) \leq K_2 C_0^w d^{1+w/2}\epsilon^{-w}. \quad (87)$$

The lower bound is less straightforward, and is given by Lemma 18:

$$\log M\big(\delta, \|\cdot\|_2, \Pi_k^{\perp}(B_J^d(C_0))\big) \geq \bar{K}_1 C_0^w d^{1+w/2}\delta^{-w}, \quad (88)$$

which holds for all $\delta \leq \bar{\delta}_1$, where $\bar{K}_1, \bar{\delta}_1 > 0$ are constants.

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 $\Pi_k^{\perp}(B_J(C_0))$, for $\delta_n > 0$ to be specified later. Fix an arbitrary estimator $\hat{f}$, and let

$$\hat{Z} = \underset{j\in\{1,\ldots,M\}}{\mathrm{argmin}}\ \|\hat{f} - f^j\|_2.$$

We will use $P_{X,f}$ and $\mathbb{E}_{X,f}$ to denote the probability and expectation operators, respectively, over i.i.d. draws $X^i \sim Q$, $i = 1, \ldots, n$ and i.i.d. draws $Y^i|X^i \sim N(f(X^i), \sigma^2)$, $i = 1, \ldots, n$. Then

$$\begin{aligned}
\sup_{f_0\in\Pi_k^{\perp}(B_J^d(C_0))} \mathbb{E}_{X,f_0}\|\hat{f} - f_0\|_2^2 &\geq \sup_{f_0\in\{f^1,\ldots,f^M\}} \mathbb{E}_{X,f_0}\|\hat{f} - f_0\|_2^2 \\
&\geq \frac{1}{M}\mathbb{E}_X\sum_{j=1}^{M}\mathbb{E}_{f^j}\|\hat{f} - f^j\|_2^2 \\
&= \frac{1}{M}\mathbb{E}_X\sum_{j=1}^{M}\Big(\mathbb{P}_{f^j}(\hat{Z}\neq j)\mathbb{E}_{f^j}\big(\|\hat{f} - f^j\|_2^2 \,\big|\, \hat{Z}\neq j\big) + \mathbb{P}_{f^j}(\hat{Z}= j)\mathbb{E}_{f^j}\big(\|\hat{f} - f^j\|_2^2 \,\big|\, \hat{Z}= j\big)\Big) \\
&\geq \frac{1}{M}\mathbb{E}_X\sum_{j=1}^{M}\mathbb{P}_{f^j}(\hat{Z}\neq j)\mathbb{E}_{f^j}\big(\|\hat{f} - f^j\|_2^2 \,\big|\, \hat{Z}\neq j\big) \\
&\geq \frac{1}{M}\mathbb{E}_X\sum_{j=1}^{M}\mathbb{P}_{f^j}(\hat{Z}\neq j)\frac{\delta_n^2}{4}, \quad (89)
\end{aligned}$$

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,

$$\frac{1}{M}\mathbb{E}_X \sum_{j=1}^M \mathbb{P}_{f^j}(\hat{Z} \neq j) \geq 1 - \frac{\frac{1}{M}\sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{q}) + \log 2}{\log M}, \tag{90}$$

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 $\Pi_k^\perp(B_J^d(C_0))$, 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

$$\begin{aligned}
\frac{1}{M}\sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{q}) &\leq \frac{1}{M}\sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{s}) \\
&= \frac{1}{M}\sum_{j=1}^M \mathbb{E}_{X,f^j} \log \frac{p(N(f^j(X), \sigma^2 I))}{\frac{1}{N}\sum_{\ell=1}^N p(N(g^\ell(X), \sigma^2 I))} \\
&\leq \frac{1}{M}\sum_{j=1}^M \bigg(\log N + \mathbb{E}_X \min_{\ell=1,\ldots,N} \mathrm{KL}(q_j \,\|\, s_\ell) \bigg) \\
&\leq \frac{1}{M}\sum_{j=1}^M \bigg(\log N + \frac{n\epsilon_n^2}{2\sigma^2} \bigg) \\
&\leq K_2 C_0^w d^{1+w/2} \epsilon_n^{-w} + \frac{n\epsilon_n^2}{2\sigma^2}.
\end{aligned} \tag{91}$$

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 simply expressed and manipulated the definition of KL divergence explicitly; 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, we recalled the entropy bound from (87). Minimizing (91) over $\epsilon_n > 0$ gives

$$\frac{1}{M}\sum_{j=1}^M \mathbb{E}_X \mathrm{KL}(q_j \,\|\, \bar{q}) \leq \bar{K}_2 C_0^{2w/(2+w)} d n^{w/(2+w)},$$

for a constant $\bar{K}_2 > 0$. Returning to Fano's inequality (89), (90), we see that a lower bound on the minimax error is

$$\frac{\delta_n^2}{4}\bigg(1 - \frac{\bar{K}_2 C_0^{2w/(2+w)} d n^{w/(2+w)} + \log 2}{\log M}\bigg),$$

Therefore, a lower bound on the minimax error is $\delta_n^2/8$, for any $\delta_n > 0$ such that

$$\log M \geq 2\bar{K}_2 C_0^{2w/(2+w)} d n^{w/(2+w)} + 2\log 2. \tag{92}$$

When n is large enough so that $2\log 2$ is dominated by the first-term on the right-hand side in (92), we see that we can choose $\delta_n = (\bar{K}_1/4\bar{K}_2)^{1/w} C_0^{w/(2+w)} \sqrt{d} n^{-1/(2+w)}$, and our log packing bound (88) ensures that (92) will be satisfied. This completes the proof.

A.11 Proof of Corollary 2

We only need to check Assumption C4 for $J(g) = \mathrm{TV}(g^{(k)})$, $w = 1/(k+1)$, and then we can apply Theorem 2. 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 $[0,1]$,

$$\mathrm{TV}(f^{(k)}) = \int_0^1 |f^{(k+1)}(t)|\,dt \le \sqrt{\int_0^1 |f^{(k+1)}(t)|^2\,dt}.$$

Hence

$$\Big\{f : \mathrm{TV}(f^{(k)}) \le 1,\ \|f\|_\infty \le 1\Big\} \supseteq \Big\{f : \int_0^1 |f^{(k+1)}(t)|^2\,dt \le 1,\ \|f\|_\infty \le 1\Big\}.$$

Results from Kolmogorov & Tikhomirov (1959) show that the space on the right satisfies the desired log packing number lower bound. This proves the result.

A.12 Proof of the linear smoother lower bound in (36)

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 orthogonality of the L_2 norm over additive functions with L_2 mean zero components, as in (25), we have for any additive linear smoother $\hat{f} = \sum_{j=1}^d \hat{f}_j$,

$$\|\hat{f} - f_0\|_2^2 = \bigg(\sum_{j=1}^d \bar{f}_j\bigg)^2 + \sum_{j=1}^d \|(\hat{f}_j - \bar{f}_j) - f_{0j}\|_{2,j}^2 \ge \sum_{j=1}^d \|(\hat{f}_j - \bar{f}_j) - f_{0j}\|_{2.j}^2,$$

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 we can restrict our attention to additive linear smoothers whose component functions have L_2 mean zero, i.e.,

$$\inf_{\hat{f} \text{ additive linear}} \ \sup_{f_0 \in \mathcal{F}_k^d(C_0)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \ = \ \sum_{j=1}^d \inf_{\hat{f}_j \text{ linear}} \ \sup_{f_0 \in \mathcal{F}_k^d(C_0)} \mathbb{E}\|\hat{f}_j(Y) - f_{0j}\|_{2,j}^2. \tag{93}$$

Now consider the jth term in the sum on the right-hand side above, and focus on an arbitrary linear smoother $\hat{f}_j$, fit to data

$$Y^i = \mu + f_{0j}(X_j^i) + \sum_{\ell \ne j} f_{0\ell}(X_\ell^i) + \epsilon^i, \quad i = 1,\ldots,n. \tag{94}$$

which depends on the component functions $f_{0\ell}$, $\ell \ne j$. This is why the supremum in the jth term of the sum in (93) must be taken over $f_0 \in \mathcal{F}_k^d(C_0)$, rather than $f_{0j} \in \mathcal{F}_k(C_0) = \{f : [0,1] \to \mathbb{R} : \mathrm{TV}(f^{(k)}) \le C_0\}$. Our notation $\hat{f}_j(Y)$ is used as a reminder to emphasize the dependence on the full data vector in (94).

A simple reformulation by averaging appropriately over the lattice will help 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 $\bar{Y}_j = (\bar{Y}_j^1, \dots, \bar{Y}_j^N) \in \mathbb{R}^N$. 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 $\tilde{f}_j$ as a linear smoother fit to data $\bar{Y}_j$, whose components follow the distribution

$$\bar{Y}_j^v = \mu_j + f_{0j}(v/N) + \bar{\epsilon}_j^v, \quad v = 1, \dots, N, \tag{95}$$

where we let $\mu_j = \mu + \frac{1}{N}\sum_{\ell \neq j}\sum_{u=1}^n f_{0\ell}(u/N)$, and $\bar{\epsilon}_j^v$, $v = 1, \dots, n$ are i.i.d. $N(0, \sigma^2/N^{d-1})$. Recalling that $f_{0j} \in \mathcal{F}_k(C_0)$, 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 $\mathcal{F}_k(C_0)$ contains the Besov space $B_{1,1}^{k+1}(C_0')$, for a radius C_0' that differs from C_0 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 (95) gives for a constant $c_0 > 0$ and N large enough,

$$\begin{aligned}
\inf_{\hat{f}_j \text{ linear}} \sup_{f_{0j} \in \mathcal{F}_k(C_0)} \mathbb{E}\|\hat{f}_j(\bar{Y}_j) - f_{0j}\|_{2,j}^2 &\geq c_0 (C_0 N^{(d-1)/2})^{2/(2k+2)} \frac{N^{-(2k+1)/(2k+2)}}{N^{d-1}} \\
&= c_0 C_0^{2/(2k+2)} N^{-d(2k+1)/(2k+2)} \\
&= c_0 C_0^{2/(2k+2)} n^{-(2k+1)/(2k+2)}.
\end{aligned} \tag{96}$$

As we have reduced the lower bound to the minimax risk of linear smoothers over a Besov ball, it is not hard to see that the same bound as in (96) indeed holds simultaneously over all $j = 1, \dots, d$. Combining this with (93) gives the desired result in (36).

A.13 Proof of Theorem 3 and derivation details for Algorithm 2

First, we show that the dual problem of (40) is indeed the additive trend filtering problem (8), and moreover that the Lagrange multipliers γ_j corresponding to the constraints $u_0 = u_j$, for $j = 1, \dots, d$, can indeed be seen as the primal parameters θ_j, $j = 1, \dots, d$. Let $M = I - \mathbb{1}\mathbb{1}^T/n$. We first rewrite problem (40) as

$$\begin{aligned}
\min_{u_0, u_1, \dots, u_d \in \mathbb{R}^n} \quad & \frac{1}{2}\|MY - Mu_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) \\
\text{subject to} \quad & Mu_0 = Mu_1, \; Mu_0 = Mu_2, \; \dots, \; Mu_0 = Mu_d,
\end{aligned}$$

We can write the Lagrangian of this problem as

$$L(u_0, u_1, \dots, u_d, \gamma_1, \dots, \gamma_d) = \frac{1}{2}\|MY - Mu_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) + \sum_{j=1}^d \gamma_j^T M(u_0 - u_j).$$

and we want to minimize this over $u_0, \dots, u_d$ to form the dual of (40). This gives

$$\max_{\gamma_1, \dots, \gamma_d \in \mathbb{R}^n} \; \frac{1}{2}\|MY\|_2^2 - \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\gamma_j\Big\|_2^2 - \sum_{j=1}^d \Big(\max_{u_j \in U_j} \; u_j^T M\gamma_j\Big). \tag{97}$$

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),

$$\max_{u_j \in U_j} \; u_j^T M\gamma_j = \max_{\|v_j\|_\leq \lambda} \; v_j^T D_j S_j M\gamma_j = \|D_j S_j M\gamma_j\|_1,$$

where recall we abbreviate $D_j = D^{(X_j, k+1)}$, for $j = 1, \dots, d$, and this allows us to rewrite the above problem (97) as

$$\min_{\gamma_1, \dots, \gamma_d \in \mathbb{R}^n} \; \frac{1}{2}\Big\|MY - \sum_{j=1}^d M\gamma_j\Big\|_2^2 - \sum_{j=1}^d \|D_j S_j M\gamma_j\|_1,$$

which is precisely the same as the original additive trend filtering problem in (8).

Hence, the Lagrange multipliers γ_j, $j = 1, \ldots, d$ in the ADMM iterations (42) are equivalent to primal variables θ_j, $j = 1, \ldots, d$. Note that in the context of the problem (40) to which ADMM is being applied, these multipliers γ_j, $j = 1, \ldots, d$ are dual variables, and under weak conditions, ADMM is known to produce convergent dual iterates. For example, in Section 3.2 of Boyd et al. (2011), the authors show that if (i) the criterion decomposes as a sum of closed and convex functions, and (ii) strong duality holds, then the ADMM algorithm produces dual iterates that converge to optimal dual solutions. (Convergence of primal iterates requires stronger assumptions.) The problem in (40) clearly meets this criteria, and thus the ADMM algorithm outlined in (42) yields $\gamma_j^{(t)}$, $j = 1, \ldots, d$ that converge to optimal solutions in the dual of (40), i.e., optimal solutions in the additive trend filtering problem (8). 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 (39). This follows by notationally swapping θ_j, $j = 1, \ldots, d$ for γ_j, $j = 1, \ldots, d$, rewriting the updates $\theta_j^{(t)} = u_0^{(t)} + \theta_j^{(t-1)} - u_j^{(t)}$, $j = 1, \ldots, d$ as

$$\theta_j^{(t)} = \mathrm{TF}_\lambda\big(u_0^{(t)} + \theta_j^{(t-1)},\ x_j\big), \quad j = 1, \ldots, d,$$

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

$$u_j^{(t-1)} = u_0^{(t-1)} + \theta_j^{(t-2)} - \theta_j^{(t-1)}, \quad j = 1, \ldots, d.$$

B Simulated homogeneously-smooth data

Figure 7 shows the results of a homogeneous simulation, as in Section 6.1 and Figure 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 $j = 1, \ldots, d$.

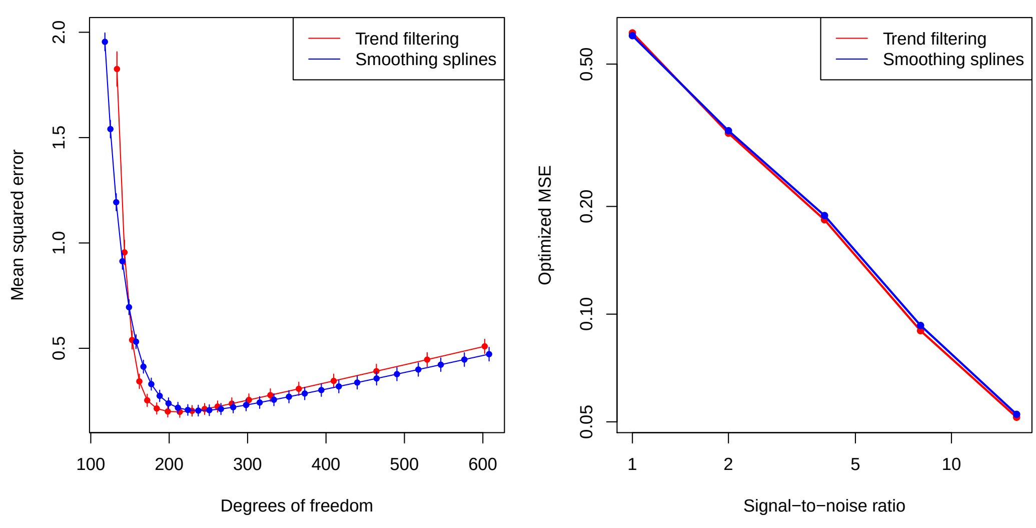

Figure 7: *Results from a simulation setup identical to that described in Section 6.1, and whose results are displayed in Figure 4, except with homogeneous smoothness in the underlying component functions.*

References

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

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

Bauschke, H. & Borwein, J. (1996), 'On projection algorithms for solving convex feasibility problems', *SIAM Review* **38**(3), 367–426.

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.

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.

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

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

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

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.

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

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

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

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

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.

Huang, J. (1998), 'Projection estimation in multiple regression with application to functional ANOVA models', *Annals of Statistics* **26**(1), 242–272.

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

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

Koltchinskii, V. & Yuan, M. (2010), 'Sparsity in multiple kernel learning', *Annals of Statistics* **38**(6), 3660–3695.

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

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.

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

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

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

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.

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

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.

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.

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. & 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.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for nondifferentiable minimization', *Journal of Optimization Theory and Applications* **109**(3), 475–494.

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 de Geer, S. & Muro, A. (2015), 'Penalized least squares estimation in the additive model with different smoothness for the components', *Journal of Statistical Planning and Inference* **162**, 43–61.

van der Burg, E. & de Leeuw, J. (1983), 'Non-linear canonical correlation', *British Journal of Mathematical and Statistical Psychological* **36**(1), 54–80.

Wainwright, M. J. (2017), *High-Dimensional Statistics: A Non-Asymptotic View*, Cambridge University Press. To appear.

Wang, Y.-X., Smola, A. & Tibshirani, R. J. (2014), 'The falling factorial basis and its statistical applications', *International Conference on Machine Learning* **31**.

Yang, Y. & Barron, A. (1999), 'Information-theoretic determination of minimax rates of convergence', *Annals of Statistics* **27**(5), 1564–1599.