The Annals of Statistics
2019, Vol. 47, No. 6, 3032–3068
https://doi.org/10.1214/19-AOS1833

ADDITIVE MODELS WITH TREND FILTERING

BY VEERANJANEYULU SADHANALA AND RYAN J. TIBSHIRANI

Carnegie Mellon University

We study additive models built with trend filtering, that is, additive models whose components are each regularized by the (discrete) total variation of their kth (discrete) derivative, for a chosen integer $k \ge 0$. 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.). Analogous to its advantages in the univariate case, additive trend filtering has favorable theoretical and computational properties, thanks in large part to the localized nature of the (discrete) total variation regularizer that it uses. On the theory side, we derive fast error rates for additive trend filtering estimates, and show these rates are minimax optimal when the underlying function is additive and has component functions whose derivatives are of bounded variation. We also show that these rates are unattainable by additive smoothing splines (and by additive models built from linear smoothers, in general). On the computational side, we use backfitting, to leverage fast univariate trend filtering solvers; we also describe a new backfitting algorithm whose iterations can be run in parallel, which (as far as we can tell) is the first of its kind. Lastly, we present a number of experiments to examine the empirical performance of trend filtering.

1. Introduction. As the dimension of the input space grows large, nonparametric regression turns into a notoriously difficult problem. In this work, we adopt the stance taken by many others, and consider an *additive model* for responses $Y^i \in \mathbb{R}$, $i = 1, \dots, n$ and corresponding 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, for example, X^i_j denotes the jth component of the ith input point. (Exceptions will occasionally be made, but the role of the index should be clear from the context.)

Received April 2018; revised February 2019.
MSC2010 subject classifications. 62G08, 62G20.
Key words and phrases. Additive models, trend filtering, smoothing splines, nonparametric regression, minimax rates, parallel backfitting.

Additive models are a special case of the more general *projection pursuit regression* model of Friedman and Stuetzle (1981). Additive models for the Cox regression and logistic regression settings were studied in Tibshirani (1983) and Hastie (1983), respectively. Some of the first asymptotic theory for additive models was developed in Stone (1985). Two algorithms closely related to (backfitting for) additive models are the *alternating least squares* and *alternating conditional expectations* methods, from van der Burg and de Leeuw (1983) and Breiman and Friedman (1985), respectively. The work of Buja, Hastie and Tibshirani (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 and Tibshirani (1990) is the definitive practical guide for additive models for exponential family data distributions, that is, 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 and Zhang (2006), Meier, van de Geer and Bühlmann (2009), Ravikumar 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, Witten and Simon (2016) study componentwise fused lasso (i.e., total variation) penalization.

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

1.1. *Review*: *Additive models and linear smoothers*.

The influential paper by Buja, Hastie and Tibshirani (1989) studies additive minimization problems of the form

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

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

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

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

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

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

Modeling the component functions as smoothing splines is arguably the most common formulation for additive models, and it is the standard in several statistical software packages like the R package `gam`. However, as Buja, Hastie and Tibshirani (1989) explain, the backfitting perspective suggests a more algorithmic approach to additive modeling: one can replace the operator $(I + \lambda Q_j)^{-1}$ in (2) by S_j, a particular (user-chosen) *linear smoother*, meaning, a linear map that performs univariate smoothing across the jth dimension of inputs X_j. The linear smoothers S_j, $j = 1, \dots, d$ could correspond to smoothing splines, regression splines (regression using a spline basis with given knots), kernel smoothing, local polynomial smoothing or a combination of these, across the input dimensions. In short, as argued in Buja, Hastie and Tibshirani (1989), the class of linear smoothers is broad enough to offer fairly flexible, interesting mechanisms for smoothing, and simple enough to understand precisely. Most of the work following Buja, Hastie and Tibshirani (1989) remains in keeping with the idea of using linear smoothers in combination with additive models.

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 and Johnstone (1998) makes this idea precise. With $d = 1$, suppose that underlying regression function f_0 lies in the univariate function class

$$\mathcal{F}_k(C) = \bigl\{f : \mathrm{TV}\bigl(f^{(k)}\bigr) \leq C\bigr\}, \tag{3}$$

for a constant $C > 0$, where $\mathrm{TV}(\cdot)$ is the total variation operator, and $f^{(k)}$ the kth weak derivative of f. The class in (3) 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 and Johnstone (1998) (see also Section 5.1 of Tibshirani (2014)) imply that the minimax error rate for estimation over $\mathcal{F}_k(C)$ is $n^{-(2k+2)/(2k+3)}$, but the minimax error rate when we consider only linear smoothers (linear transformations of Y) is $n^{-(2k+1)/(2k+2)}$. This difference is highly nontrivial, for example, for $k = 0$ this is a difference of $n^{-2/3}$ (optimal) versus $n^{-1/2}$ (optimal among linear smoothers) for estimating a function f_0 of bounded variation.

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

Two well-studied univariate estimators that are locally adaptive, that is, that attain the minimax error rate over the kth order total variation class in (3), are wavelet smoothing and locally adaptive regression splines, as developed by Donoho and Johnstone (1998) and Mammen and van de Geer (1997), respectively. There is a substantial literature on these methods in the univariate case (especially for wavelets), but fewer authors have considered them in the additive models context. Some notable exceptions are Petersen, Witten and Simon (2016), Sardy and Tseng (2004), Zhang and Wong (2003), with the latter work 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, Didas and Neumann (2006) and Kim et al. (2009), trend filtering is a relatively new approach to univariate nonparametric regression. As explained in Tibshirani (2014), it can be seen as a discrete-time analog of the locally adaptive regression spline estimator. Denoting by $X = (X^1, \dots, X^n) \in \mathbb{R}^n$ the vector of univariate input points, where we assume $X^1 < \cdots < X^n$, the trend filtering estimate of order $k \geq 0$ is defined as the solution of the optimization problem

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \|Y - \theta\|_2^2 + \lambda \|D^{(X,k+1)} \theta\|_1, \tag{4}$$

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

$$D^{(X,1)} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}, \tag{5}$$

and for $k = 1, 2, 3, \ldots,$

$$D^{(X,k+1)} = D^{(X,1)} \operatorname{diag}\left(\frac{k}{X^k - X^1}, \ldots, \frac{k}{X^n - X^{n-k+1}}\right) D^{(X,k)} \in \mathbb{R}^{(n-k-1)\times n}. \tag{6}$$

(The leading matrix $D^{(X,1)}$ in (6) is the $(n-k-1) \times (n-k)$ version of the difference operator in (5).) Intuitively, the interpretation is that the problem (4) penalizes the sum of absolute $(k+1)$st order discrete derivatives of $\theta^1, \ldots, \theta^n$ across the input points $X^1, \ldots, X^n$. Thus, at optimality, the coordinates of the trend filtering solution $\hat{\theta}^1, \ldots, \hat{\theta}^n$ obey a kth order piecewise polynomial form.

This intuition is formalized in Tibshirani (2014) and Wang, Smola and Tibshirani (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, Osher and Fatemi (1992)).

Over the kth order total variation function class defined in (3), Tibshirani (2014), Wang, Smola and Tibshirani (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), Ramdas and Tibshirani (2016), Tibshirani (2014), is that trend filtering estimates are relatively cheap to compute—much cheaper than locally adaptive regression spline estimates—owing to the bandedness of the difference operators in (5), (6), which means that specially implemented convex programming routines can solve (4) 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 $k \geq 0$, defined as a solution in the problem

$$\min_{\theta_1,\ldots,\theta_d \in \mathbb{R}^n} \frac{1}{2}\left\| Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j \right\|_2^2 + \lambda \sum_{j=1}^d \left\| D^{(X_j,k+1)} S_j \theta_j \right\|_1 \quad \text{subject to} \quad \mathbb{1}^T \theta_j = 0, \quad j = 1, \ldots, d. \tag{7}$$

As before, $Y - \bar{Y}\mathbb{1}$ is the centered response vector, $\lambda \geq 0$ is a regularization parameter, and now $S_j \in \mathbb{R}^{n\times n}$ in (7) is a permutation matrix that sorts the jth component of inputs $X_j = (X_j^1, X_j^2, \dots, X_j^n)$ into increasing order, that is,

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

Also, $D^{(X_j, k+1)}$ in (7) is the $(k+1)$st order difference operator, as in (5), (6), but defined over the sorted jth dimension of inputs $S_j X_j$, for $j = 1, \dots, d$. With backfitting (block coordinate descent), computation of a solution in (7) is still quite efficient, since we can leverage the efficient routines for univariate trend filtering.

1.4. *A motivating example.* Figure 1 shows a simulated example that compares the additive trend filtering estimates in (7) (of quadratic order, $k = 2$), to the additive smoothing spline estimates in (1) (of cubic order). In the simulation, we used $n = 3000$ and $d = 3$. We drew input points $X^i \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]^3$, $i = 1, \dots, 3000$, and drew responses $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^3 f_{0j}(X_j^i), \sigma^2)$, $i = 1, \dots, 3000$, where $\sigma = 1.72$ was set to give a signal-to-noise ratio of about 1. The underlying component functions were defined as

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

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

The first row of Figure 1 shows the estimated component functions from additive trend filtering, at a value of λ that minimizes the mean squared error (MSE), computed over 20 repetitions. The second row shows the estimates from additive smoothing splines, also at a value of λ that minimizes the MSE. We see that the trend filtering fits adapt well to the varying levels of smoothness, but the smoothing spline fits are undersmoothed, for the most part. In terms of effective degrees of freedom (df), the additive smoothing spline estimate is much more complex, having about 85 df (computed via Monte Carlo over the 20 repetitions); the additive trend filtering has only about 42 df. The third row of the figure shows the estimates from additive smoothing splines, when λ is chosen so that the resulting df roughly matches that of additive trend filtering in the first row. Now we see that the first component fit is oversmoothed, yet the third is still undersmoothed.

Figure 2 displays the MSE curves from additive trend filtering, as a function of df. We see that trend filtering achieves a lower MSE, and moreover, its MSE curve is optimized at a lower df (i.e., less complex model) than that for smoothing splines. This is analogous to what is typically seen in the univariate setting (Tibshirani (2014)).

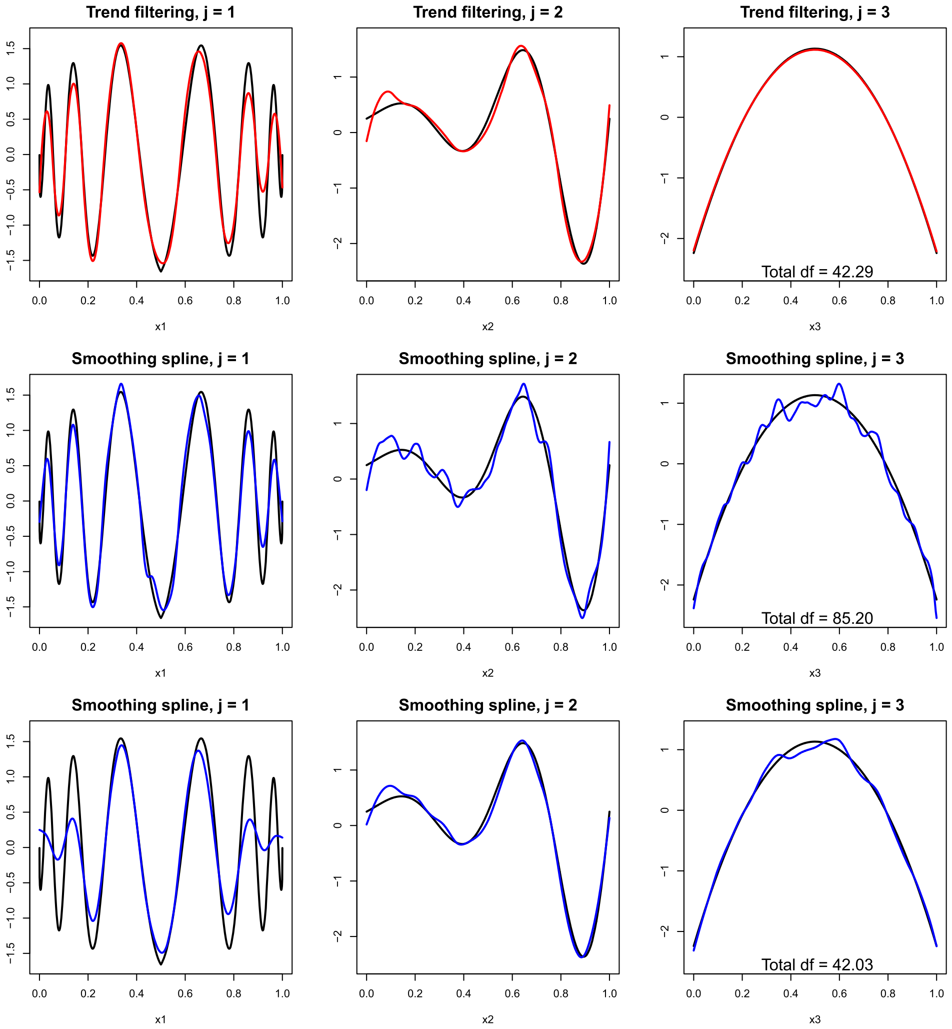

FIG. 1. *Comparing estimates from additive trend filtering* (7) (*of quadratic order*) *and additive smoothing splines* (1) (*of cubic order*), *for a simulation with* $n = 3000$ *and* $d = 2$, *as described in Section* 1.4. *In each row, the underlying component functions are plotted in black. The first row shows the estimated component functions using additive trend filtering, in red, at a value of* λ *chosen to minimize mean squared error* (*MSE*), *computed over* 20 *repetitions. The second row shows the estimates from additive smoothing splines, in blue, again at a value of* λ *that minimizes MSE. The third row shows the estimates from additive smoothing splines when* λ *is tuned so that the effective degrees of freedom* (*df*) *of the fit roughly matches that of additive trend filtering in the first row.*

We note that this motivating example is intended to elucidate the differences in what additive smoothing splines and additive trend filtering can do with a single tuning parameter each; a serious applied statistician, in just $d = 3$ dimensions,

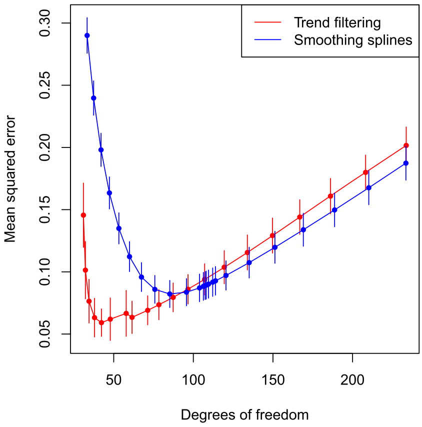

FIG. 2. *MSE curves for additive trend filtering and additive smoothing splines*, *computed over* 20 *repetitions from the same simulation setup as in Figure* 1. *Vertical segments denote* ±1 *standard deviations*. *The MSE curves are parametrized by degrees of freedom* (*computed via standard Monte Carlo methods over the* 20 *repetitions*).

would likely use REML or some related technique to fit a multiple tuning parameter smoothing spline model; see our later discussion on this topic in Section 5.2.

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

- In Section 2, we investigate basic properties of the additive trend filtering model: an equivalent continuous-time formulation, a condition for uniqueness of component function estimates, and a simple formula for the effective degrees of freedom of the additive fit.
- In Section 3, we derive error bounds for additive trend filtering. Assuming that the underlying regression function is additive, denoted by $f_0 = \sum_{j=1}^d f_{0j}$, and that $\mathrm{TV}(f_{0j}^{(k)})$ is bounded, for $j = 1, \dots, d$, we prove that the kth order additive trend filtering estimator converges to f_0 at the rate $n^{-(2k+2)/(2k+3)}$ when the dimension d is fixed (under weak assumptions), and at the rate $dn^{-(2k+2)/(2k+3)}$ when d is growing (under stronger assumptions). We prove that these rates are optimal in a minimax sense, and also show that additive smoothing splines (generally, additive models built from linear smoothers of any kind) are suboptimal over such a class of functions f_0.
- In Section 4, we study the backfitting algorithm for additive trend filtering models, and give a connection between backfitting and an alternating projections scheme in the additive trend filtering dual problem. This inspires a new parallelized backfitting algorithm.

- In Section 5, we present empirical experiments and comparisons, and we also investigate the use of multiple tuning parameter models. In Section 6, we give a brief discussion.

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 (7) as an estimation problem written in *analysis form*. The components are modeled directly by the parameters θ_j, $j = 1, \dots, d$, and the desired structure is established by regularizing the discrete derivatives of these parameters, through the penalty terms $\|D^{(X_j, k+1)} S_j \theta_j\|_1$, $j = 1, \dots, d$. Here, we present an alternative representation for (7) 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 first recall the *falling factorial basis* (Tibshirani (2014), Wang, Smola and Tibshirani (2014)). Given knot points $t^1 < \cdots < t^n \in \mathbb{R}$, the kth order falling factorial basis functions $h_1, \dots, h_n$ are defined by

$$\begin{aligned} h_i(t) &= \prod_{\ell=1}^{i-1} (t - t^\ell), \quad i = 1, \dots, k+1, \\ h_{i+k+1}(t) &= \prod_{\ell=1}^{k} (t - t^{i+\ell}) \cdot 1\{t > t^{i+k}\}, \quad i = 1, \dots, n-k-1. \end{aligned} \tag{8}$$

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 $h_1, \dots, h_n$ 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 equivalent (meaning that they have the same span). Similar to an expansion in the truncated power basis, an expansion in the falling factorial basis,

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

is a continuous piecewise polynomial function, having a global polynomial structure determined by $\alpha^1, \dots, \alpha^{k+1}$, and exhibiting a knot, that is, 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 (8) are not splines, and when g (as defined above) has a knot at a particular location, it displays a change not only in its kth derivative at this location, but also in all lower order derivatives (i.e., all derivatives of orders $1, \dots, k-1$).

Tibshirani (2014), Wang, Smola and Tibshirani (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 $j = 1, \dots, d$, let $h_1^{(X_j)}, \dots, h_n^{(X_j)}$ be the falling factorial basis in (8) with knots $(t^1, \dots, t^n) = S_j X_j$, the jth dimension of the input points, properly sorted. Then the additive trend filtering problem (7) is equivalent to the problem*

$$\begin{aligned} &\min_{\alpha_1,\dots,\alpha_d \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^n \Biggl(Y^i - \bar{Y} - \sum_{j=1}^d \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}(X_j^i) \Biggr)^2 + \lambda k! \sum_{j=1}^d \sum_{\ell=k+2}^n |\alpha_j^\ell| \\ &\quad \text{subject to} \quad \sum_{i=1}^n \sum_{\ell=1}^n \alpha_j^\ell h_\ell^{(X_j)}(X_j^i) = 0, \quad j = 1, \dots, d, \end{aligned} \tag{9}$$

in that, at any solutions in (7), (9), we have

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

An alternative way of expressing problem (9) is

$$\begin{aligned} &\min_{f_j \in \mathcal{H}_j, j=1,\dots,d} \frac{1}{2} \sum_{i=1}^n \Biggl(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \Biggr)^2 + \lambda \sum_{j=1}^d \mathrm{TV}(f_j^{(k)}) \\ &\quad \text{subject to} \quad \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \dots, d, \end{aligned} \tag{10}$$

where $\mathcal{H}_j = \mathrm{span}\{h_1^{(X_j)}, \dots, h_n^{(X_j)}\}$ is the span of the falling factorial basis over the jth dimension, and $f_j^{(k)}$ is the kth weak derivative of f_j, $j = 1, \dots, d$. In this form, at any solutions in (7), (10),

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

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

$$H_{i\ell}^{(X_j,k)} = h_\ell^{(X_j)}(X_j^i), \quad i = 1, \dots, n, \ell = 1, \dots, n. \tag{11}$$

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, Smola and Tibshirani (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 $C^{(X_j,k+1)} \in \mathbb{R}^{(k+1)\times n}$, that is,

$$(H^{(X_j,k)})^{-1} = \begin{bmatrix} C^{(X_j,k+1)} \\ \frac{1}{k!} D^{(X_j,k+1)} \end{bmatrix} S_j. \tag{12}$$

Problem (9) is given by reparameterizing (7) according to $\theta_j = H^{(X_j,k)}\alpha_j$, for $j = 1,\dots,d$. As for (10), the equivalence between this and (9) 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 (7), we can extend each component fit $\hat{\theta}_j$ to the real line, by forming an appropriate linear combination of falling factorial functions:

$$\hat{f}_j(x_j) = \sum_{\ell=1}^n \hat{\alpha}_j^\ell h_\ell^{(X_j)}(x_j), \quad x_j \in \mathbb{R}. \tag{13}$$

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 (12). 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

$$\hat{a}_j = C^{(X_j,k+1)} S_j \hat{\theta}_j, \tag{14}$$

$$\hat{b}_j = \frac{1}{k!} D^{(X_j,k+1)} S_j \hat{\theta}_j. \tag{15}$$

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.1.1 in the Supplementary Material, Sadhanala and Tibshirani (2019)). The last $n-k-1$

coefficients $\hat{b}_j$ index the knot-producing functions $h_{k+2}^{(X_j)}, \dots, h_n^{(X_j)}$, and when $(\hat{b}_j)_\ell = \frac{1}{k!}(D^{(X_j,k+1)} S_j \hat{\theta}_j)_\ell \neq 0$, the fitted function $\hat{f}_j$ exhibits a knot at the $(\ell + k)$th sorted input point among $S_j X_j$, that is, at $X_j^{(\ell+k)}$. Figure 3 gives an example.

We note that the coefficients $\hat{\alpha}_j = (\hat{a}_j, \hat{b}_j)$ in (14), (15) can be computed in $O(n)$ operations and $O(1)$ memory. This makes extrapolation of the jth fitted

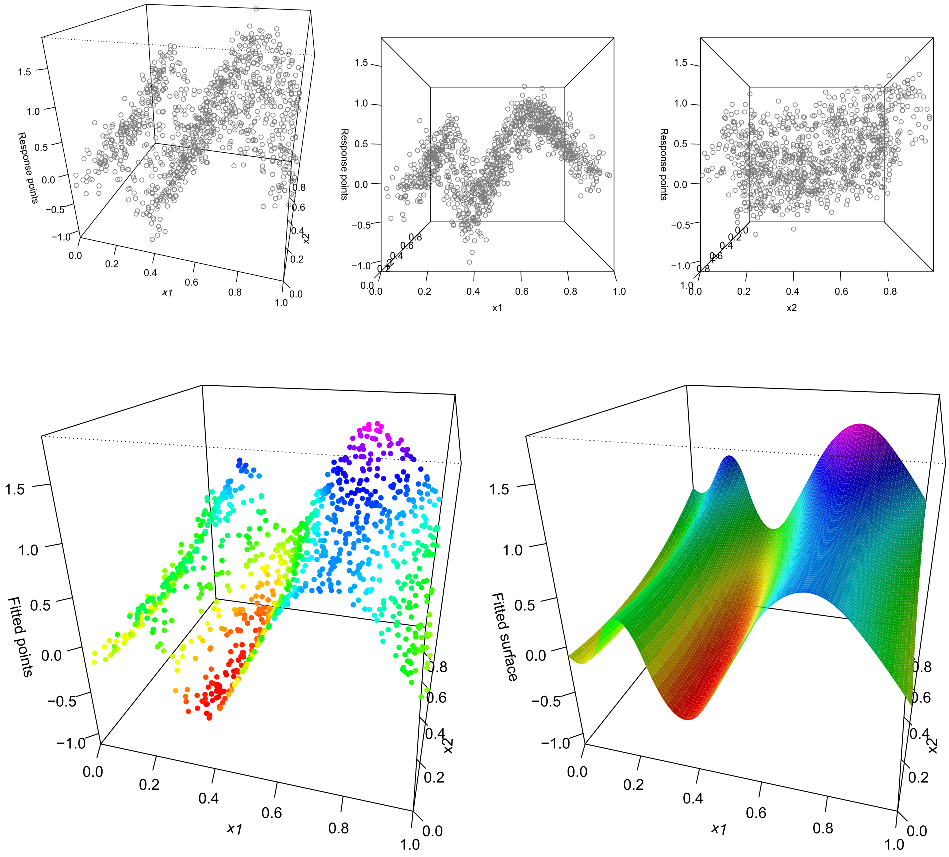

FIG. 3. *An example of extrapolating the fitted additive trend filtering model, where* $n = 1000$ *and* $d = 2$. *We generated input points* $X^i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1]^2$, $i = 1, \dots, 1000$, *and responses* $Y^i \overset{\text{i.i.d.}}{\sim} N(\sum_{j=1}^2 f_{0j}(X_j^i), \sigma^2)$, $i = 1, \dots, 1000$, *where 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* (7) (*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* (13).

function $\hat{f}_j$ in (13) highly efficient. Details are given in Appendix A.1.1 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

2.2. *Uniqueness of component fits.* It is easy to see that, for the problem (7), 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 (7), implies a unique additive fit $T\hat{\theta}$, no matter the choice of solution $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_d) \in \mathbb{R}^{nd}$.

On the other hand, when $d > 1$, the criterion in (7) is not strictly convex in θ, and hence there need not be a unique solution $\hat{\theta}$, that is, the individual components fits $\hat{\theta}_j$, $j = 1, \dots, d$ need not be uniquely determined. We show next that uniqueness of the component fits can be guaranteed under some conditions on the input matrix $X = [X_1 \cdots X_d] \in \mathbb{R}^{n \times d}$. We will rely on the falling factorial representation for additive trend filtering, introduced in the previous subsection, and on the notion of *general position*: a matrix $A \in \mathbb{R}^{m \times p}$ is said to have columns in general position provided that, for any $\ell < \min\{m, p\}$, subset of $\ell + 1$ columns denoted $A_{i_1}, \dots, A_{i_{\ell+1}}$, and signs $s_1, \dots, s_{\ell+1} \in \{-1, 1\}$, the affine span of $\{s_1 A_{i_1}, \dots, s_{\ell+1} A_{i_{\ell+1}}\}$ does not contain any element of $\{\pm A_i : i \neq i_1, \dots, i_{\ell+1}\}$.

LEMMA 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 (11). 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*

$$\tilde{P} = M\left[\tilde{P}^{(X_1,k)} \quad \dots \quad \tilde{P}^{(X_d,k)}\right] \in \mathbb{R}^{n \times dk}, \tag{16}$$

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

$$\tilde{K} = U^T M\left[K^{(X_1,k)} \quad \dots \quad K^{(X_d,k)}\right] \in \mathbb{R}^{(n-kd-1) \times (n-k-1)d}, \tag{17}$$

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 (7) 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, that is, 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 $\tilde{P}$ has full column rank, then the polynomial parts of component fits are uniquely determined, that is, 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.1.2 in the Supplementary Material (Sadhanala and Tibshirani (2019)). 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.

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 (7). Basic arguments in convex analysis, deferred to Appendix A.1.3 in the Supplementary Material (Sadhanala and Tibshirani (2019)), show that the dual of problem (7) can be expressed as

$$\begin{aligned} &\min_{u \in \mathbb{R}^n} \|Y - \bar{Y}\mathbb{1} - u\|_2^2 \quad \text{subject to } u \in U = U_1 \cap \cdots \cap U_d, \\ &\quad \text{where } U_j = \{S_j D_j^T v_j : \|v_j\|_\infty \le \lambda\}, j = 1, \dots, d, \end{aligned} \tag{18}$$

and that primal and dual solutions in (7), (18) are related by

$$\sum_{j=1}^{d} \hat{\theta}_j = Y - \bar{Y}\mathbb{1} - \hat{u}. \tag{19}$$

From the form of (18), 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 (19), 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 4, where we return to the dual perspective, and argue that the backfitting algorithm for the additive trend filtering problem (7) can be seen as a type of alternating projections algorithm for its dual problem (18).

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 and Tibshirani (1990)):

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

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 a class of estimators is well understood (Tibshirani and Taylor (2011, 2012)). The next result is a consequence of existing generalized lasso theory, proved in Appendix A.1.4 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

LEMMA 3 (Degrees of freedom). *Assume the conditions of Lemma 2, that is, that the matrix $\tilde{P}$ in (16) has full column rank, and the matrix $\tilde{K}$ in (17) is in general position. Assume also that the response is Gaussian, $Y \sim N(\eta, \sigma^2 I)$, and treat the input points $X^i \in \mathbb{R}^d$, $i = 1, \dots, n$ as fixed and arbitrary, as well as the tuning parameter value $\lambda \geq 0$. Then the additive trend filtering fit from (7) has degrees of freedom*

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

REMARK 1 (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 $D^{(X_j,k+1)}\hat{\theta}_j$, that is, the number of changes in the discrete $(k+1)$st derivative), add them up over $j = 1, \dots, d$, and add kd. This may seem surprising, as these knot locations are chosen adaptively based on the data Y. But, such adaptivity is counterbalanced by the shrinkage induced by the ℓ_1 penalty in (7) (i.e., for each component fit $\hat{\theta}_j$, there is shrinkage in the differences between the attained kth derivatives on either side of a selected knot). See Tibshirani (2015) for a study of this phenomenon.

2.5. *Two related additive spline estimators.* From its equivalent formulation in (10), additive trend filtering is seen to be closely related to two other additive spline estimators, which we introduce here. Consider, for univariate function classes $\mathcal{S}_j$, $j = 1, \dots, d$, the problem

$$\begin{aligned}&\min_{f_j \in \mathcal{S}_j \, j=1,\dots,d} \frac{1}{2} \sum_{i=1}^{n} \left(Y^i - \bar{Y} - \sum_{j=1}^{d} f_j(X_j^i) \right)^2 + \lambda \sum_{j=1}^{d} \mathrm{TV}(f_j^{(k)}) \\ &\text{subject to} \quad \sum_{i=1}^{n} f_j(X_j^i) = 0, \quad j = 1, \dots, d.\end{aligned} \tag{20}$$

When each $\mathcal{S}_j$, $j = 1, \dots, d$ is the set of k times weakly differentiable functions, we call the solution in (20) the *additive locally adaptive regression spline* of order $k \geq 0$, as it is the natural extension of the univariate estimator considered in Mammen and van de Geer (1997). Denote by $\hat{f}_j$, $j = 1, \dots, d$ this solution; the representation arguments used by these authors apply immediately to the additive setting, and imply that each $\hat{f}_j$, $j = 1, \dots, d$ is indeed a spline of degree k (justifying the choice of name). The same arguments show that, for $k = 0$ or $k = 1$, the knots of the spline $\hat{f}_j$ lie among the jth dimension of the input points $X_j^1, \dots, X_j^n$, for $j = 1, \dots, d$, but for $k \geq 2$, this need not be true, and in general the components will be splines with knots at locations other than the inputs.

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

$$T_j = \begin{cases} \{t_j^{k/2+2}, \dots, t_j^{n-k/2}\} & \text{if } k \text{ is even,} \\ \{t_j^{(k+1)/2+1}, \dots, t_j^{n-(k+1)/2}\} & \text{if } k \text{ is odd,} \end{cases} \tag{21}$$

that is, defined by removing $k + 1$ input points at the boundaries, for $j = 1, \dots, d$. Setting $\mathcal{S}_j = \mathcal{G}_j$, $j = 1, \dots, d$ makes (20) a finite-dimensional problem, just like (10). When $k = 0$ or $k = 1$, as is evident from their form in (8), the falling factorial functions are simply splines, which means that $\mathcal{H}_j = \mathcal{G}_j$ for $j = 1, \dots, d$, hence additive trend filtering and restricted additive locally adaptive regression splines are the same estimator. When $k \geq 2$, this is no longer true, and they are not the same. Additive trend filtering will be much easier to compute, since $\mathrm{TV}(g^{(k)})$ does not admit a nice representation in terms of discrete derivatives for a kth order spline (and yet it does for a kth order falling factorial function, as seen in (7)).

To summarize, additive locally adaptive splines, restricted additive locally adaptive splines, and additive trend filtering all solve a problem of the form (20) for different choices of function classes $\mathcal{S}_j$, $j = 1, \dots, d$. For $k = 0$ or $k = 1$, these three estimators are equivalent. For $k \geq 2$, they will be generically different, though our intuition tells us that their differences should not be too large: the unrestricted problem admits a solution that is a spline in each component; the restricted problem simply forces these splines to have knots at the input points; and the trend filtering problem swaps splines for falling factorial functions, which are highly similar in form. Next, we give theory that confirms this intuition, in large samples.

3. Error bounds.

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

3.1. *Notation.*

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

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

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

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

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

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

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

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

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

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

3.2. *Error bounds for a fixed dimension d.* We consider error bounds for the generic roughness-penalized estimator defined as a solution of

$$\begin{aligned} &\min_{f_j \in \mathcal{S}_j, j=1,\dots,d} \frac{1}{2} \sum_{i=1}^n \left(Y^i - \bar{Y} - \sum_{j=1}^d f_j(X_j^i) \right)^2 + \lambda \sum_{j=1}^d J(f_j) \\ &\text{subject to} \quad \sum_{i=1}^n f_j(X_j^i) = 0, \quad j = 1, \dots, d, \end{aligned} \tag{22}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We now state our main result in the fixed d case, which is proved in Appendices A.1.5, A.1.6 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

THEOREM 1. *Assume A1, B1 on the data distribution, and assume C1, C2, C3 on the seminorm J. Also, assume that the dimension d of the input space is fixed. Let $C_n \ge 1$ be an arbitrary sequence. There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on d, σ, k, L, K, w, such that for all $c \ge c_1$, $n \ge n_0$, and tuning parameter values $\lambda \ge cn^{w/(2+w)}C_n^{-(2-w)/(2+w)}$, any solution in (22) satisfies*

$$\left\|\sum_{j=1}^d \hat{f}_j - f_0\right\|_n^2 \le \left\|\sum_{j=1}^d \tilde{f}_j - f_0\right\|_n^2 + \frac{6\lambda}{n}\max\left\{C_n, \sum_{j=1}^d J(\tilde{f}_j)\right\}, \tag{23}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$, simultaneously over all $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, feasible for the problem (22), such that $\|\tilde{f} - f_0\|_n \le \max\{C_n, \sum_{j=1}^d J(\tilde{f}_j)\}$.

REMARK 2 (Error bound for additive, J-smooth f_0). Assume $f_0 = \sum_{j=1}^d f_{0j}$, where $f_{0j} \in \mathcal{S}_j$, $j = 1, \dots, d$, and $\sum_{j=1}^d J(f_{0j}) \le C_n$. Letting $\tilde{f} = f_0$, the approximation error term in (23) (the first term on the right-hand side) is zero, and for $\lambda = cn^{w/(2+w)}C_n^{-(2-w)/(2+w)}$, the result in the theorem reads

$$\left\|\sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j}\right\|_n^2 \le 6cn^{-2/(2+w)}C_n^{2w/(2+w)}, \tag{24}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$. As we will see in the minimax lower bound in Theorem 3 (plugging in $c_n = C_n/d$, and taking d to be a constant), the rate $n^{-2/(2+w)} C_n^{2w/(2+w)}$ is optimal for such a class of functions.

REMARK 3 (Distance to best additive, J-smooth approximation of f_0). The arguments used to establish the oracle-type inequality (23) also imply a result on the empirical norm error between $\hat{f}$ and the best additive approximation of f_0. To be precise, let $(f_1^{\text{best}}, \dots, f_d^{\text{best}})$ denote a solution in the population-level problem

$$\begin{aligned} &\min_{f_j \in \mathcal{S}_j, j=1,\dots,d} \frac{1}{2} \sum_{i=1}^{n} \Bigg(f_0(X^i) - \sum_{j=1}^{d} f_j(X_j^i) \Bigg)^2 + \frac{\lambda}{2} \sum_{j=1}^{d} J(f_j) \\ &\text{subject to} \quad \sum_{i=1}^{n} f_j(X_j^i) = 0, \quad j = 1, \dots, d. \end{aligned} \tag{25}$$

We note that (25) has "half" of the regularization of problem (22), as it uses a penalty parameter of $\lambda/2$ versus λ. We can think of $f^{\text{best}} = \sum_{j=1}^d f_j^{\text{best}}$ as the best additive, J-smooth approximation of f_0, where λ as usual controls the level of smoothness. The following is a consequence of the proof of Theorem 1, verified in Appendix A.1.7 in the Supplementary Material (Sadhanala and Tibshirani (2019)): assume that $\|f^{\text{best}} - f_0\|_n \le \max\{C_n, \sum_{j=1}^d J(f_j^{\text{best}})\}$ almost surely (with respect to Q), for sufficiently large λ; then any solution in (22) satisfies for all $c \ge c_1$, $n \ge n_0$, and $\lambda \ge c n^{w/(2+w)} C_n^{-(2-w)/(2+w)}$,

$$\Bigg\| \sum_{j=1}^{d} \hat{f}_j - \sum_{j=1}^{d} f_j^{\text{best}} \Bigg\|_n^2 \le \frac{6\lambda}{n} \max\Bigg\{ C_n, \sum_{j=1}^{d} J\big(f_j^{\text{best}}\big) \Bigg\}, \tag{26}$$

with probability at least $1 - \exp(-c_2 c) - \exp(-c_3\sqrt{n})$, where as before $c_1, c_2, c_3, n_0 > 0$ are constants that depend only on d, σ, k, L, K, w. Notably, the right-hand side in the bound (26) does not depend on the approximation error; in particular, we do not even require $\|f^{\text{best}} - f_0\|_n$ to converge to zero. This is analogous to classical results from Stone (1985).

We examine a special case of the generic problem (22) when the regularizer is $J(g) = \mathrm{TV}(g^{(k)})$, and derive implications of the above Theorem 1 for additive locally regression adaptive splines and additive trend filtering, corresponding to different choices of the function classes $\mathcal{S}_j$, $j = 1, \dots, d$ in (22). We must introduce an additional (weak) assumption on the input distribution, for the results on restricted locally adaptive regression splines and trend filtering.

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

Here is our result for additive locally adaptive splines and additive trend filtering. The proof is given in Appendices A.1.8, A.1.9 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

COROLLARY 1. *Assume* A1, B1 *on the data distribution. Also, assume that the dimension d of the input space is fixed, and that the underlying regression function is additive, $f_0 = \sum_{j=1}^d f_{0j}$, where the components f_{0j}, $j = 1, \dots, d$ are k times weakly differentiable, such that $\sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}) \le C_n$ for a sequence $C_n \ge 1$. For $J(g) = \mathrm{TV}(g^{(k)})$, Assumptions* C1, C2, C3 *hold with $L = 1$ and $w = 1/(k+1)$. Furthermore, the following is true of the estimator defined by problem* (22):

(a) *Let $\mathcal{S}_j$ be the set of all k times weakly differentiable functions, for each $j = 1, \dots, d$. There are constants $c_1, c_2, c_3, n_0 > 0$, depending only on d, σ, k, such that for all $c \ge c_1$ and $n \ge n_0$, any solution in the additive locally adaptive regression spline problem* (22), *with tuning parameter value $\lambda = cn^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)}$, satisfies*

$$\left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_n^2 \le cn^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}, \tag{27}$$

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

(b) *Let $\mathcal{S}_j = \mathcal{G}_j$, the set of kth degree splines with knots in the set T_j in* (21), *for $j = 1, \dots, d$, and assume* A2 *on the input density. Then there are constants $c_1, c_2, c_3, n_0 > 0$, that depend only on d, b_0, σ, k, such that for all $c \ge c_1$ and $n(\log n)^{-(1+1/k)} \ge n_0 C_n^{(2k+2)/(2k^2+2k-1)}$, any solution in the restricted additive locally adaptive spline problem* (22), *with $\lambda = cn^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)}$, satisfies the same result in* (27), *with probability at least $1 - \exp(-c_2 c) - c_3/n$.*

(c) *Let $\mathcal{S}_j = \mathcal{H}_j$, the set of kth degree falling factorial functions defined over X_j (the jth dimension of inputs), for $j = 1, \dots, d$, and assume* A2. *Then there exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on d, b_0, σ, k, such that for all $c \ge c_1$ and $n(\log n)^{-(2k+3)} \ge n_0 C_n^{4k+4}$, any solution in the additive trend filtering problem* (22), *with $\lambda = cn^{1/(2k+3)} C_n^{-(2k+1)/(2k+3)}$, satisfies* (27), *with probability at least $1 - \exp(-c_2 c) - c_3/n$.*

REMARK 4 (Spline and falling factorial approximants). For part (a) of the corollary, the approximation error (the first term on the right-hand side) in (24) is zero by definition, and we need only verify Assumptions C1, C2, C3 for the regularizer $J(g) = \mathrm{TV}(g^{(k)})$. Parts (b) and (c) require control over the approximation error, because the underlying regression function $f_0 = \sum_{j=1}^d f_{0j}$ need not have components that lie in the chosen function spaces $\mathcal{S}_j$, $j = 1, \dots, d$. To be clear: for $k = 0$ or $k = 1$, as discussed in Section 2.5, all three problems considered in parts (a), (b), (c) are equivalent; hence parts (b) and (c) really only concern the case

$k \geq 2$. For both of these parts, we control the approximation error by controlling the univariate approximation error and then applying the triangle inequality. For part (b), we use a special spline quasi-interpolant from Proposition 7 in Mammen and van de Geer (1997) (who in turn construct this using results from de Boor (1978)); for part (c), we develop a new falling factorial approximant that may be of independent interest.

3.3. *Error bounds for a growing dimension d.* In this subsection, we allow the input dimension d to grow with the sample size n. To keep our analysis as clean as possible, we consider a constrained version of the problem (22), namely

$$\begin{aligned} &\min_{f_j \in \mathcal{S}_j, j=1,\dots,d} \frac{1}{2} \sum_{i=1}^{n} \Bigg(Y^i - \bar{Y} - \sum_{j=1}^{d} f_j(X_j^i) \Bigg)^2 \\ &\text{subject to} \quad \sum_{i=1}^{n} f_j(X_j^i) = 0, \qquad J(f_j) \leq \delta, \quad j = 1, \dots, d, \end{aligned} \tag{28}$$

for a tuning parameter $\delta > 0$. (The penalized problem (22) can also be analyzed in the setting of growing d, but we find that the analysis is messier and requires more assumptions in order to obtain the same results.) Instead of A1, we now use the following assumption in the input distribution.

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

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

$$\Bigg\| \sum_{j=1}^{d} m_j \Bigg\|_2^2 = \sum_{j=1}^{d} \|m_j\|_2^2. \tag{29}$$

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 (29), is only true because of the product form $Q = Q_1 \times \cdots \times Q_d$. Such decomposability is not generally possible with the empirical norm. In the proof of Theorem 2, we move from considering the empirical norm of the error vector to the L_2 norm, in order to leverage the property in (29), which eventually leads to an error rate that has a linear dependence on the dimension d. In the absence of L_2 decomposability, the same error rate can be achieved with a weaker incoherence bound, as in (34); see Remark 7 after the theorem.

We now state our main result in the growing d case, whose proof is in Appendices A.1.10, A.1.11 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

THEOREM 2. *Assume* A3, B1 *on the data distribution, and assume* C1, C2, C3 *on the seminorm* J. *Let* $\delta \geq 1$ *be arbitrary. There are constants* $c_1, c_2, c_3, n_0 > 0$, *that depend only on* $b_1, b_2, \sigma, k, L, K, w$, *such that for all* $c \geq c_1$ *and* $n \geq n_0(d\delta)^{1+w/2}$, *any solution in* (28) *satisfies both*

$$\left\|\sum_{j=1}^d \hat{f}_j - f_0\right\|_n^2 \leq \left\|\sum_{j=1}^d \tilde{f}_j - f_0\right\|_n^2 + cdn^{-2/(2+w)}\delta, \tag{30}$$

$$\left\|\sum_{j=1}^d \hat{f}_j - f_0\right\|_2^2 \leq 2\left\|\sum_{j=1}^d \tilde{f}_j - f_0\right\|_2^2 + 24\left\|\sum_{j=1}^d \tilde{f}_j - f_0\right\|_n^2 + cdn^{-2/(2+w)}\delta^2, \tag{31}$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$, *simultaneously over all functions* $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$, *feasible for the problem* (28).

REMARK 5 (Error bound for additive, J-smooth f_0). Assume $f_0 = \sum_{j=1}^d f_{0j}$, where $f_{0j} \in \mathcal{S}_j$ and $J(f_{0j}) \leq c_n$, $j = 1, \ldots, d$, for a sequence $c_n \geq 1$. Letting $\tilde{f} = f_0$, and $\delta = c_n$, the results in (30), (31) translate to

$$\begin{aligned} &\left\|\sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j}\right\|_n^2 \leq cdn^{-2/(2+w)}c_n \quad \text{and} \\ &\left\|\sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j}\right\|_2^2 \leq cdn^{-2/(2+w)}c_n^2, \end{aligned} \tag{32}$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$, provided that $n \geq n_0(dc_n)^{1+w/2}$. From the minimax lower bound in Theorem 3, we can see that the optimal rate for such a class of functions is in fact $dn^{-2/(2+w)}c_n^{2w/(2+w)}$, which reveals that the rates in (32) are tight when c_n is a constant, but not when c_n grows with n. It is worth noting that the dependence of the bounds on c_n in Theorem 2 (and hence in (32)) can be improved to have the optimal scaling of $c_n^{2w/(2+w)}$ by assuming that f_0 is sup norm bounded, and additionally placing a sup norm bound on the components in (28). This feels like an unnecessary restriction, so we prefer to present results without it, as in Theorem 2 (and (32)).

REMARK 6 (Distance to best additive, J-smooth approximation of f_0). A consequence of the proof of (30) is a bound on the empirical norm error between $\hat{f}$ and the best additive approximation of f_0. To be precise, let $f^{\text{best}} = \sum_{j=1}^d f_j^{\text{best}}$ minimize $\| \sum_{j=1}^d \tilde{f}_j - f_0\|_n^2$ over all additive functions $\tilde{f} = \sum_{j=1}^d \tilde{f}_j$ feasible for problem (28). Then following directly from (A.36) in the proof of Theorem 2, we have for all $c \geq c_1$ and $n \geq n_0(d\delta)^{1+w/2}$,

$$\left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_j^{\text{best}} \right\|_n^2 \leq cdn^{-2/(2+w)}\delta, \tag{33}$$

with probability at least $1 - \exp(-c_2 c) - c_3/n$, where again $c_1, c_2, c_3, n_0 > 0$ are constants that depend on $b_1, b_2, \sigma, k, L, K, w$. Just as we saw in fixed d case, the right-hand side in (33) does not depend on the approximation error $\|f^{\text{best}} - f_0\|_n$, which is analogous to classical results from Stone (1985).

REMARK 7 (L_2 decomposability and incoherence). The decomposability property in (29) is critical in obtaining the sharp (linear) dependence on d in the error rates (30), (31). However, it is worth noting that all that is needed in the proof is in fact a lower bound of the form

$$\left\| \sum_{j=1}^d m_j \right\|_2^2 \geq \phi_0 \sum_{j=1}^d \|m_j\|_2^2, \tag{34}$$

for a constant $\phi_0 > 0$, rather than an equality, as in (29). The above is an incoherence condition that can hold for nonproduct distributions Q, over an appropriate class of functions (additive functions with smooth components), provided that the correlations between components of Q are not too large. See Meier, van de Geer and Bühlmann (2009), van de Geer (2014) for similar incoherence conditions.

Next, we present our results for additive locally adaptive regression splines (both unrestricted and restricted variants) and additive trend filtering. The proof is in Appendix A.1.12 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

COROLLARY 2. *Assume A3, B1 on the data distribution. Also, assume that the underlying regression function is additive, $f_0 = \sum_{j=1}^d f_{0j}$, where the components f_{0j}, $j = 1, \ldots, d$ are k times weakly differentiable, such that* $\mathrm{TV}(f_{0j}^{(k)}) \leq c_n$, *$j = 1, \ldots, d$, for a sequence $c_n \geq 1$. Then for $J(g) = \mathrm{TV}(g^{(k)})$, the following is true of the estimator defined by problem* (28):

(a) *Let S_j be the space of all k times weakly differentiable functions, for each $j = 1, \ldots, d$. There exist constants $c_1, c_2, c_3, n_0 > 0$, that depend only on b_1, b_2, σ, k, such that for all $c \geq c_1$ and $n \geq n_0(dc_n)^{(2k+3)/(2k+2)}$, any solution in* *the constrained-form additive locally adaptive spline problem* (28), *with tuning parameter* $\delta = c_n$, *satisfies*

$$\left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_n^2 \le cdn^{-(2k+2)/(2k+3)} c_n \quad \textit{and}$$
$$(35)\qquad \left\| \sum_{j=1}^d \hat{f}_j - \sum_{j=1}^d f_{0j} \right\|_2^2 \le cdn^{-(2k+2)/(2k+3)} c_n^2,$$

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

(b) *Let* $\mathcal{S}_j = \mathcal{G}_j$, *the set of kth degree splines with knots in the set* T_j *in* (21), *for* $j = 1, \dots, d$. *There exist constants* $c_1, c_2, c_3, n_0 > 0$, *that depend only on* b_1, b_2, σ, k, *such that for* $c \ge c_1$ *and* $n \ge (dc_n)^{(2k+3)/(2k+2)}$, *any solution in the constrained-form restricted additive locally adaptive spline problem* (28), *with tuning parameter* $\delta = a_k c_n$, *where* $a_k \ge 1$ *is a constant that depends only on* k, *satisfies* (35), *with probability at least* $1 - \exp(-c_2 c) - c_3 d/n$.

(c) *Let* $\mathcal{S}_j = \mathcal{H}_j$, *the set of kth degree falling factorial functions defined over* X_j (*the jth dimension of input points*), *for* $j = 1, \dots, d$. *Then there are constants* $c_1, c_2, c_3, n_0 > 0$, *depending only on* b_1, b_2, σ, k, *such that for all* $c \ge c_1$ *and* $n \ge n_0(dc_n)^{(2k+3)/(2k+2)}$, *any solution in the constrained-form additive trend filtering problem* (28), *with tuning parameter* $\delta = a_k c_n$, *where* $a_k \ge 1$ *is a constant depending only on* k, *satisfies* (35), *with probability at least* $1 - \exp(-c_2 c) - c_3 d/n$.

3.4. *Minimax lower bounds.*

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

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

ASSUMPTION B2. The responses Y^i, $i = 1, \dots, n$ follow

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

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

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

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

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

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

Let us introduce the notation

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

Now we state our main minimax lower bound. The proof is given in Appendices A.1.13, A.1.14 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

THEOREM 3. *Assume A4, B2 on the data distribution, and C1, C2, C4 on the seminorm J. Then there exist constants $c_0, n_0 > 0$, that depend only on $\sigma, k, L, K_1, K_2, w$, such that for all $c_n \geq 1$ and $n \geq n_0 d^{1+w/2} c_n^w$, we have*

$$\inf_{\hat f} \sup_{f_0 \in B_J^d(c_n)} \mathbb{E}\|\hat f - f_0\|_2^2 \geq c_0 d n^{-2/(2+w)} c_n^{2w/(2+w)}. \tag{36}$$

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

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

and Theorem 3 implies the following result, whose proof is in Appendix A.1.15 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

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

$$\inf_{\hat f} \sup_{f_0 \in \mathcal{F}_k^d(c_n)} \mathbb{E}\|\hat f - f_0\|_2^2 \geq c_0 d n^{-(2k+2)/(2k+3)} c_n^{2/(2k+3)}. \tag{37}$$

REMARK 8 (Optimality for a fixed dimension d). For a fixed d, the estimator defined by (22) is minimax rate optimal over the class of additive functions f_0 such that $\sum_{j=1}^d J(f_{0j}) \leq C_n$. To see this, note that such a class of functions contains $B_J^d(C_n/d)$, therefore plugging $c_n = C_n/d$ into the right-hand side in (36) yields a lower bound rate of $n^{-2/(2+w)} C_n^{2w/(2+w)}$, which matches the upper bound rate in (24).

Furthermore, when $J(g) = \mathrm{TV}(g^{(k)})$, the lower bound rate given by plugging $c_n = C_n/d$ into the right-hand side in (37) is $n^{-(2k+2)/(2k+3)} C_n^{2/(2k+3)}$, matching the upper bound rate in (27). Hence additive locally adaptive regression splines, restricted additive locally adaptive regression splines, and additive trend filtering all achieve the minimax rate over the space of additive functions f_0 such that $\sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}) \le C_n$.

REMARK 9 (Optimality for a growing dimension d). For growing d, the estimator defined by (28) is minimax rate optimal over the class of additive functions f_0 such that $J(f_{0j}) \le c$, $j = 1, \dots, d$, where $c > 0$ is a constant. This is verified by noting that the lower bound rate of $dn^{-2/(2+w)}$ in (36) matches the upper bound rates in (30), (31).

When $J(g) = \mathrm{TV}(g^{(k)})$, and again, $c_n = c$ (a constant), the lower bound rate of $dn^{-(2k+2)/(2k+3)}$ in (37) matches the upper bound rates in (35). Thus additive locally adaptive regression splines, restricted additive locally adaptive regression splines, and additive trend filtering all attain the minimax rate over the space of additive functions f_0 with $\mathrm{TV}(f_{0j}^{(k)}) \le c$, $j = 1, \dots, d$.

For growing c_n, we note that the upper bounds in (32) and (35) have an inflated dependence on c_n, compared to (36) and (37). It turns out that the latter (lower bounds) are tight, and the former (upper bounds) are loose. The upper bounds can be tightened under further boundedness assumptions (see Remark 5).

REMARK 10 (Suboptimality of additive linear smoothers). Seminal theory from Donoho and Johnstone (1998) on minimax linear rates over Besov spaces shows that, under Assumption B2, and with the inputs X^i, $i = 1, \dots, n$ being now nonrandom and occurring over the regular d-dimensional lattice $\{1/N, 2/N, \dots, 1\}^d \subseteq [0, 1]^d$ with $N = n^{1/d}$, we have

$$\inf_{\hat{f} \text{ additive linear}} \ \sup_{f_0 \in \mathcal{F}_k^d(c_n)} \mathbb{E}\|\hat{f} - f_0\|_2^2 \ge c_0 d n^{-(2k+1)/(2k+2)} c_n^{2/(2k+2)}, \tag{38}$$

for all $n \ge n_0$, where $c_0, n_0 > 0$ are constants, depending only on σ, k. On the left-hand side in (38), the infimum is taken over all additive linear smoothers, that is, estimators $\hat{f} = \sum_{j=1}^d \hat{f}_j$ such that each component $\hat{f}_j$ is a linear smoother, for $j = 1, \dots, d$. The additive linear smoother lower bound (38) is verified in Appendix A.1.16 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

For a fixed d, we can see that all additive linear smoothers, for example, additive smoothing splines, additive kernel smoothing estimators, additive RKHS estimators, etc. are suboptimal over the class of additive functions f_0 with $\sum_{j=1}^d \mathrm{TV}(f_{0j}^{(k)}) \le C_n$, as the optimal linear rate in (38) (set $c_n = C_n/d$) is

$n^{-(2k+1)/(2k+2)} C_n^{2/(2k+2)}$, slower than the optimal rate $n^{-(2k+2)/(2k+3)} C_n^{2/(2k+2)}$ of additive locally adaptive splines and additive trend filtering in (27).

For growing d, and $c_n = c$ (a constant), we also see that additive linear smoothers are suboptimal over the class of additive functions f_0 such that $\mathrm{TV}(f_{0j}^{(k)}) \le c$, $j = 1, \dots, d$, as the optimal linear rate in (38) is $dn^{-(2k+1)/(2k+2)}$, slower than the optimal rate $dn^{-(2k+2)/(2k+3)}$ of additive locally adaptive regression splines and additive trend filtering in (35).

4. Backfitting and the dual. We now examine computational approaches for the additive trend filtering problem (7). This is a convex optimization problem, and many standard approaches can be applied. For its simplicity and its ubiquity in additive modeling, we focus on the backfitting algorithm in particular.

4.1. *Backfitting.* The backfitting approach for problem (7) is described in Algorithm 1. We write $\mathrm{TF}_\lambda(r, X_j)$ for the univariate trend filtering fit, with a tuning parameter $\lambda > 0$, to a response vector $r = (r^1, \dots, r^n) \in \mathbb{R}^n$ over an input vector $X_j = (X_j^1, \dots, X_j^n) \in \mathbb{R}^n$. In words, the algorithm cycles over $j = 1, \dots, d$, and at each step updates the estimate for component j by applying univariate trend filtering to the jth partial residual (i.e., the current residual excluding component j). Centering in Step 2b part (ii) is optional, because the fit $\mathrm{TF}_\lambda(r, X_j)$ will have mean zero whenever r has mean zero, but centering can still be performed for numerical stability. In general, the efficiency of backfitting hinges on the efficiency of the univariate smoother employed; to implement Algorithm 1 in practice we can use fast interior point methods (Kim et al. (2009)) or fast operator splitting methods (Ramdas and 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 (7) over the coordinate blocks θ_j,

Algorithm 1 Backfitting for additive trend filtering

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

1. Set $t = 0$ and initialize $\theta_j^{(0)} = 0$, $j = 1, \dots, d$.
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) $\theta_j^{(t)} = \theta_j^{(t)} - \frac{1}{n}\mathbb{1}^T \theta_j^{(t)}$
3. Return $\hat{\theta}_j$, $j = 1, \dots, d$ (parameters $\theta_j^{(t)}$, $j = 1, \dots, d$ at convergence).
```

$j = 1, \dots, d$. 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 (7), all limit points of the sequence of iterates produced by BCD are optimal solutions. We are primarily interested in developing a connection between BCD for problem (7) and alternating projections in its dual problem (18), which is the topic of the next subsection.

4.2. *Dual alternating projections.* Using the additive trend filtering problem (7) and its dual (18), related by the transformation (19), we see that for any dimension $j = 1, \dots, d$, the univariate trend filtering fit with response vector $r = (r^1, \dots, r^n)$ and input vector $X_j = (X_j^1, \dots, X_j^n)$ becomes

$$\mathrm{TF}_\lambda(r, X_j) = (\mathrm{Id} - \Pi_{U_j})(r), \tag{39}$$

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

$$\begin{aligned} u_0^{(t)} &= Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j^{(t-1)}, \\ u_j^{(t)} &= \Pi_{U_j}\big(u_{j-1}^{(t)} + \theta_j^{(t-1)}\big), \quad j = 1, \dots, d, \\ \theta_j^{(t)} &= \theta_j^{(t-1)} + u_{j-1}^{(t)} - u_j^{(t)}, \quad j = 1, \dots, d. \end{aligned} \tag{40}$$

Thus the backfitting algorithm for (7), as expressed above in (40), is seen to be a particular type of *alternating projections* method applied to the dual problem (18), cycling through projections onto U_j, $j = 1, \dots, d$. Interestingly, as opposed to the classical alternating projections approach, which would repeatedly project the current iterate $u_{j-1}^{(t)}$ onto U_j, $j = 1, \dots, d$, the steps in (40) repeatedly project an "offset" version $u_{j-1}^{(t)} + \theta_j^{(t-1)}$ of the current iterate, for $j = 1, \dots, d$.

4.3. *Parallelized backfitting.* We have seen that backfitting is a special type of alternating projections algorithm, applied to the dual problem (18). For set intersection problems (where we seek a point in the intersection of given closed, convex sets), the optimization literature offers a variety of *parallel projections* methods (in contrast to alternating projections methods) that are provably convergent. One such method can be derived using ADMM (e.g., see Section 5.1 of Boyd et al. (2011)), and a similar construction can be used for the dual problem (18). We first rewrite this problem as

$$\begin{aligned} \min_{u_0, u_1, \dots, u_d \in \mathbb{R}^n} \; & \frac{1}{2}\|Y - \bar{Y}\mathbb{1} - u_0\|_2^2 + \sum_{j=1}^d I_{U_j}(u_j) \\ \text{subject to} \quad & u_0 = u_1, u_0 = u_2, \dots, u_0 = u_d, \end{aligned} \tag{41}$$

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

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

The ADMM steps for (41) are now given by repeating, for $t = 1, 2, 3, \dots,$

$$\begin{aligned} u_0^{(t)} &= \frac{1}{\rho d + 1}\Biggl(Y - \bar{Y}\mathbb{1} + \rho \sum_{j=1}^d \bigl(u_j^{(t-1)} - \gamma_j^{(t-1)}\bigr) \Biggr), \\ u_j^{(t)} &= \Pi_{U_j}\bigl(u_0^{(t)} + \gamma_j^{(t-1)}\bigr), \quad j = 1, \dots, d, \\ \gamma_j^{(t)} &= \gamma_j^{(t-1)} + u_0^{(t)} - u_j^{(t)}, \quad j = 1, \dots, d. \end{aligned} \tag{42}$$

Now compare (42) to (40)—the key difference is that in (42), the updates to u_j, $j = 1, \dots, d$, that is, 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 (40).

Of course, for our current study, the dual problem (41) is really only interesting insofar as it is connected to the additive trend filtering problem (7). In Algorithm 2, we transcribe the iterations in (42) into an equivalent primal form, and we provide a convergence guarantee in the next theorem. For details, see Appendix A.1.17 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

THEOREM 4. *Initialized arbitrarily, the ADMM steps* (42) *produce parameters* $\hat{\gamma}_j$, $j = 1, \dots, d$ (*i.e., the iterates* $\gamma_j^{(t)}$, $j = 1, \dots, d$ *at convergence*) *such that*

Algorithm 2 Parallel backfitting for additive trend filtering

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

1. Initialize $u_0^{(0)} = 0$, $\theta_j^{(0)} = 0$ and $\theta_j^{(-1)} = 0$ for $j = 1, \dots, d$.
2. For $t = 1, 2, 3, \dots$ (until convergence):
   a. $u_0^{(t)} = \frac{1}{\rho d+1}(Y - \bar{Y}\mathbb{1} - \sum_{j=1}^d \theta_j^{(t-1)}) + \frac{\rho d}{\rho d+1}(u_0^{(t-1)} + \frac{1}{\rho d}\sum_{j=1}^d(\theta_j^{(t-2)} - \theta_j^{(t-1)}))$
   b. For $j = 1, \dots, d$ (in parallel):
         (i) $\theta_j^{(t)} = \rho \cdot \mathrm{TF}_\lambda(u_0^{(t)} + \theta_j^{(t-1)}/\rho, X_j)$
         (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).
```

the scaled parameters $\rho\hat{\gamma}_j$, $j = 1, \ldots, d$ *solve additive trend filtering* (7). *Further, the outputs* $\hat{\theta}_j$, $j = 1, \ldots, d$ *of Algorithm* 2 *solve additive trend filtering* (7).

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

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

then the update steps for $\theta_j^{(t)}$, $j = 1, \ldots, d$ that follow would be just given by applying univariate trend filtering to each partial residual (without sequentially updating the partial residuals between trend filtering runs). This naive parallel method has no convergence guarantees, and can fail even in simple practical examples to produce optimal solutions. Importantly, Algorithm 2 does not take $u_0^{(t)}$ to be the full residual, but as Step 2a shows, uses a less greedy choice: it basically takes $u_0^{(t)}$ to be a convex combination of the residual $r^{(t-1)}$ and its previous value $u_0^{(t-1)}$, with higher weight on the latter. The subsequent parallel updates for $\theta_j^{(t)}$, $j = 1, \ldots, d$ are still given by univariate trend filtering fits, and though these steps do not exactly use partial residuals (since $u_0^{(t)}$ is not exactly the full residual), they are guaranteed to produce additive trend filtering solutions upon convergence (as per Theorem 4). An example of cyclic versus parallelized backfitting is given in Appendix A.1.18 in the Supplementary Material (Sadhanala and Tibshirani (2019)).

5. Experiments. Through empirical experiments, we examine the performance of additive trend filtering relative to additive smoothing splines. We also examine the efficacy of cross-validation for choosing the tuning parameter λ, as well as the use of multiple tuning parameters. All experiments were performed in R. For the univariate trend filtering solver, we used the `trendfilter` function in the `glmgen` package; for the univariate smoothing spline solver, we used the `smooth.spline` function in base R.

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

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

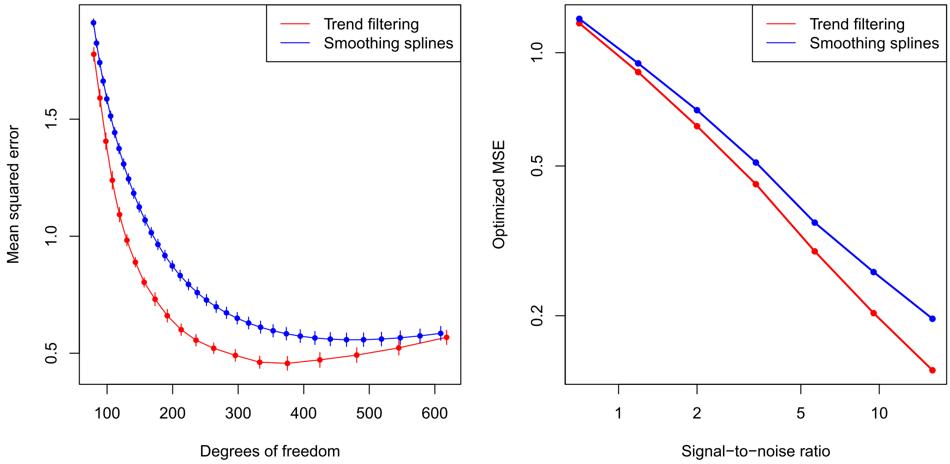

FIG. 4. *The left panel shows the MSE curves for additive trend filtering* (7) (*of quadratic order*) *and additive smoothing splines* (1) (*of cubic order*), *computed over* 10 *repetitions from the heterogeneous smoothness simulation with* $n = 2500$ *and* $d = 10$, *described in Section* 5.1, *where the SNR is set to* 4. *Vertical segments denote* ± 1 *standard deviations. The right panel displays the best-case MSE for each method* (*the minimum MSE over its regularization path*), *in a problem setup with* $n = 1000$ *and* $d = 6$, *as the signal-to-noise ratio* (*SNR*) *varies from* 0.7 *to* 16, *in equally spaced values on the log scale.*

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

The left panel of Figure 4 shows a comparison of the MSE curves from additive trend filtering in (7) (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, and does so at a less complex model (lower df).

The right panel of Figure 4 shows the best-case MSEs for additive trend filtering and additive smoothing splines (i.e., the minimum MSE over their regularization paths) as the noise level σ^2 is varied so that the SNR ranges from 0.7 to 1.6, in equally spaced values on the log scale. The results were again averaged over 10 repetitions of data drawn from a simulation setup essentially the same as the one described above, except that we considered a smaller problem size, with $n = 1000$

and $d = 6$. The plot reveals that additive trend filtering performs increasingly well (in comparison to additive smoothing splines) as the SNR grows—not surprising, as here it is able to better capture the heterogeneity in the component functions.

Lastly, in Appendix A.1.19 in the Supplementary Material (Sadhanala and Tibshirani (2019)), we present results from an experimental setup mimicking that in this subsection, except with the component functions f_{0j}, $j = 1, \dots, d$ having homogeneous smoothness throughout. Here, additive trend filtering and additive smoothing splines perform very similarly.

5.2. *Cross-validation and multiple tuning parameters.* Sticking to the simulation setup from the last subsection, but at the smaller problem size, $n = 1000$ and $d = 6$ (used to produce the right panel of Figure 4), we study in the left panel of Figure 5 the use of 5-fold cross-validation (CV) to select the tuning parameter λ for additive trend filtering and additive smoothing splines. Displayed are the resulting MSE curves as the SNR varies from 0.7 to 16. Also shown on the same plot are the oracle MSE curves (which are the same as those the right panel of Figure 4), in which λ has been chosen to minimize the MSE for each method. We can see that the performance of each method degrades using CV, but not by much.

In the right panel of the figure, we examine the use of multiple tuning parameters for additive smoothing splines and additive trend filtering, that is, replacing the

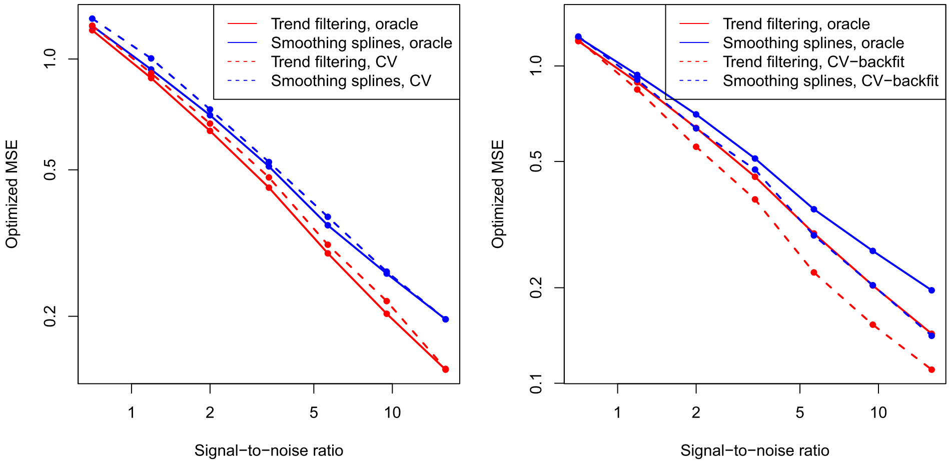

FIG. 5. *Both panels display results from the same simulation setup as that in the right panel of Figure 4. The left panel shows MSE curves when the estimators are tuned by 5-fold cross-validation (CV), and also by the oracle (reflecting the minimum possible MSE). The right panel displays MSE curves when we allow each estimator to have d tuning parameters, tuned by a hybrid backfit-CV method explained in the text, versus the oracle MSE curves for a single tuning parameter.*

penalties in (1) and (7) by

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

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

When d is moderate (even just for $d = 6$), cross-validation over a d-dimensional grid of values for λ_j, $j = 1, \dots, d$ can be prohibitive. However, as pointed out by a referee of this article, there has been a considerable amount of work dedicated to this problem by authors studying additive models built from splines (or other linear smoothers), for example, Fahrmeir and Lang (2001), Gu and Wahba (1991), Kim and Gu (2004), Rue, Martino and Chopin (2009), Ruppert, Wand and Carroll (2003), Wood (2000, 2004, 2011), Wood, Goude and Shaw (2015), Wood, Pya and Säfken (2016). Many of these papers use an efficient computational approach based on restricted maximum likelihood (REML) for selecting λ_j, $j = 1, \dots, d$; see also Wood (2017) for a nice introduction and description of this approach. Unfortunately, as far as we see it, REML does not easily apply to additive trend filtering.

We thus use the following simple approach for multiple tuning parameter selection: within each backfitting loop, for each component $j = 1, \dots, d$, we use (univariate) CV to choose λ_j. While this does not solve a particular convex optimization problem, and is not guaranteed to converge in general, we have found it to work quite well in practice. The right panel of Figure 5 compares the performance of this so-called backfit-CV tuning to the oracle, that chooses just a single tuning parameter. Both additive trend filtering and additive smoothing splines are seen to improve with d tuning parameters, tuned by backfit-CV, in comparison to the oracle choice of tuning parameter. Interestingly, we also see that additive smoothing splines with d tuning parameters performs on par with additive trend filtering with the oracle choice of tuning parameter. (In this example, REML tuning for additive smoothing splines—as implemented by the `mgcv` R package—performed worse than backfit-CV tuning, and so we only show results from the latter.)

6. Discussion. We have studied additive models built around the univariate trend filtering estimator, that is, defined by penalizing according to the sum of ℓ_1 norms of discrete derivatives of the component functions. We examined basic properties of these additive models, such as extrapolation of the fitted values to a d-dimensional surface, and uniqueness of the component fits. When the underlying regression function is additive, with components whose kth derivatives are of bounded variation, we derived error rates for kth order additive trend filtering: $n^{-(2k+2)/(2k+3)}$ for a fixed input dimension d (under weak assumptions), and $dn^{-(2k+2)/(2k+3)}$ for a growing dimension d (under stronger assumptions). We showed these rates are sharp by establishing matching minimax lower bounds. On the computational side, we devised a provably convergent parallel backfitting algorithm for additive trend filtering. It is worth noting that our parallel backfitting method is not specific to additive trend filtering, but it can be embedded in a more general parallel coordinate descent framework (Tibshirani (2017)).

A natural extension of our work is to consider the high-dimensional case, where d is comparable or possibly even much larger than n, and we fit a *sparse additive model* by employing an additional sparsity penalty in problem (7). Another natural extension is to consider responses $Y^i|X^i$, $i = 1, \dots, n$ from an exponential family distribution, and we fit a *generalized additive model* by changing the loss in (7). After we completed an initial version of this paper, both extensions have been pursued: Tan and Zhang (2017) develop a suite of error bounds for sparse additive models, with various forms of penalties (which include total variation on derivatives of components); and Haris, Simon and Shojaie (2018) give comprehensive theory for sparse generalized additive models, with various types of penalties (which again include total variation on derivatives of components).

Acknowledgments. We are very thankful to Garvesh Raskutti for his generous help and insight on various issues, and Martin Wainwright for generously sharing his unpublished book with us. We are also grateful to an anonymous referee whose thoughtful comments improved our paper.

SUPPLEMENTARY MATERIAL

Supplement to "Additive models with trend filtering" (DOI: 10.1214/19-AOS1833SUPP; .pdf). Proofs and additional simulations.

REFERENCES

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternative direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. MR0803258

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249

DE BOOR, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470. MR0845884

FAHRMEIR, L. and LANG, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *J. Roy. Statist. Soc. Ser. C* **50** 201–220. MR1833273

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892

GU, C. and WAHBA, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.* **12** 383–398. MR1087766

HARIS, A., SIMON, N. and SHOJAIE, A. (2018). Generalized sparse additive models. Available at https://arxiv.org/abs/1903.04641.

HASTIE, T. (1983). Non-parametric logistic regression. Technical report, Stanford Univ., Stanford, CA.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

KIM, Y.-J. and GU, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 337–356. MR2062380

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500

LOU, Y., BIEN, J., CARUANA, R. and GEHRKE, J. (2016). Sparse partially linear additive models. *J. Comput. Graph. Statist.* **25** 1026–1040. MR3572032

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

PETERSEN, A., WITTEN, D. and SIMON, N. (2016). Fused lasso additive model. *J. Comput. Graph. Statist.* **25** 1005–1025. MR3572026

RAMDAS, A. and TIBSHIRANI, R. J. (2016). Fast and flexible ADMM algorithms for trend filtering. *J. Comput. Graph. Statist.* **25** 839–858. MR3533641

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268. MR3363401

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

SADHANALA, V. and TIBSHIRANI, R. J (2019). Supplement to "Additive Models with Trend Filtering." DOI:10.1214/19-AOS1833SUPP.

SARDY, S. and TSENG, P. (2004). AMlet, RAMlet, and GAMlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *J. Comput. Graph. Statist.* **13** 283–309. MR2063986

STEIDL, G., DIDAS, S. and NEUMANN, J. (2006). Splines in higher order TV regularization. *Int. J. Comput. Vis.* **70** 214–255.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

TAN, Z. and ZHANG, C.-H. (2017). Penalized estimation in additive regression with high-dimensional data. Available at arXiv:1704.07229.

TIBSHIRANI, R. J. (1983). Non-parametric estimation of relative risk. Technical report, Stanford Univ., Stanford, CA.

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487

TIBSHIRANI, R. J. (2015). Degrees of freedom and model search. *Statist. Sinica* **25** 1265–1296. MR3410308

TIBSHIRANI, R. J. (2017). Dykstra's algorithm, ADMM, and coordinate descent: Connections, insights, and extensions. *Adv. Neural Inf. Process. Syst.* **30**.

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

TIBSHIRANI, R. J. and TAYLOR, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. MR2985948

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

VAN DE GEER, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.* **8** 543–574. MR3211024

VAN DER BURG, E. and DE LEEUW, J. (1983). Non-linear canonical correlation. *Br. J. Math. Stat. Psychol.* **36** 54–80.

WANG, Y.-X., SMOLA, A. and TIBSHIRANI, R. J. (2014). The falling factorial basis and its statistical applications. *Int. Conf. Mach. Learn.* **31**.

WOOD, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 413–428. MR1749600

WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99** 673–686. MR2090902

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC Press, Boca Raton, FL. MR3726911

WOOD, S. N., GOUDE, Y. and SHAW, S. (2015). Generalized additive models for large data sets. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 139–155. MR3293922

WOOD, S. N., PYA, N. and SÄFKEN, B. (2016). Smoothing parameter and model selection for general smooth models. *J. Amer. Statist. Assoc.* **111** 1548–1563. MR3601714

ZHANG, S. and WONG, M.-Y. (2003). Wavelet threshold estimation for additive regression models. *Ann. Statist.* **31** 152–173. MR1962502

MACHINE LEARNING DEPARTMENT
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVENUE
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: vsadhana@cs.cmu.edu

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVENUE
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: ryantibs@stat.cmu.edu