The Annals of Statistics
2020, Vol. 48, No. 1, 205–229
https://doi.org/10.1214/18-AOS1799

ADAPTIVE RISK BOUNDS IN UNIVARIATE TOTAL VARIATION DENOISING AND TREND FILTERING

BY ADITYANAND GUNTUBOYINA[1,*], DONOVAN LIEU[1,**], SABYASACHI CHATTERJEE[2] AND BODHISATTVA SEN[3]

[1]*Department of Statistics, University of California, Berkeley,* [*]*aditya@stat.berkeley.edu;* [**]*dlieu333@berkeley.edu*
[2]*Department of Statistics, University of Illinois at Urbana–Champaign, sc1706@illinois.edu*
[3]*Department of Statistics, Columbia University, bodhi@stat.columbia.edu*

We study trend filtering, a relatively recent method for univariate nonparametric regression. For a given integer $r \geq 1$, the rth order trend filtering estimator is defined as the minimizer of the sum of squared errors when we constrain (or penalize) the sum of the absolute rth order discrete derivatives of the fitted function at the design points. For $r = 1$, the estimator reduces to total variation regularization which has received much attention in the statistics and image processing literature. In this paper, we study the performance of the trend filtering estimator for every $r \geq 1$, both in the constrained and penalized forms. Our main results show that in the strong sparsity setting when the underlying function is a (discrete) spline with few "knots," the risk (under the global squared error loss) of the trend filtering estimator (with an appropriate choice of the tuning parameter) achieves the *parametric* n^{-1}-rate, up to a logarithmic (multiplicative) factor. Our results therefore provide support for the use of trend filtering, for every $r \geq 1$, in the strong sparsity setting.

1. Introduction. Consider the nonparametric regression problem where we observe data generated according to the model:

$$Y_i = f^*(i/n) + \xi_i, \quad i = 1, \ldots, n, \tag{1}$$

where $f^* : [0, 1] \to \mathbb{R}$ is the unknown regression function, and $\xi_1, \ldots, \xi_n$ are unobserved independent errors having the normal distribution with mean zero and variance σ^2. The goal is to recover the underlying function f^* from the measurements $Y_1, \ldots, Y_n$. Alternatively, in the Gaussian sequence formulation, (1) can be expressed as

$$Y = \theta^* + \xi, \tag{2}$$

where $\xi \sim N_n(0, \sigma^2 I_n)$, and $\theta^* := (f^*(1/n), f^*(2/n), \ldots, f^*(1))$ is unknown. Here $N_n(0, \sigma^2 I_n)$ denotes the multivariate normal distribution with mean vector zero and covariance matrix $\sigma^2 I_n$.

In this paper, we study the performance of *trend filtering*, a relatively new method for nonparametric regression with special emphasis on its risk properties. For a given integer $r \geq 1$, the rth order trend filtering estimator is defined as the minimizer of the sum of squared errors when we constrain or penalize the sum of the absolute rth order discrete derivatives of the fitted function at the design points. Formally, given a fixed integer $r \geq 1$ and a tuning parameter $V \geq 0$, the rth order trend filtering estimator for θ^* in the constrained form is given by

$$\hat{\theta}_V^{(r)} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Y - \theta\|^2 : \left\| D^{(r)} \theta \right\|_1 \leq V n^{1-r} \right\}, \tag{3}$$

Received February 2017; revised June 2018.
MSC2010 subject classifications. 62G08, 62J05, 62J07.
Key words and phrases. Adaptive splines, discrete splines, fat shattering, higher order total variation regularization, metric entropy bounds, nonparametric function estimation, risk bounds, subdifferential, tangent cone.

where $V > 0$ is a tuning parameter (the multiplicative factor n^{1-r} is just for normalization), $D^{(0)}\theta := \theta$, $D^{(1)}\theta := (\theta_2 - \theta_1, \ldots, \theta_n - \theta_{n-1})$ and $D^{(r)}\theta$, for $r \geq 2$, is recursively defined as $D^{(r)}\theta := D^{(1)}(D^{(r-1)}\theta)$. Also $\|\cdot\|_1$ denotes the usual L^1 norm defined by $\|x\|_1 := \sum_{i=1}^k |x_i|$ for $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$. Note that $\|D^{(r)}\theta\|_1$ also equals $V(D^{(r-1)}\theta)$ where $V(\alpha) := \sum_{i=2}^k |\alpha_i - \alpha_{i-1}|$ denotes the variation of a vector $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$. For simplicity, we denote the operator $D^{(1)}$ by simply D.

Alternatively, the trend filtering estimator in the penalized form is

$$\hat{\theta}_\lambda^{(r)} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^n} \left(\frac{1}{2}\|Y - \theta\|^2 + \sigma n^{r-1} \lambda \|D^{(r)}\theta\|_1 \right) \tag{4}$$

for $r \geq 1$ and tuning parameter $\lambda \geq 0$. There is an abuse of notation here in that we are using the same notation for both the constrained and the penalized estimators. It may be noted, however, that when the subscript of $\hat{\theta}^{(r)}$ is V, we are referring to the constrained estimator (3) while when the subscript is λ, we are referring to the penalized estimator (4).

For $r = 1$, (4) reduces to the one-dimensional discrete version of total variation regularization or total variation denoising which was first proposed by Rudin, Osher and Fatemi [33] and has since been heavily used in the image processing community. The penalized estimator (4), for general $r \geq 1$, was first proposed by Steidl, Didas and Neumann [34] in the image processing literature who termed it *higher order total variation regularization*. The same estimator was later rediscovered by Kim et al. [19] who coined the name *trend filtering* for it. Many properties of the estimator have been studied in Tibshirani [36] and Wang, Smola and Tibshirani [39]. It should also be mentioned here that a continuous version of (4), where the discrete differences are replaced by continuous derivatives, was proposed much earlier in the statistics literature by Mammen and van de Geer [24] under the name *locally adaptive regression splines*.

The presence of the L^1 norm in the constraint in (3) (resp., penalty in (4)) promotes sparsity of the vector $D^{(r)}\hat{\theta}_V^{(r)}$ (resp., $D^{(r)}\hat{\theta}_\lambda^{(r)}$). Now for every vector $\theta \in \mathbb{R}^n$, $\|D^{(r)}\theta\|_0 = k$ if and only if θ equals $(f(1/n), \ldots, f(n/n))$ for a *discrete spline* function f that is made of $k+1$ polynomials each of degree $(r-1)$ (here $\|x\|_0$ denotes the number of entries of the vector x that are nonzero). Discrete splines are piecewise polynomials with regularity at the knots. They differ from the usual (continuous) splines in the form of the regularity condition at the knots: for splines, the regularity condition translates to (higher order) derivatives of adjacent polynomials agreeing at the knots, while for discrete splines it translates to discrete differences of adjacent polynomials agreeing at the knots; see Mangasarian and Schumaker [25] for details. This fact about the connection between $\|D^{(r)}\theta\|_0$ and discrete splines is standard (see, e.g., Steidl, Didas and Neumann [34]) but we included a proof in Subsection D.3 of the supplementary file [15] for the convenience of the reader.

Thus the presence of the L^1 norm in (3) (resp., (4)) implies that $\hat{\theta}_V^{(r)}$ (resp., $\hat{\theta}_\lambda^{(r)}$) can be written as $(\hat{f}(1/n), \ldots, \hat{f}(n/n))$ for a discrete spline $\hat{f}$ of degree $(r-1)$ made up of not too many polynomial pieces. Trend filtering thus presents a way of fitting (discrete) splines to the data. Note that the knots of the discrete splines are automatically chosen by the optimization algorithms underlying (3) and (4) without any input from the user (except for the value of the tuning parameter V or λ). Because of this automatic selection of the knots, trend filtering can be regarded as a spatially adaptive method (in the terminology of Donoho and Johnstone [7]). Note that such spatial adaptation is not exhibited by classical nonparametric regression methods such as local polynomials, kernels and splines, with a fixed tuning parameter. On the other hand, methods such as CART (Breiman et al. [3]), MARS (Friedman [11]), variable-bandwidth kernel/spline methods (see, e.g., Müller and Stadtmüller [26], Brockmann, Gasser and Herrmann [4], Pintore, Speckman and Holmes [29] and Zhou and Shen [41]) and wavelets (Donoho and Johnstone [7]) are also spatially adaptive.

The present paper studies the performance of the estimators $\hat{\theta}_V^{(r)}$ and $\hat{\theta}_\lambda^{(r)}$ as estimators of θ^* under the multivariate Gaussian model (2). We shall use the squared error loss under which the risk of an estimator $\hat{\theta}$ is defined as

$$R(\hat{\theta}, \theta^*) := \frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta} - \theta^*\|^2. \tag{5}$$

Under natural sparsity assumptions on θ^*, we provide upper bounds on the risks $R(\hat{\theta}_V^{(r)}, \theta^*)$ and $R(\hat{\theta}_\lambda^{(r)}, \theta^*)$ as well as high probability upper bounds on the random loss functions $\|\hat{\theta}_V^{(r)} - \theta^*\|^2/n$ and $\|\hat{\theta}_\lambda^{(r)} - \theta^*\|^2/n$.

It is natural to study the risk properties of (3) and (4) under the following two kinds of assumptions on θ^*: (a) $n^{r-1}\|D^{(r)}\theta^*\|_1 \le V$ for some $V > 0$ (possibly dependent on n), and (b) $\|D^{(r)}\theta^*\|_0 \le k$ for some k that is much smaller than n. We shall refer to these two regimes as *weak sparsity* and *strong sparsity*, respectively. This breakdown into weak and strong sparsity settings is inspired by corresponding terminology in the study of risk properties of thresholding based estimators in Gaussian sequence models [18] and the prediction risk properties of the LASSO estimators in regression [5]. Indeed, as demonstrated in Tibshirani [36], there is a close connection between the trend filtering estimators and LASSO (more details are provided in Section 5.4).

A thorough study on the performance of the penalized trend filtering estimator (4) under weak sparsity has been done by Tibshirani [36] and Wang, Smola and Tibshirani [39] building on earlier results of Mammen and van de Geer [24]. It is proved there that, when the tuning parameter λ is appropriately chosen, the penalized estimator (4) is minimax optimal in the weak sparsity setting. Actually, the weak sparsity results of [36, 39] are broader and hold under more general settings (see Remark 2.1 for more details).

The present paper focuses on the strong sparsity setting. Compared to available results in the weak sparsity setting, relatively little is known about the performance of the trend filtering estimators in the strong sparsity setting. In fact, all existing results [6, 17, 22, 23, 27, 37] for strong sparsity deal with the case $r = 1$ (where trend filtering is the same as total variation denoising). To the best of our knowledge, the present paper is the first to prove risk bounds for trend filtering under strong sparsity for arbitrary $r \ge 1$. We also improve, in certain aspects, existing results for $r = 1$.

In order to motivate our results, let us consider the strong sparsity setting where it is assumed that $D^{(r)}\theta^*$ is sparse. If $\|D^{(r)}\theta^*\|_0 = k$, then, as mentioned previously, $\theta^* = (f(1/n), \ldots, f((n-1)/n), f(1))$ for a discrete spline function f that is made of $k+1$ polynomials each of degree $(r-1)$. Given data $Y \sim N_n(\theta^*, \sigma^2 I_n)$, an oracle piecewise polynomial estimator (having access to locations of the knots of θ^*) would put knots corresponding to θ^* and then fit a polynomial of degree $(r-1)$ in each of the partitions given by the knots. This would be a linear estimator with at most $(k+1)r$ degrees of freedom and its risk (defined as in (5)) will be bounded by $r\sigma^2(k+1)/n$. This motivates the following question which is the focus of this paper: When $\|D^{(r)}\theta^*\|_0 = k$, how do the risks of properly tuned trend filtering estimators (3) and (4) compare with the oracle risk of $r\sigma^2(k+1)/n$?

The main results of this paper for constrained trend filtering (Theorem 2.2 and Corollary 2.3) imply that when $\|D^{(r)}\theta^*\|_0 = k$, the risk of $\hat{\theta}_V^{(r)}$ satisfies

$$R(\hat{\theta}_V^{(r)}, \theta^*) \le C_r(c)\sigma^2\frac{k+1}{n}\log\frac{en}{k+1}, \tag{6}$$

provided:

(i) the tuning parameter V is nonrandom and close to $V^* := n^{r-1}\|D^{(r)}\theta^*\|_1$, and

(ii) (minimum length condition) each of the polynomial pieces of θ^* have length bounded below by $cn/(k+1)$ for a constant $c > 0$ (in fact, our result requires a weaker version of this condition; see (13) and Remark 2.4).

Here $C_r(c)$ is a positive constant that depends only on r and the constant c from the second assumption above.

We also prove results for the penalized estimators. For $r = 1$, our main result (Corollary 2.8) states that the risk of $\hat{\theta}^{(1)}_\lambda$ is also bounded by the right-hand side of (6) under the minimum length condition provided λ is close to a theoretical choice λ^* and $\lambda \geq \lambda^*$. This choice λ^* depends on θ^* and is defined in (27). We provide an explicit upper bound for λ^* in Lemma 2.9 which gives risk bounds for $\hat{\theta}^{(1)}_\lambda$ under more explicit choices of λ (see Corollary 2.10). A comparison of these results to existing results is given in Remarks 2.6 and 2.7.

For $r \geq 2$, we prove, in Corollary 2.11, that the penalized estimator satisfies

$$R(\hat{\theta}^{(r)}_\lambda, \theta^*) \leq C_r(c)\sigma^2\left(\frac{k+1}{n}\log\frac{en}{k+1} + \frac{(k+1)^{2r}}{n}\right) \tag{7}$$

under the minimum length condition provided that λ is close to λ^* (defined in (27)) and $\lambda \geq \lambda^*$. Explicit upper bounds for λ^* are in Lemma 2.12 and risk bounds for $\hat{\theta}^{(r)}_\lambda$ with explicit penalty choices are in Corollary 2.13. Note that (7) is weaker compared to (6) in terms of the dependence on k.

The implication of our results is the following. As mentioned earlier, the trend filtering estimators are given by discrete spline functions of degree $r-1$. The knots of these splines are chosen automatically by the algorithm (the user only needs to specify the tuning parameter V or λ). Our results indicate that under the assumption $\|D^{(r)}\theta^*\|_0 = k$ (i.e., θ^* is a discrete spline of degree $r-1$ with $k+1$ polynomial pieces) with a minimum length condition on the polynomial pieces of θ^*, the automatic selection of knots by the trend filtering estimators (when appropriate choices of V or λ) happens in a way that the overall risk is comparable to the oracle risk of $r\sigma^2(k+1)/n$. In fact, when $k = O(1)$, the risks of the ideally tuned trend filtering estimators is only off compared to the oracle risk by a factor that is logarithmic in n (we also prove in Lemma 2.4 that this logarithmic factor cannot be completely removed in general). The automatic knot selection of trend filtering can therefore be interpreted as being done *adaptively* depending on the structure of the unknown θ^* in order to approximate the oracle risk. This is the reason why we refer to our results as adaptive risk bounds. It should be mentioned here that a similar adaptation story can also be used to describe the weak sparsity results [36, 39] where the knots are adaptively chosen to attain the minimax rate under the L^1 constraint on $D^{(r)}\theta^*$. Therefore, our results (together with those of [36, 39]) provide support for the use of the trend filtering estimators in both weak and strong sparsity settings.

We would like to mention here that theoretical analysis of spatially adaptive nonparametric regression methods under strong sparsity is nontrivial. Indeed, among various such methods including CART, MARS, variable-bandwidth kernel/spline methods and wavelets, rigorous theoretical risk results under strong sparsity only exist for wavelets [7] and variable-bandwidth kernel methods [13, 21]. The analysis of trend filtering estimators is more involved compared to estimators based on wavelets and variable-bandwidth kernels because the trend filtering estimators are given by the output of an optimization algorithm and have no closed form expressions.

The rest of this paper is organized as follows. Our main results are described in Section 2: Section 2.1 deals with the constrained estimator where we provide risk bounds under both weak sparsity (which was not known previously) and strong sparsity. Section 2.2 deals with the penalized estimator and here we separate our presentation into two parts: results for $r = 1$ and results for $r \geq 2$; our results for $r \geq 2$ are weaker (there is an additional $(k+1)^{2r}/n$ term

in the risk) than the results for $r = 1$. Throughout, we focus on nonasymptotic upper bounds for the risk (expected loss) although all our results can be converted into high probability upper bounds on the loss (see Remark 2.3). Because of space constraints, all proofs are given in the supplementary file [15]. However, a high level overview of the proofs is provided in Section 3. Section 4 contains some simulation studies supporting some of our theoretical results. Finally, several interesting issues related to our results are described in Section 5.

2. Main results. Throughout C_r will denote a positive constant that depends on r alone although its precise value will change from equation to equation. We shall assume that $n \geq 2r$ throughout the paper (many of our results also hold under the weaker condition $n \geq r + 1$).

2.1. *Results for the constrained estimator.* We start with the bound of $n^{-2r/(2r+1)}$ for risk of $\hat{\theta}_V^{(r)}$ under the condition that the tuning parameter V satisfies $\|D^{(r)}\theta^*\|_1 \leq Vn^{1-r}$. This result is similar to results in Mammen and van de Geer [24], Tibshirani [36] and Wang, Smola and Tibshirani [39] who focussed on the penalized estimator (4) (see Remark 2.1 for details). We also explicitly state the dependence of the bound on V and σ.

THEOREM 2.1. *Fix $r \geq 1$. Suppose that the tuning parameter V is chosen so that $n^{r-1}\|D^{(r)}\theta^*\|_1 \leq V$. Then there exists a positive constant C_r depending on r alone such that*

$$R(\hat{\theta}_V^{(r)}, \theta^*) \leq C_r \max\left(\left(\frac{\sigma^2 V^{1/r}}{n}\right)^{2r/(2r+1)}, \frac{\sigma^2}{n}\log(en)\right). \tag{8}$$

Also for every $x > 0$, we have

$$\frac{1}{n}\|\hat{\theta}_V^{(r)} - \theta^*\|^2 \leq C_r \max\left(\left(\frac{\sigma^2 V^{1/r}}{n}\right)^{2r/(2r+1)}, \frac{\sigma^2}{n}\log(en)\right) + \frac{4\sigma^2 x}{n} \tag{9}$$

with probability at least $1 - e^{-x}$.

REMARK 2.1. As mentioned earlier, bounds similar to (8) and (9) have been proved in Mammen and van de Geer [24], Tibshirani [36] and Wang, Smola and Tibshirani [39] for the penalized trend filtering estimator. Actually, the bounds in these earlier papers hold under more general assumptions than the assumptions of the current paper. For example, their analyses also hold under the assumption that the (continuous) variation norm of the function $(f^*)^{(r-1)}$ (this is the $(r-1)$th derivative of f^*) is at most V, where f^* is the true function with $\theta^* = (f^*(1/n), \ldots, f^*(1))$. Note that there is subtle difference between this and our assumption of an upper bound on $\|D^{(r)}\theta^*\|_1$ in the sequence model (2). An assumption on the variation norm of $(f^*)^{(r-1)}$ does not directly lead to a bound on $\|D^{(r)}\theta^*\|_1$ which makes the analysis difficult (see Wang, Smola and Tibshirani [39] for more details on the relation between the two variation norms). Also, the results in these earlier papers studied the general setting with $\theta^* := (f^*(x_1), \ldots, f^*(x_n))$ where $x_1, \ldots, x_n$ are design points that are not necessarily equally spaced. We restrict ourselves to the equally spaced design setting in this paper (see Section 5.1).

REMARK 2.2. $n^{-2r/(2r+1)}$ is the minimax rate of estimation over the class of $\theta \in \mathbb{R}^n$ with $\|D^{(r)}\theta\|_1 \leq Vn^{1-r}$ (see, e.g., Donoho and Johnstone [8]). This means that the constrained trend filtering estimator with tuning parameter V is minimax optimal over $\{\theta \in \mathbb{R}^n : \|D^{(r)}\theta\|_1 \leq Vn^{1-r}\}$. This result was known previously for the penalized estimator; see Tibshirani [36]. Note also that V here can change with n as well and inequality (8) implies that $\hat{\theta}_V^{(r)}$ is minimax optimal even in terms of the dependence of the rate on V.

Before we state results for strong sparsity, we need some notation. Fix an integer $r \geq 1$ and let $n \geq r+1$. For a vector $\theta \in \mathbb{R}^n$ and an index $2 \leq j \leq n-r+1$, we say that j is an rth *order knot* (or *knot of order* r) of θ provided $(D^{(r-1)}\theta)_{j-1} \neq (D^{(r-1)}\theta)_j$. Note that first-order knots are just jumps and second-order knots are points of change of slope. We also say that an rth order knot j has *sign* $+1$ if $(D^{(r-1)}\theta)_{j-1} < (D^{(r-1)}\theta)_j$ and *sign* -1 if $(D^{(r-1)}\theta)_{j-1} > (D^{(r-1)}\theta)_j$. For $\theta \in \mathbb{R}^n$, we let

$$\mathbf{k}_r(\theta) := \left\|D^{(r)}\theta\right\|_0 \quad \text{and} \quad V^{(r)}(\theta) := n^{r-1}\left\|D^{(r)}\theta\right\|_1. \tag{10}$$

When $r=1$, note that $V^{(1)}(\theta) = \|D\theta\|_1 = |\theta_2 - \theta_1| + \cdots + |\theta_n - \theta_{n-1}|$ which is simply the variation of θ. We therefore simply denote $V^{(1)}(\theta)$ by $V(\theta)$. It also follows then that $V^{(r)}(\theta) = n^{r-1}V(D^{(r-1)}\theta)$.

It may be observed that $\mathbf{k}_r(\theta)$ equals precisely the number of rth order knots of θ. When the value of r and $\theta \in \mathbb{R}^n$ are clear from the context, we simply denote $\mathbf{k}_r(\theta)$ by k. Also, note that as $D^{(r)}\theta$ is a vector of length $n-r$, we necessarily have $\mathbf{k}_r(\theta) = \|D^{(r)}\theta\|_0 \leq n-r \leq n-1$.

Suppose $\mathbf{k}_r(\theta) = k$ and let $2 \leq j_1 < \cdots < j_k \leq n-r+1$ denote all the rth order knots of θ with associated signs $\mathfrak{r}_1, \ldots, \mathfrak{r}_k \in \{-1, 1\}$. Also let $\mathfrak{r}_0 = \mathfrak{r}_{k+1} = 0$. Further, let $n_0 := j_1 + r - 2$, $n_i := j_{i+1} - j_i$, for $1 \leq i \leq k-1$, and $n_k := n - r + 2 - j_k$, and observe that $\sum_{i=0}^{k} n_i = n$. Finally, let

$$n_{i*} := \min\left(n_i, \frac{n}{k+1}\right) \quad \text{for } i = 0, 1, \ldots, k.$$

We now define two quantities $\delta_r(\theta)$ and $\Delta_r(\theta)$ in the following way:

$$\delta_r(\theta) := \left(n_{0*}^{1-2r} + n_{k*}^{1-2r} + \sum_{i=1}^{k-1} n_{i*}^{1-2r} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}\right)^{1/2} \tag{11}$$

and

$$\Delta_r(\theta) := \frac{k+1}{n}\log\frac{en}{k+1} + \frac{\delta_r^2(\theta)}{n}\left(\frac{n}{k+1}\right)^{2r-1}\log\frac{en}{k+1} + \left(\frac{\delta_r(\theta)}{\sqrt{n}}\right)^{1/r}, \tag{12}$$

where, in the definition of $\delta_r(\theta)$, the quantity $I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}$ denotes the indicator variable that equals 1 if $\mathfrak{r}_i \neq \mathfrak{r}_{i+1}$ and 0 if $\mathfrak{r}_i = \mathfrak{r}_{i+1}$. Note that trivially $\Delta_r(\theta) \geq (k+1)/n \geq 1/n$.

Our results will show that the risk of the estimator $\hat{\theta}_V^{(r)}$ for θ^* will essentially be controlled by $\Delta_r(\theta^*)$. The key point to note about $\Delta_r(\theta)$ is the fact (easy to check) that when

$$\min_{0 \leq i \leq k : \mathfrak{r}_i \neq \mathfrak{r}_{i+1}} n_i \geq \frac{cn}{k+1} \tag{13}$$

for a positive constant $c \leq 1$ (here $\mathfrak{r}_1, \ldots, \mathfrak{r}_k \in \{-1, 1\}$ are the signs of the rth order knots of θ while $\mathfrak{r}_0$ and $\mathfrak{r}_{k+1}$ are taken to be zero), then

$$\delta_r^2(\theta) \leq \left(\frac{cn}{k+1}\right)^{1-2r}(k+1)$$

and consequently

$$\begin{aligned}\Delta_r(\theta) &\leq \left\{1 + c^{1-2r}\right\}\frac{k+1}{n}\log\frac{en}{k+1} + c^{(1-2r)/(2r)}\frac{k+1}{n} \\ &\leq \left\{1 + c^{1-2r} + c^{(1-2r)/(2r)}\right\}\frac{k+1}{n}\log\frac{en}{k+1}. \end{aligned} \tag{14}$$

We say that θ satisfies the *minimum length condition* with constant c if condition (13) holds. We have just observed that when θ satisfies the minimum length condition with constant c then $\Delta_r(\theta) \leq C_r(c)\frac{k+1}{n}\log\frac{en}{k+1}$ for a constant $C_r(c)$ depending only on c and r.

The following is our main result for the constrained trend filtering estimator.

THEOREM 2.2. *Fix $r \geq 1$ and $n \geq 2r$. Consider the estimator $\hat{\theta}_V^{(r)}$ defined in (3) with tuning parameter $V \geq 0$. Then for every $\theta^* \in \mathbb{R}^n$, we have*

$$R\big(\hat{\theta}_V^{(r)}, \theta^*\big) \leq \inf_{\theta \in \mathbb{R}^n : V^{(r)}(\theta) = V} \left(\frac{1}{n} \big\|\theta^* - \theta\big\|^2 + C_r \sigma^2 \Delta_r(\theta) \right) \tag{15}$$

for a positive constant C_r, depending only on r.

REMARK 2.3 (High-probability bound). Note that Theorem 2.2 gives an upper bound for $R(\hat{\theta}_V^{(r)}, \theta^*)$ which is the expectation of $\frac{1}{n}\|\hat{\theta}_V^{(r)} - \theta^*\|^2$. Similarly, as in Theorem 2.1, the risk bound (15) can be supplemented by the following high probability bound: for every $x > 0$, we have

$$\frac{1}{n}\big\|\hat{\theta}_V^{(r)} - \theta^*\big\|^2 \leq \inf_{\theta \in \mathbb{R}^n : V^{(r)}(\theta) = V} \left(\frac{1}{n} \big\|\theta^* - \theta\big\|^2 + C_r \sigma^2 \Delta_r(\theta) \right) + \frac{4\sigma^2 x}{n} \tag{16}$$

with probability at least $1 - e^{-x}$. This will be true in all the results of this paper (namely that the bound on $R(\hat{\theta}, \theta^*)$ plus $4\sigma^2 x/n$ will dominate $\frac{1}{n}\|\hat{\theta} - \theta^*\|^2$ with probability at least $1 - e^{-x}$). Thus, for ease of presentation, we shall omit high probability statements and only report risk results (i.e., bounds on $R(\hat{\theta}, \theta^*)$) in the rest of the paper.

Theorem 2.2 applies to every $\theta^* \in \mathbb{R}^n$ and is stated in the sharp oracle form. It implies that the risk of $\hat{\theta}_V^{(r)}$ is small provided there exists some $\theta \in \mathbb{R}^n$ with $V^{(r)}(\theta) = V$ such that (a) $\|\theta - \theta^*\|$ is small, and (b) $\Delta_r(\theta)$ is small.

Theorem 2.2 yields the following corollary which is a nonoracle inequality and is more readily interpretable. Recall from (14) that $\Delta_r(\theta)$ is bounded from above by a constant multiple of $\frac{k+1}{n} \log \frac{en}{k+1}$ with $\mathbf{k}_r(\theta) = k$ provided θ satisfies (13).

COROLLARY 2.3. *Consider the estimator $\hat{\theta}_V^{(r)}$ with tuning parameter V. Suppose θ^* satisfies the minimum length condition (13) with constant c, then*

$$R\big(\hat{\theta}_V^{(r)}, \theta^*\big) \leq \big(V - V^{(r)}(\theta^*)\big)^2 + C_r(c) \frac{\sigma^2(\mathbf{k}_r(\theta^*) + 1)}{n} \log \frac{en}{\mathbf{k}_r(\theta^*) + 1}, \tag{17}$$

where $C_r(c)$ is a positive constant that depends on r and c alone. Further, if V is chosen so that

$$\big(V - V^{(r)}(\theta^*)\big)^2 \leq C \frac{\sigma^2(\mathbf{k}_r(\theta^*) + 1)}{n} \log \frac{en}{\mathbf{k}_r(\theta^*) + 1}$$

for a positive constant C, then we have

$$R\big(\hat{\theta}_V^{(r)}, \theta^*\big) \leq C_r(c, C) \frac{\sigma^2(\mathbf{k}_r(\theta^*) + 1)}{n} \log \frac{en}{\mathbf{k}_r(\theta^*) + 1} \tag{18}$$

for a positive constant $C_r(c, C)$ that depends on r, c and C alone.

Note that Theorem 2.2 and Corollary 2.3 both apply to every $r \geq 1$. On the other hand, existing adaptation results for trend filtering all deal with the case $r = 1$ (which corresponds to total variation regularization). Even for $r = 1$, our results are stronger, in some respects, compared to the existing results in the literature (see Remark 2.6 for a precise comparison).

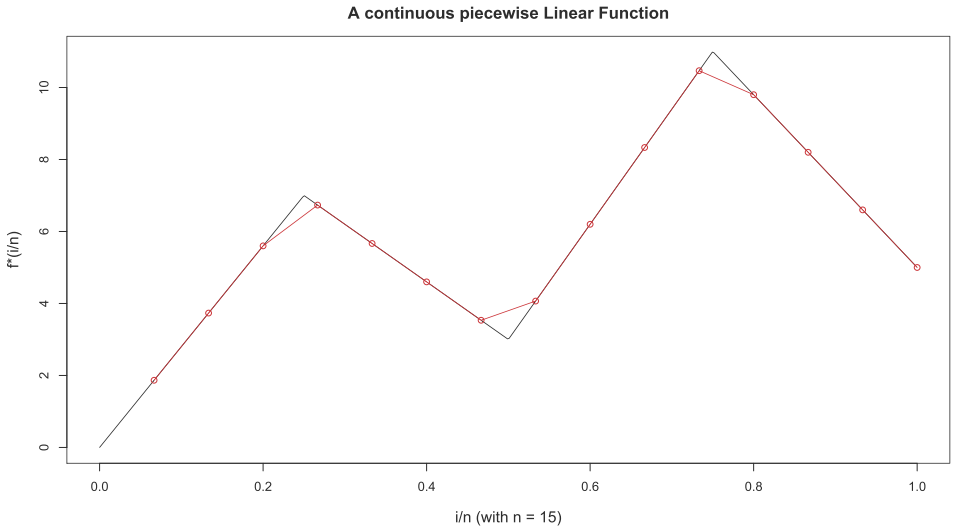

FIG. 1. *A piecewise linear function f^* on $[0,1]$ together with the vector $\theta^* := (f^*(1/n), \ldots, f^*(1))$ for $n = 15$ plotted in red. Note that f^* has three knots while θ^* has six first-order knots.*

REMARK 2.4 (On the minimum length condition). The minimum length condition (13) required for Corollary 2.3 is weaker than existing minimum length conditions in the literature (this comparison is only for $r = 1$ because no results exist for $r \geq 2$) which are all of the form

$$\min_{0 \leq i \leq k} n_i \geq \frac{cn}{k+1} \quad \text{where } k = \mathbf{k}_1(\theta^*). \tag{19}$$

Indeed our condition (13) requires that $n_i \geq cn/(k+1)$ be true only for those i for which $\mathfrak{r}_i \neq \mathfrak{r}_{i+1}$ while (19) requires this for all i. To see why our condition can be substantially weaker, consider, for example, the situation when $D^{(r-1)}\theta^*$ is a monotonic vector (for $r = 1$, this means that θ^* is itself monotone while for $r = 2$, this means that θ^* is convex/concave). In this case, condition (13) is equivalent to requiring that $n_i \geq cn/(k+1)$ only for $i = 0$ and $i = k$ which is much weaker than requiring it for all $0 \leq i \leq k$.

The fact that our minimum length condition involves only those i for which $\mathfrak{r}_i \neq \mathfrak{r}_{i+1}$ as opposed to involving all $i \in \{0, 1, \ldots, k\}$ is especially crucial for $r \geq 2$. To see this, consider the piecewise linear function f^* on $[0,1]$ shown in Figure 1. This function clearly has three knots (points of change of slope) in $(0,1)$. However, the vector θ^* obtained as $(f^*(1/n), \ldots, f^*(n/n))$ (with $n = 15$) has six second-order knots. The reason for the additional knots is due to the fact that the original knots of f^* are not at the design points $1/n, \ldots, n/n$. Note however that because of these additional knots, the minimum length condition will not be satisfied over all $i = 0, 1, \ldots, k$. On the other hand, it should be clear that (13) will still be satisfied because the additional linear pieces satisfy the property that $\mathfrak{r}_i = \mathfrak{r}_{i+1}$.

REMARK 2.5 (The minimum length condition cannot be removed). We shall argue here via simulations that the minimum length condition in Corollary 2.3 cannot be removed. Suppose that θ^* is given by

$$\theta_1^* = \cdots = \theta_{n-1}^* = 0 \quad \text{and} \quad \theta_n^* = 5 \tag{20}$$

and consider estimating θ^* from an observation $Y \sim N_n(\theta^*, I_n)$ (i.e., $\sigma = 1$) by $\hat{\theta}_V^{(1)}$ (i.e., $r = 1$) with tuning parameter $V = V^{(1)}(\theta^*) = 5$. It is clear here that $\mathbf{k}_1(\theta^*) = 1$. The minimum length condition (13) is not satisfied because $n_0 = n - 1$ and $n_1 = 1$. The risk $R(\hat{\theta}_V^{(1)}, \theta^*)$

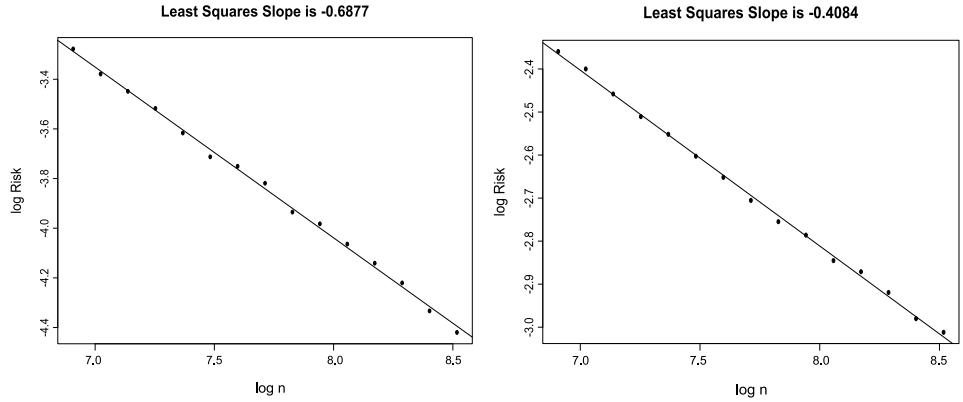

FIG. 2. Left: *plot of* $\log R(\hat{\theta}_V^{(1)}, \theta^*)$ *against* $\log n$ *for* θ^* *as in* (20). *The least squares slope is close to* $-2/3$ *which suggests that the risk decays as* $n^{-2/3}$ *instead of the faster rate given by Corollary* 2.3. Right: *plot of* $\log R(\hat{\theta}_V^{(2)}, \theta^*)$ *against* $\log n$ *for* θ^* *defined in* (21). *The slope is close to* $-2/5$ *which suggests that the risk decays as* $n^{-2/5}$ *instead of the faster rate given by Corollary* 2.3.

can be computed via simulation. In Figure 2 (left panel), we have plotted $\log R(\hat{\theta}_V^{(1)}, \theta^*)$ against $\log n$ for values of n between 1000 and 5000 (chosen to be equally spaced on the log-scale). For each value of n, we calculated the risk using 100 Monte Carlo replications. The slope of the least squares line through these points turned out to be close to $-2/3$ which indicates that the risk $R(\hat{\theta}_V^{(1)}, \theta^*)$ decays at the rate $n^{-2/3}$. This rate is slower than the rate given by Corollary 2.3 indicating that inequality (17) is not true for this θ^*. On the other hand, the $n^{-2/3}$ rate here makes sense in light of Theorem 2.1. Therefore, even though the vector $D\theta^*$ is sparse (with $\|D\theta^*\|_0 = 1$), the rate of convergence of $\hat{\theta}^{(1)}$ is equal to the $n^{-2/3}$ and not the faster rate given by Corollary 2.3. This points to the necessity of the minimum length condition (13).

Another counterexample for the necessity of (13) for Corollary 2.3 is

$$\theta_1^* = \cdots = \theta_{\lfloor n/2 \rfloor}^* = 0 \quad \text{and} \quad \theta_{\lfloor n/2 \rfloor + 1}^* = \theta_{\lfloor n/2 \rfloor + 2}^* = \cdots = \theta_n^* = 5. \tag{21}$$

Here consider the problem of estimating θ^* by the estimator $\hat{\theta}_V^{(2)}$ (i.e., $r = 2$) with tuning parameter $V = V^{(2)}(\theta^*) = 10n$. It is clear that $\mathbf{k}_2(\theta^*) = 2$, $n_0 = \lfloor n/2 \rfloor$, $n_1 = 1$ and $n_2 = n - \lfloor n/2 \rfloor - 1$. The minimum length condition (13) is not satisfied as n_1 is too small. The risk $\log R(\hat{\theta}_V^{(2)}, \theta^*)$ is plotted against $\log n$ in the right panel of Figure 2 (the values of n are chosen as before). The slope of the least squares line here is close to $-2/5$ which suggests that the risk decays slowly than what is given by Corollary 2.3. Note that $n^{-2/5}$ is exactly the rate given by Theorem 2.1 (take $r = 2$ and $V = 10n$ in (8)).

It is natural to ask if the bound given by inequality (18) can be improved further by dropping the $\log \frac{en}{\mathbf{k}_r(\theta^*)+1}$ term. The following simple result shows that this cannot be done in general.

LEMMA 2.4. *Suppose* $\theta^* := (0, \ldots, 0, 1, \ldots, 1)$ *with jump at* $j = \lceil n/2 \rceil$. *Let* $\hat{\theta}_{V=1}^{(1)}$ *denote the estimator* (3) *with* $V = 1$. *Then*

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R\big(\hat{\theta}_{V=1}^{(1)}, \theta^*\big) \geq \frac{\log(n/2)}{2n}.$$

2.2. *Results for the penalized estimator.* In this section, we present risk results for the penalized estimator defined in (4). An important role in these results will be played by the subdifferential of the convex function $f(\theta) := \|D^{(r)}\theta\|_1$ at the true parameter value θ^*. Recall that the subdifferential of a convex function $g : \mathbb{R}^n \to \mathbb{R}$ at a point $\theta \in \mathbb{R}^n$ is the set consisting of all subgradients of g at θ and will be denoted by $\partial g(\theta)$. For every finite convex function g on $\mathbb{R}^n$ and $\theta \in \mathbb{R}^n$, the subdifferential $\partial g(\theta)$ is nonempty, closed, convex and bounded (see, e.g., Rockafellar [31], page 218).

The following is the reason why $\partial f(\theta^*)$ (for $f(\theta) := \|D^{(r)}\theta\|_1$) plays a key role in understanding the risk of (4). It has been proved by Oymak and Hassibi [28], Theorem 2.2, that for a general penalized estimator:

$$\hat{\theta}^g_\lambda := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^n}\Big(\frac{1}{2}\|Y-\theta\|^2 + \sigma\lambda g(\theta)\Big),$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is convex, its risk under the model $Y \sim N_n(\theta^*, \sigma^2 I_n)$ satisfies:

$$R\big(\hat{\theta}^g_\lambda, \theta^*\big) \le \frac{\sigma^2}{n}\mathbb{E}\Big(\inf_{v\in\lambda\partial g(\theta^*)}\|Z-v\|^2\Big), \tag{22}$$

where $\lambda\partial g(\theta^*) := \{\lambda v : v \in \partial g(\theta^*)\}$ and the expectation on the right-hand side is with respect to the standard Gaussian vector $Z \sim N_n(0, I_n)$. Moreover, inequality (22) cannot in general be improved, because, as proved in [28], Proposition 4.2, it is tight in the low σ limit, that is, the limit (as $\sigma \to 0$) of the left-hand side of (22) scaled by σ^2/n equals the expectation on the right-hand side of (22). Inequality (22) will be our main technical tool for studying the risk of (4), and thus it will be important to understand the subdifferentials of the function $\theta \mapsto \|D^{(r)}\theta\|_1$.

The next result (proved in Subsection C.4 of the supplementary material [15]) characterizes the subdifferential of $f(\theta) := \|D^{(r)}\theta\|_1$.

PROPOSITION 2.5. *Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\alpha) := \|D^{(r)}\alpha\|_1$. Fix $\theta \in \mathbb{R}^n$. Then a vector $v \in \mathbb{R}^n$ belongs to the subdifferential $\partial f(\theta)$ if and only if the following two conditions hold*:

$$\sum_{i=j}^{n}\binom{r+i-j-1}{r-1}v_i = 0 \quad \text{for } 1 \le j \le r, \tag{23}$$

and

$$\sum_{i=j}^{n}\binom{r+i-j-1}{r-1}v_i = \begin{cases}\operatorname{sgn}\big((D^{(r)}\theta)_{j-r}\big) & \text{if } (D^{(r)}\theta)_{j-r} \neq 0,\\ \in [-1,1] & \text{otherwise}\end{cases} \tag{24}$$

for $r < j \le n$. Here $\operatorname{sgn}(x)$ *denotes the sign of x for $x \neq 0$.*

It should be clear from the above proposition that $\partial f(\theta^*)$ is always a convex polyhedron and is of a different nature when $D^{(r)}\theta^* \neq 0$ as opposed to when $D^{(r)}\theta^* = 0$. For example, when $D^{(r)}\theta^* = 0$, the zero vector belongs to $\partial f(\theta^*)$ and moreover, the sets $\lambda\partial f(\theta^*) := \{\lambda v : v \in \partial f(\theta^*)\}$ are increasing as λ increases. Both these facts are not true when $D^{(r)}\theta^* \neq 0$. We thus separate our risk results into the two cases: $D^{(r)}\theta^* \neq 0$ and $D^{(r)}\theta^* = 0$. First we deal with the case $D^{(r)}\theta^* \neq 0$. The other (simpler) case is in Lemma 2.14.

Assume therefore that $D^{(r)}\theta^* \neq 0$. The following quantities (all defined in terms of the subdifferential $\partial f(\theta^*)$) will play a key role in our risk bounds for the penalized estimator (4). Let

$$v^* := \operatorname*{arg\,min}_{v\in\partial f(\theta^*)}\|v\| \quad \text{and} \quad v_0 := \operatorname*{arg\,min}_{v\in\operatorname{aff}(\partial f(\theta^*))}\|v\|, \tag{25}$$

where $\operatorname{aff}(\partial f(\theta^*))$ denotes the affine hull of $\partial f(\theta^*)$ (recall that for a subset $S \subseteq \mathbb{R}^n$, its affine hull $\operatorname{aff}(S)$ consists of all vectors $w_1x_1 + \cdots + w_mx_m$ such that $m \geq 1$, $x_i \in S$ and $w_1 + \cdots + w_m = 1$). Note that v^* and v_0 are uniquely defined because they are simply the projections of the zero vector onto the closed convex sets $\partial f(\theta^*)$ and $\operatorname{aff}(\partial f(\theta^*))$ respectively. Moreover, they are both nonzero vectors because every vector v in $\partial f(\theta^*)$ (and consequently $\operatorname{aff}(\partial f(\theta^*))$) is nonzero as it satisfies

$$\sum_{i=j}^{n} \binom{r+i-j-1}{r-1} v_i = \operatorname{sgn}\big((D^{(r)}\theta^*)_{j-r}\big)$$

whenever $(D^{(r)}\theta^*)_{j-r} \neq 0$ (it should be kept in mind that we are working under the assumption that $D^{(r)}\theta^* \neq 0$). It is helpful to note here that $v_0 = v^*$ when $r = 1$ (see Lemma 2.7) but for $r \geq 2$, they are not necessarily the same.

In addition to v^* and v_0, we need the following quantity:

$$\lambda_{\theta^*}(z) := \operatorname*{arg\,min}_{\lambda \geq 0} \inf_{v \in \partial f(\theta^*)} \|z - \lambda v\| \quad \text{for } z \in \mathbb{R}^n. \tag{26}$$

In words, $\lambda_{\theta^*}(z)$ is the value of λ which minimizes the distance of the vector z from the set $\lambda\partial f(\theta^*)$. Lemma B.5 in the supplementary material [15] proves that $\lambda_{\theta^*}(z)$ is uniquely defined for each $z \in \mathbb{R}^n$ (under the assumption that $D^{(r)}\theta^* \neq 0$) and also that $\mathbb{E}\lambda_{\theta^*}(Z) < \infty$ where the expectation is taken with respect to $Z \sim N_n(0, I_n)$. We are now ready to state our first result on the risk of the penalized trend filtering estimators (recall $\Delta_r(\theta)$ from (12)).

THEOREM 2.6. *Fix $r \geq 1$ and suppose $\theta^* \in \mathbb{R}^n$ with $D^{(r)}\theta^* \neq 0$. Let*

$$\lambda^* := n^{1-r}\bigg(\mathbb{E}\lambda_{\theta^*}(Z) + \frac{2}{\|v_0\|}\bigg), \tag{27}$$

where the expectation is taken with respect to the standard Gaussian vector $Z \sim N_n(0, I_n)$. Then for every regularization parameter $\lambda \geq \lambda^*$, we have

$$R\big(\hat{\theta}_\lambda^{(r)}, \theta^*\big) \leq C_r\sigma^2\Delta_r\big(\theta^*\big) + \frac{64\sigma^2}{n}\frac{\|v^*\|^2}{\|v_0\|^2} + \frac{4\sigma^2}{n^{3-2r}}\big(\lambda - \lambda^*\big)^2\big\|v^*\big\|^2 \tag{28}$$

for a constant C_r that only depends on r.

The bound (28) (which holds for every $\lambda \geq \lambda^*$) is clearly smallest when $\lambda = \lambda^*$. To simplify the right-hand side of (28) further, we need to bound $\|v^*\|$ from above and $\|v_0\|$ from below. This is done in the next result.

LEMMA 2.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\theta) := \|D^{(r)}\theta\|_1$ and let $\theta^* \in \mathbb{R}^n$ be such that $D^{(r)}\theta^* \neq 0$.*

1. *Suppose $r = 1$. Then $v_0 = v^*$. Further suppose that θ^* has $k \geq 1$ jumps (first order knots) with signs $\mathfrak{r}_1, \ldots, \mathfrak{r}_k$ and let $n_0, n_1, \ldots, n_k$ denote the lengths of the constant pieces of θ^*. Then*

$$\|v_0\|^2 = \big\|v^*\big\|^2 = \frac{1}{n_0} + \frac{1}{n_k} + 4\sum_{i=1}^{k-1} \frac{I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}}{n_i}. \tag{29}$$

2. *For $r \geq 2$, we have*

$$\|v_0\| \geq \frac{(r-1)!}{(r+1)2^{r-1}} n^{-r+1/2}. \tag{30}$$

3. *Suppose $r \geq 2$ and θ^* satisfies the minimum length condition (13) with constant c, then*

$$\|v^*\| \leq C_r c^{-r+1/2}(k+1)^r n^{-r+1/2}, \tag{31}$$

where C_r is a constant depending only on r.

We shall now present more explicit risk bounds by combining Theorem 2.6 and Lemma 2.7. Since the information provided by Lemma 2.7 about $\|v_0\|$ and $\|v^*\|$ is much more precise for $r = 1$ compared to $r \geq 2$, we find it natural to state our risk results separately in the two cases $r = 1$ and $r \geq 2$. The following result deals with the $r = 1$ case.

COROLLARY 2.8. *Suppose $\theta^* \in \mathbb{R}^n$ has $k \geq 1$ jumps with signs $\mathfrak{r}_1, \ldots, \mathfrak{r}_k$ and suppose that $n_0, n_1, \ldots, n_k$ denote the lengths of the constant pieces of θ^*. Then, with λ^* as in (27), we have*

$$R(\hat{\theta}^{(1)}_\lambda, \theta^*) \leq C\sigma^2\left(\Delta_1(\theta^*) + \frac{(\lambda-\lambda^*)^2}{n}\sum_{i=0}^{k}\frac{I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}}{n_i}\right) \tag{32}$$

for every $\lambda \geq \lambda^*$. Here C is a universal constant. Also, we use our usual convention $\mathfrak{r}_0 = \mathfrak{r}_{k+1} = 0$.

Further, if θ^* satisfies the minimum length condition (13) with constant c, then

$$R(\hat{\theta}^{(1)}_\lambda, \theta^*) \leq C(c)\sigma^2\left(\frac{k+1}{n}\log\frac{en}{k+1} + (\lambda-\lambda^*)^2\frac{k+1}{n^2}\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}\right), \tag{33}$$

where $C(c)$ depends on c alone.

Inequality (33) implies that, under the minimum length condition, we have

$$R(\hat{\theta}^{(1)}_\lambda, \theta^*) \leq C(c)\sigma^2\frac{k+1}{n}\log\frac{en}{k+1} \quad \text{for } \lambda = \lambda^*, \tag{34}$$

where k is the number of jumps of θ^*, that is, $k = \mathbf{k}_1(\theta^*)$. Moreover, the logarithmic term above cannot be removed in general. This is due to the following reason. First note that, for every nonrandom λ possibly depending on λ^*, the penalized estimator $\hat{\theta}^{(1)}_\lambda$ has worse risk compared to the ideally tuned constrained estimator, that is, $\hat{\theta}^{(1)}_V$ with $V = V^{(r)}(\theta^*)$. This fact (which is noted and explained in Section 5.2), together with Lemma 2.4, implies clearly that the logarithmic factor in (34) cannot be removed in general.

REMARK 2.6 (Comparison to existing results). Among the class of existing results for the risk of $\hat{\theta}^{(1)}_\lambda$, the strongest (in terms of giving the smallest bound on the risk) is due to Lin et al. [23] who proved that, when λ is appropriately selected (depending on θ^*), $\hat{\theta}^{(1)}_\lambda$ satisfies

$$R(\hat{\theta}^{(1)}_\lambda, \theta^*) \leq C\frac{\sigma^2(k+1)}{n}\left(\left[\log(k+1) + \log\log n\right]\log n + \sqrt{k+1}\right) \tag{35}$$

provided

$$\min_{0\leq i\leq k} n_i \geq \frac{cn}{k+1} \tag{36}$$

for a positive constant c. Here $n_0, \ldots, n_k$ are the lengths of the constant pieces of θ^*. This bound from Lin et al. [23] is smaller compared to an earlier result of Dalalyan, Hebiri and Lederer [6] and to a very recent result of Ortelli and van de Geer [27] (although the results of [6, 27] apply to a universal choice of the tuning parameter λ; see Remark 2.7). The bound

(35) is weaker than (34) in two respects: (a) there are additional terms in (35) involving $\log n$ and k compared to (34), and (b) our minimum length condition (13) is weaker than (36): (13) requires that $n_i \geq cn/(k+1)$ only for those i for which $\mathfrak{r}_i \neq \mathfrak{r}_{i+1}$ while (36) requires this for all i.

Note that the regularization parameter λ^* (for which the near parametric risk bound (34) holds) depends on θ^*. Further, the exact nature of its dependence on θ^* is not apparent from its definition (27). In the next result, we provide a more explicit upper bound for λ^*. For this, we require a stronger length condition than (13). Note that we are still in the $r = 1$ case.

LEMMA 2.9. *Consider the same setting as in Corollary 2.8. Assume that the length condition*:

$$\min_{0 \leq i \leq k : \mathfrak{r}_i \neq \mathfrak{r}_{i+1}} n_i \geq \frac{c_1 n}{k+1} \quad \text{and} \quad \max_{0 \leq i \leq k : \mathfrak{r}_i \neq \mathfrak{r}_{i+1}} n_i \leq \frac{c_2 n}{k+1} \tag{37}$$

holds for two positive constants $c_1 \leq 1$ and $c_2 \geq 1$. Let λ^* be as defined in (27). Then there exists a positive constant $C^*(c_1, c_2)$ (which depends only on c_1 and c_2) such that

$$\lambda^* \leq C^*(c_1, c_2) \sqrt{\frac{n}{\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}} \log\left(\frac{en}{k+1}\right)}. \tag{38}$$

Lemma 2.9 can be used, in conjunction with the risk bound (33) (which holds for every $\lambda \geq \lambda^*$) to yield the following result which provides bounds similar to (34) for explicit choices of λ.

COROLLARY 2.10. *Consider the same setting as in Lemma 2.9 and assume the length condition* (37). *Then if the regularization parameter λ satisfies*

$$\lambda = \Gamma \sqrt{\frac{n}{\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}} \left(\log \frac{en}{k+1}\right)}, \tag{39}$$

we have

$$R\big(\hat{\theta}_\lambda^{(1)}, \theta^*\big) \leq C(c_1) \sigma^2 \big(1 + \Gamma^2\big) \frac{k+1}{n} \log \frac{en}{k+1} \tag{40}$$

for every $\Gamma \geq C^*(c_1, c_2)$ (where $C^*(c_1, c_2)$ is the constant given by Lemma 2.9). Also $C(c_1)$ depends only on c_1.

Also, if the regularization parameter λ satisfies

$$\lambda = \Gamma \sqrt{n \log(en)}, \tag{41}$$

we have

$$R\big(\hat{\theta}_\lambda^{(1)}, \theta^*\big) \leq C(c_1) \frac{\sigma^2 (k+1)(\log(en))}{n} \left(1 + \Gamma^2 \sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}\right) \tag{42}$$

for every $\Gamma \geq C^*(c_1, c_2)$.

In the bound (42), the term $\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}$ can be further bounded by its maximum possible value of $k+1$. However, in certain instances (such as when θ^* is monotone), $\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}$ can be much smaller than $k+1$.

REMARK 2.7 (Comparison to existing results). We now compare Corollary 2.10 to existing results for the penalized estimator in Lin et al. [23], Dalalyan, Hebiri and Lederer [6] and Ortelli and van de Geer [27]. Note first that the choice (39) of λ depends on certain aspects of θ^*: in particular, it depends on k, $\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}$ and the values c_1 and c_2 in the length condition (37). The bound (35) of Lin et al. [23] holds for $\lambda_1 = (n \min_{0\le i\le k} n_i)^{1/4}$ which also depends on the true vector θ^* through the lengths $n_1, \ldots, n_k$. If we assume that each n_i is of order $n/(k+1)$, then

$$\lambda_1 \sim \sqrt{\frac{n}{\sqrt{k+1}}}. \tag{43}$$

Note that the leading term in our choice (39) of λ as well as in λ_1 is $\sqrt{n}$. Corollary 2.10 also applies to the choice (41) for which the bound (42) holds. Note that (41) has considerably less dependence on θ^* as it only depends on the constants c_1 and c_2 appearing in the length condition (37). On the other hand, the bound (42) is weaker compared to (40). However, (42) needs to be compared to the results of Dalalyan, Hebiri and Lederer [6], Proposition 3, and Ortelli and van de Geer [27], Corollary 4.4. Indeed, Dalalyan, Hebiri and Lederer [6] considered the choice

$$\lambda_2 := 2\sqrt{2n \log(n/\delta)} \tag{44}$$

and proved that the following loss bound holds with probability at least $1-\delta$:

$$\frac{1}{n}\|\hat{\theta}^{(1)}_{\lambda} - \theta^*\|^2 \le C(c_1)\left(\frac{(k+1)^2}{n}\log\frac{en}{\delta} + \frac{k+1}{n}\log(en)\log\frac{en}{\delta}\right). \tag{45}$$

This result has been improved slightly in the very recent paper Ortelli and van de Geer [27] (see also van de Geer [37]) where the $\log(en)\log(en/\delta)$ term in the right-hand side above is replaced by $\log(en/(k+1))\log(en/\delta)$ (i.e., one of the $\log(en)$ terms is replaced by $\log(en/(k+1))$). An expectation (risk) bound has not been proved in these two papers. Note the the choice of λ in (41) is similar to that of λ_2 in (44) although our choice needs Γ to be sufficiently large while the choice λ_2 is universal (although it depends on δ). On the other hand, the high probability bound implied by (42) is (see Remark 2.3) the statement that

$$\frac{1}{n}\|\hat{\theta}^{(1)}_{\lambda} - \theta^*\|^2 \le C(c_1)\frac{\sigma^2(k+1)(\log(en))}{n}\left(1 + \Gamma^2\sum_{i=0}^{k} I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}\right) + \frac{4\sigma^2}{n}\log(\delta^{-1})$$

holds with probability at least $1-\delta$. This is stronger compared to (45) because the right-hand side of (45) has a $\log(en)\log(en/\delta) \ge (\log(en))^2$ term.

We reiterate here that our length condition (37) involves an upper bound on n_i for $\mathfrak{r}_i \neq \mathfrak{r}_{i+1}$. From an examination of the proof of Lemma 2.9, it will be clear that we will obtain a weaker upper bound for λ^* in the sense of having additional multiplicative factors involving k if this upper bound assumption on n_i is removed. No such upper bound is needed for the results in Lin et al. [23], Dalalyan, Hebiri and Lederer [6], Ortelli and van de Geer [27]. On the other hand, our lower bound (and our upper bound in (37)) involves only those i satisfying $\mathfrak{r}_i \neq \mathfrak{r}_{i+1}$ while the assumptions in these earlier papers required a lower bound on every n_i.

We now state our risk results for (4) with $r \ge 2$ when $D^{(r)}\theta^* \neq 0$. The following result is obtained by combining Theorem 2.6 and Lemma 2.7.

COROLLARY 2.11. *Fix $r \geq 2$. Suppose $D^{(r)}\theta^* \neq 0$ and θ^* satisfies the minimum length condition (13) with constant c. Then, with λ^* as in (27), we have*

$$R\big(\hat{\theta}^{(r)}_\lambda, \theta^*\big) \leq C_r(c)\sigma^2\bigg(\frac{k+1}{n}\log\frac{en}{k+1} + \frac{(k+1)^{2r}}{n} + \big(\lambda - \lambda^*\big)^2\frac{(k+1)^{2r}}{n^2}\bigg) \tag{46}$$

for every $\lambda \geq \lambda^*$. Here $k := \mathbf{k}_r(\theta^*)$ and $C_r(c)$ depends only on c.

Corollary 2.11 implies that when θ^* satisfies the minimum length condition (13), then (with $k = \mathbf{k}_r(\theta^*)$)

$$R\big(\hat{\theta}^{(r)}_\lambda, \theta^*\big) \leq C_r(c)\sigma^2\bigg(\frac{k+1}{n}\log\frac{en}{k+1} + \frac{(k+1)^{2r}}{n}\bigg) \quad \text{for } \lambda = \lambda^*. \tag{47}$$

It may be noted that the above result is weaker than our corresponding risk bound for the constrained trend filtering estimator (Corollary 2.3) because of the additional term involving $(k+1)^{2r}$. We believe that this term is redundant and is an artifact of our proof. Specifically, this additional term comes from the fact that our upper bound for $\|v^*\|$ and lower bound for $\|v_0\|$ in Lemma 2.7 are off by a factor of $(k+1)^r$.

With the aim of providing an explicit value for λ for which the bound (47) holds, the next result gives an upper bound for λ^*. As in the case of Lemma 2.9, we need a stronger length condition (compared to (13)) for this result.

LEMMA 2.12. *Fix $r \geq 2$. Suppose $D^{(r)}\theta^* \neq 0$ and θ^* satisfies the length condition*:

$$\min_{0 \leq i \leq k : \mathfrak{r}_i \neq \mathfrak{r}_{i+1}} n_i \geq \frac{c_1 n}{k+1} \quad \text{and} \quad \max_{0 \leq i \leq k : \mathfrak{r}_i \neq \mathfrak{r}_{i+1}} n_i \leq \frac{c_2 n}{k+1} \tag{48}$$

for two positive constants $c_1 \leq 1$ and $c_2 \geq 1$. Here $n_0, \ldots, n_k$ have the same meaning as in (13). Then λ^* (defined as in (27)) satisfies

$$\lambda^* \leq C^*_r(c_1, c_2)\sqrt{n\log\bigg(\frac{en}{k+1}\bigg)}, \tag{49}$$

where $C^*_r(c_1, c_2)$ depends on r, c_1 and c_2 alone.

Note that even though (48) and (37) look exactly the same, the difference is that (37) applies to $r = 1$ while (48) applies to $r = 2$. The meaning of $n_0, \ldots, n_k$ depends on r. Indeed, the n_i's refer to the lengths of the constant pieces for $r = 1$, the lengths of the linear pieces for $r = 2$, etc.

Compared to (38), the bound (49) is weaker because there is no $\sum_{i=0}^k I\{\mathfrak{r}_i \neq \mathfrak{r}_{i+1}\}$ in the denominator in (49).

Combining Lemma 2.12 with the risk bound (46), we obtain the following result which provides bounds similar to (47) for explicit choices of λ.

COROLLARY 2.13. *Consider the same setting as in Lemma 2.12 and assume the length condition (48). Then if the regularization parameter satisfies*

$$\lambda = \Gamma\sqrt{n\log\bigg(\frac{en}{k+1}\bigg)}, \tag{50}$$

we have

$$R(\hat{\theta}_\lambda^{(r)}, \theta^*) \le C_r(c_1)\sigma^2(2+\Gamma^2)\frac{(k+1)^{2r}}{n}\log\frac{en}{k+1} \tag{51}$$

for every $\Gamma \ge C_r^*(c_1, c_2)$ *(where* $C_r^*(c_1, c_2)$ *is the constant given by Lemma 2.9). Also* $C_r(c_1)$ *only depends only on* r *and* c_1.

Further, if the regularization parameter λ *satisfies*

$$\lambda = \Gamma\sqrt{n\log(en)}, \tag{52}$$

we have

$$R(\hat{\theta}_\lambda^{(r)}, \theta^*) \le C_r(c_1)\sigma^2(2+\Gamma^2)\frac{(k+1)^{2r}}{n}\log(en) \tag{53}$$

for every $\Gamma \ge C_r^*(c_1, c_2)$.

Finally, we deal with the risk of the penalized estimator when $D^{(r)}\theta^* = 0$. Here we have the following result which proves that the risk is parametric (without any logarithmic factors) as long as the tuning parameter λ is larger than or equal to $\sqrt{6n\log(en)}$. This result holds for every $r \ge 1$.

LEMMA 2.14. *Suppose* $D^{(r)}\theta^* = 0$. *Then for every* $\lambda \ge \sqrt{6n\log(en)}$, *we have*

$$R(\hat{\theta}_\lambda^{(r)}, \theta^*) \le \frac{C_r\sigma^2}{n},$$

for a constant C_r *that depends on* r *alone.*

3. Proof ideas. In this section, we provide a brief overview of the main ideas underlying our proofs. Full proofs are in the supplementary material [15]. For studying the constrained trend filtering estimator $\hat{\theta}_V^{(r)}$, we invoke the general theory of convex-constrained least squares estimators. Convex-constrained least squares estimators are estimators of the form

$$\hat{\theta} := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^n}\left\{\frac{1}{2}\|Y-\theta\|^2 : \theta \in K\right\},$$

for a closed convex set K. Clearly, $\hat{\theta}_V^{(r)}$ is a special case of this estimator when K is taken to be the set $K^{(r)}(V)$ defined as

$$K^{(r)}(V) := \{\theta \in \mathbb{R}^n : \|D^{(r)}\theta\|_1 \le Vn^{1-r}\}.$$

The general theory of convex-constrained least squares estimators (summarized in Section A of the supplementary material [15]) states that the accuracy of $\hat{\theta}_V^{(r)}$ as an estimator for θ^* under the model $Y \sim N_n(\theta^*, \sigma^2 I_n)$ can be deduced from bounds on the quantity:

$$\mathbb{E}\sup_{\theta\in K_V^{(r)}:\|\theta-\theta^*\|\le t}\langle\xi, \theta-\theta^*\rangle, \tag{54}$$

where $\xi \sim N_n(0, \sigma^2 I_n)$. To prove Theorem 2.1, we prove bounds on (54) in [15], Lemma B.1. Our strategy involves using Dudley's entropy bound to control (54) in terms of the metric entropy of the set:

$$S_r(V, t) := \{\alpha \in \mathbb{R}^n : \|\alpha\| \le t, \|D^{(r)}\alpha\|_1 \le Vn^{1-r}\}.$$

We then bound the metric entropy of $S_r(V,t)$ via its fat-shattering dimension (it is well known that fat-shattering dimension can be used to control metric entropy; see e.g., Rudelson and Vershynin [32]). Metric entropy and fat-shattering dimension are formally defined in the supplementary material [15]. Our idea of using fat shattering to establish the metric entropy of $S_r(V,t)$ and thereby bounding (54) seems novel. Previous bounds on quantities similar to (54) in the context of trend filtering used eigenvector incoherence (see, e.g., Wang et al. [38]) and the ideas here are quite different from our methods.

To prove the strong sparsity risk bound, Theorem 2.2, we use another strand of results from the general theory of convex-constrained least squares estimators. Specifically, a result from Oymak and Hassibi [28] implies that the risk of $\hat{\theta}_V^{(r)}$ at $V = V^* := V^{(r)}(\theta^*)$ can be obtained by controlling the *Gaussian width* of the *tangent cone* of the convex set $K^{(r)}(V^*)$ at θ^*. These general results, along with the definitions of tangent cones and Gaussian width, are again recalled in [15], Subsection A. Understanding the tangent cone to $K^{(r)}(V^*)$ at θ^* then becomes key to proving Theorem 2.2.

We provide a precise characterization of the tangent cones of $K^{(r)}(V^*)$ in [15], Lemma C.3. These tangent cones have a complicated structure (especially for $r \geq 2$) and calculating their Gaussian width is nontrivial. Our idea behind these calculations is the fact (proved in Lemma B.2) that, under a unit norm constraint, every vector α in the tangent cone of $K^{(r)}(V^*)$ at θ^* is nearly made up of two $(r-1)$th order convex/concave sequences in each polynomial part of θ^* (note that a sequence $\theta \in \mathbb{R}^n$ is said to be $(r-1)$th order convex/concave if the vector $D^{(r-1)}\theta$ is monotone; see, e.g., Kuczma [20]). The special case of this observation for $r = 1$ implies that every vector α with $\|\alpha\| \leq 1$ in the tangent cone to $K^{(1)}(V^*)$ at θ^* is nearly made up of two monotonic sequences in each constant piece of θ^*. For $r = 2$, it means that every vector α with $\|\alpha\| \leq 1$ in the tangent cone to $K^{(2)}(V^*)$ at θ^* is nearly made up of two convex/concave sequences in each linear piece of θ^*.

The above observation allows us to compute the Gaussian width of these tangent cones using metric entropy results (established again via connections between metric entropy and fat shattering) and also available results (from Bellec [2]) on the Gaussian widths of shape constrained cones. The set of all $(r-1)$th order convex sequences in $\mathbb{R}^n$ forms a convex cone in $\mathbb{R}^n$ and these cones have been studied in the literature on shape constrained estimation.

For $r = 1$, the above idea bears strong similarities with the method employed in Lin et al. [23] for studying the penalized estimator (4) for $r = 1$. In this paper, they use the key observation that for appropriate λ, the vector $(I - P_0)(\hat{\theta}_\lambda^{(1)} - \theta^*)$ is well approximated by a vector which is made of two monotonic sequences in each constant piece of θ^*. Here P_0 is the projection matrix onto the piecewise constant structure determined by θ^* and I is the identity matrix. This idea is similar in spirit to our observation on the tangent cone of $K^{(1)}(V^*)$ at θ^*. The details differ though as we are working with the vectors in the tangent cone while Lin et al. [23] focus on a functional of $\hat{\theta}_\lambda^{(1)} - \theta^*$ (note though that if $\hat{\theta}$ has variation $\leq V^*$, then $\hat{\theta} - \theta^*$ does indeed belong to the tangent cone). Also our method for dealing with the Gaussian width of the set of these piecewise monotonic vectors is sharper than the analysis of Lin et al. [23] and our analysis also extends to every $r \geq 2$.

The results in Section 2.2 for the penalized estimator are all based on (22). We use the precise characterization of the subdifferential of the penalty function $\theta \mapsto \|D^{(r)}\theta\|_1$ given in Proposition 2.5 to control the right-hand side of (22). Our idea here is to relate the right-hand side of (22) to the risk of the constrained estimator (we use and extend ideas from Foygel and Mackey [10] for this). This allows us to derive risk results for the penalized trend filtering estimator as a corollary to our results for the constrained estimator.

4. Simulations. In this section, we present numerical evidence for our theoretical results. We generate data from a piecewise constant function f_1^* and a continuous piecewise

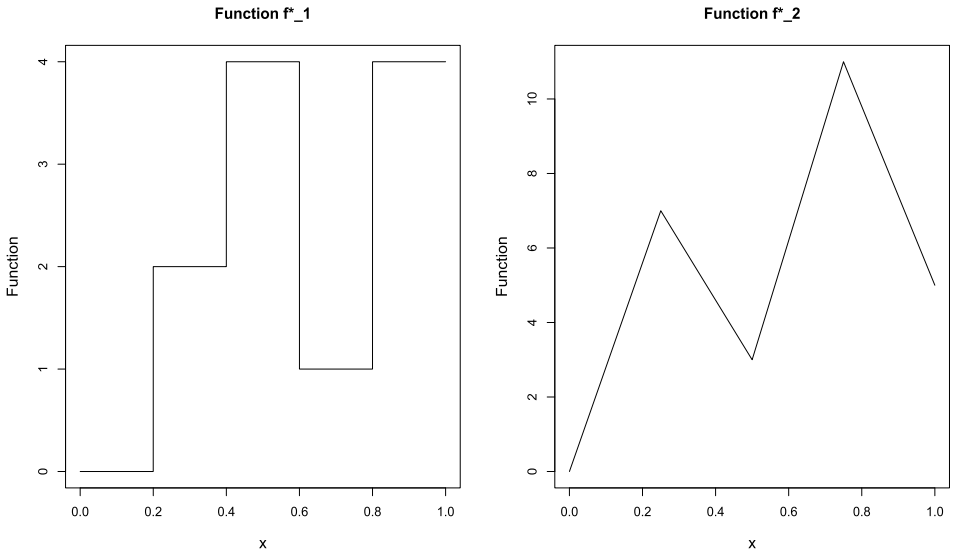

FIG. 3. *The two functions f_1^* and f_2^*.*

affine function f_2^* on $[0, 1]$ and evaluate the performance of the trend filtering estimators for $r = 1$ (total variation denoising) and $r = 2$, respectively. The functions f_1^* and f_2^* (see Figure 3) are given by

$$f_1^*(x) := 2I_{(0.2,0.4]}(x) + 4I_{(0.4,0.6]}(x) + I_{(0.6,0.8]}(x) + 4I_{(0.8,1]}(x)$$

and

$$f_2^*(x) := -44\max(x - 0.25, 0) + 48\max(x - 0.5, 0) - 56\max(x - 0.75, 0) + 28x.$$

The function f_1^* was used in the simulation study of Lin et al. [23]. In addition to these functions, we also performed a simulation study on another piecewise constant function f_3^* which is similar to the blocks function of Donoho and Johnstone [7]; for space constraints we have moved the results for f_3^* to the supplementary material (see Section E in [15]).

From f_1^* and a value of n (chosen from a grid of size 30 between 100 and 10,000; the grid being equally spaced on the logarithmic scale), we generated an $n \times 1$ observation vector $Y \sim N_n(\theta^*, I_n)$ where θ^* is the vector obtained by sampling f_1^* at n equally spaced points with end-points 0 and 1. We then computed the following six estimators on the data vector Y: (a) the ideal constrained estimator (3) with $V = V^* = \|D\theta^*\|_1$, (b) the ideal penalized estimator (4) with $\lambda = \lambda^*$ (as defined in (27)), (c) two cross-validation (CV) based estimators, (d) the penalized estimator (4) with λ of the form (39) with $\Gamma = 1$ and (e) the penalized estimator (4) with λ of the form (41) with $\Gamma = 0.5$. Corollary 2.10 proves that the risk with these λ choices decays as $(\log n)/n$ (ignoring terms involving k) provided Γ is taken to be a large enough constant. In our simulations for f_1^*, we found that $\Gamma = 1$ in (39) and $\Gamma = 0.5$ in (41) were large enough to yield the desired performance. Higher values of Γ led to similar rates of decay of the risk with n (even though the risk itself seemed to become larger with Γ).

Here are some details behind the computation of these estimates. The constrained estimator was computed by the convex optimization software MOSEK (via the R package `Rmosek`). The penalized estimators were computed via the R package `tvd` for total variation denoising. The computation of the ideal penalized estimator requires computing the value of λ^* and, for this, we need to compute $\mathbb{E}\lambda_{\theta^*}(Z)$ (where $Z \sim N_n(0, I_n)$) and $2/\|v_0\|$ (see (27)). $2/\|v_0\|$ was calculated by the formula (29). For $\mathbb{E}\lambda_{\theta^*}(Z)$, we used the fact that $\lambda_{\theta^*}(z)$ can be calculated by convex optimization for each $z \in \mathbb{R}^n$ which implies that the expectation can be computed by

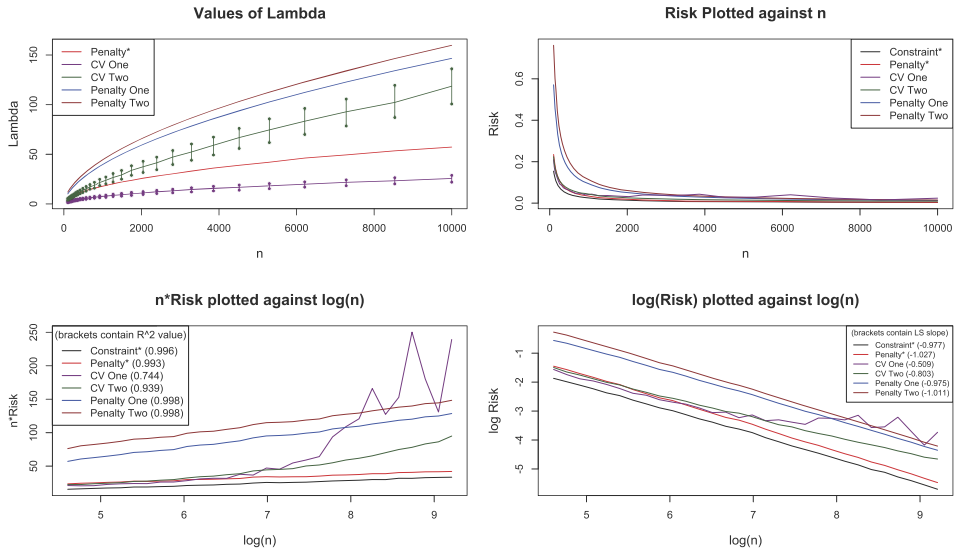

FIG. 4. *Plots when the true function is* f_1^*. *The top-left plot shows the* λ^* *values, the CV* λ *values (median and the first and third quartiles over* 200 *replications) and the values corresponding to the explicit penalties* (39) *with* $\Gamma = 1$ *and* (41) *with* $\Gamma = 0.5$. *The other three figures show the behavior of the risk as a function of* n. *In the last two plots, the legend shows the value of* R^2 *and the slope respectively for the curves corresponding to each estimator.*

Monte Carlo averaging. More details behind this are provided in the supplementary material (Section E in [15]). The CV estimators were calculated using the R package `genlasso` which provides two penalized estimates based on CV: one based on choosing λ so as to minimize the CV error (CV_1) and the other based on choosing λ via the one standard error rule (CV_2).

For each data set, we computed the value of the loss $\|\hat{\theta} - \theta^*\|^2/n$ for each of these six estimates. We generated 600 replications of the data for each value of n to compute the average value of the loss which is an approximation of the risk of each estimator. Our results are provided in Figure 4. The top-left plot shows the different values of λ employed by the estimators based on (4). Here we plotted the λ^* values as well as those corresponding to (39) with $\Gamma = 1$ (penalty one) and (41) with $\Gamma = 0.5$ (penalty two). In addition, we also plotted here the penalty levels chosen by the CV estimators. These are random so we plotted their median and quartile values over the 600 replications. The remaining three plots in Figure 4 show the risks of the six estimators. In the top-right plot, the risk is simply plotted as a function of n (from our theoretical results, the risk is supposed to decay like the curve $n \mapsto (t_1/n)\log(t_2 n)$ for two constants t_1 and t_2). In the bottom-left plot, we plotted n times the risk against $\log n$. These curves are supposed to be linear so we provided the squared correlation (R^2) values of each of the curves in this plot. One can see that the R^2 values are close to one for every estimator except CV_1. Finally, in the bottom-right plot, we plotted the logarithm of the risk against $\log n$. We expect the curves here to have a near-linear relationship with negative slope of -1. The least squares slope values for the different curves are given in the legend in this and it is clear that, for the non-CV estimators, the slope is indeed close to -1.

The numerical results in Figure 4 for the non-CV estimates therefore clearly support our theoretical results. On the other hand, the behavior of the CV estimators seems more complicated and a theoretical study of their risk performance is beyond the scope of the present paper.

We also show results for f_2^* where we evaluated the performance of trend filtering for $r = 2$. We did a simplified study here with the three estimators: (a) the ideal constrained

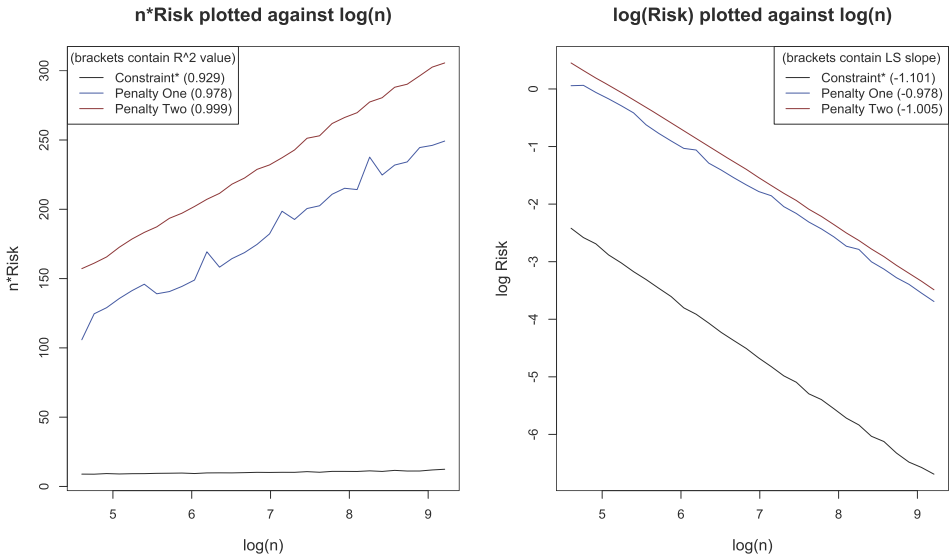

FIG. 5. *Risk plots when the true function is* f_2^*.

estimator (3) with $V = V^* = n\|D^2\theta^*\|_1$, (b) the penalized estimator (4) with λ taken to be (50) with $\Gamma = 1/16$ and (c) the penalized estimator (4) with λ taken to be (52) with $\Gamma = 1/16$. Note that our theoretical results apply to (50) and (52) for a sufficiently large Γ. For f_2^*, we found in simulations that $\Gamma = 1/16$ was large enough to yield the desired rates. Higher values of Γ inflated risk but gave similar risk decay rates. We could not compute the ideal penalized estimator with $\lambda = \lambda^*$ (defined in (27)) here as the convex optimization problem to compute $\lambda_{\theta^*}(z)$ was highly ill-conditioned for $n \geq 1000$ so that MOSEK seemed unable to find the global minimum (see Section E of [15] for more details). We also did not compute CV estimates here as these are not the focus of this paper.

Our results are given in Figure 5. The left plot shows n times the risk plotted against $\log n$. Our theory indicates that the curve corresponding to each estimator should be linear so we provided the squared correlation (R^2) values which are all close to 1. The right plot shows the behavior of log risk against $\log n$. These curves are expected to have a near-linear relationship with negative slope of -1. The legend shows the least squares slopes which are all close to -1. These plots therefore support our theoretical results.

5. Discussion. In this section, we address various issues that are naturally linked to our main results.

5.1. *Weakening our assumptions.* We emphasized the vector estimation setting (2) in this paper. Our results can also be interpreted in the function estimation setting in the following way. There is an unknown function f^* and we observe data $Y_1, \ldots, Y_n$ according to the model:

$$Y_i = f^*(x_i) + \xi_i \quad \text{for } i = 1, \ldots, n,$$

where $f^* : [0, 1] \to \mathbb{R}$ is the unknown regression function and $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0, \sigma^2)$. We focused on the situation where $x_i = i/n$ for $i = 1, \ldots, n$. We can estimate f^* by any discrete spline $\hat{f}$ of degree $r - 1$ whose values at i/n, $1 = 1, \ldots, n$, are given by $\hat{\theta}_1, \ldots, \hat{\theta}_n$ (with $\hat{\theta}$ defined as in (3) or (4)). We then evaluate the performance of $\hat{f}$ as an estimator for f^* via the loss $\frac{1}{n}\sum_{i=1}^n (f^*(x_i) - \hat{f}(x_i))^2$ and prove bounds for the risk when f^* is a discrete spline in terms of the number of polynomials that make up f^*.

This basic setting (which is standard and used in many theoretical papers on univariate nonparametric regression) can be generalized in many ways and we mention two extensions involving the design points $x_1, \dots, x_n$ below. One is the situation where $x_1, \dots, x_n$ are not equally spaced. In this case, note that the penalty terms in (3) and (4) need to be changed for $r \geq 2$; see for example, Tibshirani [36]. We believe that our results will still hold in this case provided $x_1, \dots, x_n$ satisfy $\kappa_1/n \leq x_i - x_{i-1} \leq \kappa_2/n$ for two constants κ_1 and κ_2. However, this would make the notation in our proofs quite cumbersome.

One can also study the setting where $x_1, \dots, x_n$ are generated independently from a common distribution ν on $[0, 1]$ and/or we measure the loss via $\int (\hat{f}(x) - f^*(x))^2 \, d\nu(x)$. Analyzing this situation will require handling additional approximation error terms and we will leave it for future work.

5.2. *Constrained and penalized estimators.* As mentioned in the Introduction, we have studied both constrained and penalized versions of trend filtering while previous papers have focused on the penalized estimator alone. When the noise level σ tends to zero, it can be proved that the constrained estimator with $V = V^* := V^{(r)}(\theta^*)$ is better than the penalized estimator for every choice of the tuning parameter λ. More precisely,

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R\big(\hat{\theta}^{(r)}_{V^*}, \theta^*\big) < \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R\big(\hat{\theta}^{(r)}_{\lambda}, \theta^*\big) \quad \text{for every } \lambda \in [0, \infty). \tag{55}$$

Here λ is even allowed to depend on θ^* as long as it is nonrandom. Inequality (55) follows from the results of Oymak and Hassibi [28] as described below. Oymak and Hassibi [28], Theorem 2.1, implies

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R\big(\hat{\theta}^{(r)}_{V^*}, \theta^*\big) = \frac{1}{n} \mathbb{E}\Big(\inf_{v \in \operatorname{cone}(\partial g(\theta^*))} \|Z - v\|^2 \Big) \tag{56}$$

and Oymak and Hassibi [28], Theorem 1.1, implies

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R\big(\hat{\theta}^{(r)}_{\lambda}, \theta^*\big) = \frac{1}{n} \mathbb{E}\Big(\inf_{v \in \lambda \partial g(\theta^*)} \|Z - v\|^2 \Big) \tag{57}$$

for every $\lambda \geq 0$. Here $g(\theta) := n^{r-1} \|D^{(r)}\theta\|_1$, $\lambda \partial g(\theta^*) := \{\lambda v : v \in \partial g(\theta^*)\}$, $\operatorname{cone}(\partial g(\theta^*)) := \bigcup_{\lambda \geq 0} \lambda \partial g(\theta^*)$ and $Z \sim N_n(0, I_n)$. As $\operatorname{cone}(\partial g(\theta^*))$ is strictly larger than $\lambda g(\theta^*)$ for every fixed $\lambda > 0$, the right-hand side of (56) will be strictly smaller than the right-hand side of (57) which proves (55).

The implication of this inequality is that there exist settings (where σ is small) where the constrained estimator with $V = V^*$ is better than every penalized estimator. Therefore, it makes sense to study the constrained estimator in addition to the penalized estimator.

5.3. *Results for data-dependent tuning parameters.* From a practical point of view, a major limitation of the results of this paper is that they only hold for ideal or oracle choices of the tuning parameters. Indeed, our strong sparsity risk bounds for the constrained estimator require V to be close to $V^* := V^{(r)}(\theta^*)$. On the other hand, our risk bounds for the penalized estimator require knowledge of the noise level σ (note that the tuning parameter in (4) involves σ) as well as certain aspects of θ^*. For example, the choices (27), (39) and (50) depend on certain properties of the locations and signs of the knots of θ^*. The choices (41) and (52) have lesser dependence on θ^* but they still depend on the constants c_1 and c_2 from the condition (48).

We would like to note that this feature is also present in earlier papers on the trend filtering estimators. The strong sparsity risk results of Lin et al. [23] hold for the tuning choice (43) which depends on θ^*. The results of Dalalyan, Hebiri and Lederer [6] and Ortelli and van de

Geer [27] hold for the tuning choice (44) which does not depend on θ^* but depends on the noise level σ and the probability level δ (note that these results of [6, 27] give only high probability statements and not expectation (risk) bounds).

We would like to highlight the problem of proving risk bounds under strong sparsity for completely data-dependent choices of the tuning parameters as a major open problem. One can approach this problem via the constrained estimator which would require estimation of the variation functional $V^{(r)}(\theta^*)$. Alternatively, one can approach this problem via the penalized estimator which would require estimation of σ and λ^* (defined in (27)). It will be interesting to see if the risk of $\log(en)/n$ (up to multiplicative factors depending on k) will be achieved for a completely data dependent method of tuning.

5.4. *Connections to results for the LASSO.* The trend filtering estimators are closely related to the LASSO estimator of Tibshirani [35]. Indeed, for $r = 1$, it is easy to see that the constrained estimator $\hat{\theta}_V^{(1)}$ is exactly equal to $X\hat{\beta}_V$ where X is the $n \times n$ matrix whose (i, j)th entry equals $I\{i \geq j\}$ and $\hat{\beta}_V := \arg\min_{\theta \in \mathbb{R}^n}\{\|Y - X\beta\|^2 : \sum_{i=2}^n |\beta_i| \leq V\}$. Therefore, our strong sparsity risk results for $\hat{\theta}_V^{(1)}$ can simply be seen as results for the LASSO estimator for this special design matrix X. This connection to LASSO also holds for $r \geq 2$ (see Tibshirani [36]).

Based on this link to the LASSO, it might seem possible to believe that our results might be derivable from general theorems about the LASSO. However, existing strong sparsity risk bounds for the LASSO impose stringent conditions on the design matrix (such as the compatibility condition or the restricted eigenvalue condition) which do not hold for this particular design matrix X (see Dalalyan, Hebiri and Lederer [6]). The relaxed compatibility condition of [6] does hold for this X and the authors of [6] use this observation to prove rates under strong sparsity but their argument is not strong enough to yield the $\frac{k+1}{n} \log \frac{en}{k+1}$ bound. More importantly, it is not clear if the relaxed compatibility condition or a modified version of it holds for $r \geq 2$.

5.5. *Comparison to the L^0 estimators.* It is natural to compare the performance of the trend-filtering estimators to the estimators obtained by replacing the L^1 norm in (3) by the L^0 norm:

$$\hat{\theta}_k^{(r)} := \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|Y - \theta\|^2 : \left\|D^{(r)}\theta\right\|_0 \leq k \right\}. \tag{58}$$

Under strong sparsity, that is, $\|D^{(r)}\theta^*\|_0 \leq k$, it should be possible to prove that

$$R\big(\hat{\theta}_k^{(r)}, \theta^*\big) \leq C_r \frac{\sigma^2(k+1)}{n} \log \frac{en}{k+1}. \tag{59}$$

A proof of this result for $r = 1$ can be found in the recent paper Gao, Han and Zhang [12], Theorem 2.1. We could not find an exact reference for $r \geq 2$ but we believe that (59) should be true based on the regression connection described in the previous subsection and existing results for L^0-penalized estimators in linear regression (see, e.g., [30], Theorem 4).

From a comparison of (59) with (18), it might seem that the constrained trend filtering estimator (with $V = V^*$) has similar performance under strong sparsity as that of the L^0 estimator. However, it must be kept in mind here that (18) requires the minimum length condition (13) while the bound (59) for the L^0 estimator does not require any such minimum length condition. Without the minimum length condition, the L^1 estimator performs much worse compared to the L^0 estimator as proved in the recent paper Fan and Guan [9]. Note, however, that the minimum length condition is quite natural from the point of view of estimating piecewise polynomial functions.

From a computational viewpoint, (58) can be efficiently computed for $r = 1$ via dynamic programming (see, e.g., Winkler and Liebscher [40]) but it is not clear how to compute it for $r \geq 2$. On the other hand, the trend filtering estimators are efficiently computable for every $r \geq 2$ via convex optimization (see, e.g., Arnold and Tibshirani [1] and Kim et al. [19] for details).

5.6. *Connection to shape constrained estimators.* Shape constrained regression estimators are closely related to the trend filtering estimators. Indeed, if one takes the constrained trend filtering estimator (3) and replaces the L^1 constraint by a nonnegativity constraint on $D^{(r)}\theta$, then we obtain shape constrained estimators. Specifically, consider

$$\hat{\theta}^{(r)}_{\text{shape}} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|Y - \theta\|^2 : D^{(r)}\theta \geq 0 \right\}. \tag{60}$$

Here $D^{(r)}\theta \geq 0$ means that each component of the vector $D^{(r)}\theta$ is nonnegative. When $r = 1$, (60) coincides with the classical isotonic least squares estimator and when $r = 2$, (60) coincides with the convex least squares estimator (see Groeneboom and Jongbloed [14] for an introduction to shape constrained estimation). Like the trend filtering estimators, the shape constrained estimators enjoy the property that $D^{(r)}\hat{\theta}^{(r)}_{\text{shape}}$ is sparse. However, unlike the trend filtering estimators, there is no tuning parameter in (60) (of course, (60) is only applicable in situations where θ^* satisfies the constraint $D^{(r)}\theta^* \geq 0$ exactly or in some approximate sense).

The risk of (60) under the strong sparsity assumption (and the shape assumption $D^{(r)}\theta \geq 0$) has received much recent attention (see Guntuboyina and Sen [16] for a recent survey). In Bellec [2], it was proved that

$$R\big(\hat{\theta}^{(r)}_{\text{shape}}, \theta^*\big) \leq \inf_{\theta : D^{(r)}\theta \geq 0} \left(\frac{1}{n}\|\theta^* - \theta\|^2 + C_r \frac{\sigma^2(k+1)}{n} \log \frac{en}{k+1} \right), \tag{61}$$

where $k := \mathbf{k}_r(\theta) = \|D^{(r)}\theta\|_0$. This result is very similar to our risk bounds for the constrained trend filtering estimator with the important difference that no minimum length condition is required for (61). It is interesting to note that we use the above result in the proof of Theorem 2.2.

Acknowledgments. We thank Ryan Tibshirani for informing us about the reference Steidl, Didas and Neumann [34] and for many other helpful comments. We are also extremely thankful to the Associate Editor and the anonymous referees for very detailed comments on an earlier version of the paper. Their feedback greatly improved the quality of the paper.

The first author was supported by NSF CAREER Grant DMS-1654589.

The fourth author was supported by NSF Grant DMS-1150435.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive risk bounds in univariate total variation denoising and trend filtering" (DOI: 10.1214/18-AOS1799SUPP; .pdf). This supplementary material contains additional results and omitted proofs.

REFERENCES

[1] ARNOLD, T. B. and TIBSHIRANI, R. J. (2016). Efficient implementations of the generalized lasso dual path algorithm. *J. Comput. Graph. Statist.* **25** 1–27. MR3474034 https://doi.org/10.1080/10618600.2015.1008638

[2] BELLEC, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

[3] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[4] BROCKMANN, M., GASSER, T. and HERRMANN, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *J. Amer. Statist. Assoc.* **88** 1302–1309. MR1245363

[5] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[6] DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. MR3556784 https://doi.org/10.3150/15-BEJ756

[7] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

[8] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414 https://doi.org/10.1214/aos/1024691081

[9] FAN, Z. and GUAN, L. (2017). l_0-estimation of piecewise-constant signals on graphs. Preprint. Available at arXiv:1703.01421.

[10] FOYGEL, R. and MACKEY, L. (2014). Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Trans. Inform. Theory* **60** 1223–1247. MR3164972 https://doi.org/10.1109/TIT.2013.2293654

[11] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141. With discussion and a rejoinder by the author. MR1091842 https://doi.org/10.1214/aos/1176347963

[12] GAO, C., HAN, F. and ZHANG, C.-H. (2017). Minimax risk bounds for piecewise constant models. Preprint. Available at arXiv:1705.06386.

[13] GOLDENSHLUGER, A. and NEMIROVSKI, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6** 135–170. MR1466625

[14] GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*: *Estimators*, *Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

[15] GUNTUBOYINA, A., LIEU, D., CHATTERJEE, S. and SEN, B. (2020). Supplement to "Adaptive risk bounds in univariate total variation denoising and trend filtering." https://doi.org/10.1214/18-AOS1799SUPP.

[16] GUNTUBOYINA, A. and SEN, B. (2017). Nonparametric shape-restricted regression. Preprint. Available at arXiv:1709.05707.

[17] HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565 https://doi.org/10.1198/jasa.2010.tm09181

[18] JOHNSTONE, I. M. (2015). *Gaussian Estimation*: *Sequence and Wavelet Models*. Available at http://statweb.stanford.edu/~imj/GE09-08-15.pdf.

[19] KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

[20] KUCZMA, M. (2009). *An Introduction to the Theory of Functional Equations and Inequalities*: *Cauchy's Equation and Jensen's Inequality*, 2nd ed. Birkhäuser, Basel. Edited and with a preface by Attila Gilányi. MR2467621 https://doi.org/10.1007/978-3-7643-8749-5

[21] LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734 https://doi.org/10.1214/aos/1069362731

[22] LÉVY-LEDUC, C. and HARCHAOUI, Z. (2008). Catching change-points with lasso. In *Advances in Neural Information Processing Systems* 617–624.

[23] LIN, K., SHARPNACK, J., RINALDO, A. and TIBSHIRANI, R. J. (2016). Approximate recovery in change-point problems, from ℓ_2 estimation error rates. Preprint. Available at arXiv:1606.06746.

[24] MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

[25] MANGASARIAN, O. L. and SCHUMAKER, L. I. (1971). Discrete splines via mathematical programming. *SIAM J. Control* **9** 174–183. MR0304937

[26] MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201. MR0885731 https://doi.org/10.1214/aos/1176350260

[27] ORTELLI, F. and VAN DE GEER, S. (2018). On the total variation regularized estimator over the branched path graph. Preprint. Available at arXiv:1806.01009.

[28] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Found. Comput. Math.* **16** 965–1029. MR3529131 https://doi.org/10.1007/s10208-015-9278-4

[29] PINTORE, A., SPECKMAN, P. and HOLMES, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93** 113–125. MR2277744 https://doi.org/10.1093/biomet/93.1.113

[30] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

[31] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. *Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. MR0274683

[32] RUDELSON, M. and VERSHYNIN, R. (2006). Combinatorics of random processes and sections of convex bodies. *Ann. of Math.* (2) **164** 603–648. MR2247969 https://doi.org/10.4007/annals.2006.164.603

[33] RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

[34] STEIDL, G., DIDAS, S. and NEUMANN, J. (2006). Splines in higher order TV regularization. *Int. J. Comput. Vis.* **70** 241–255.

[35] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[36] TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

[37] VAN DE GEER, S. (2018). On tight bounds for the Lasso. Preprint. Available at arXiv:1804.00989.

[38] WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.* **17** Paper No. 105. MR3543511

[39] WANG, Y.-X., SMOLA, A. J. and TIBSHIRANI, R. J. (2014). The falling factorial basis and its statistical applications. In *ICML* 730–738.

[40] WINKLER, G. and LIEBSCHER, V. (2002). Smoothers for discontinuous signals. *J. Nonparametr. Stat.* **14** 203–222. MR1905594 https://doi.org/10.1080/10485250211388

[41] ZHOU, S. and SHEN, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *J. Amer. Statist. Assoc.* **96** 247–259. MR1952735 https://doi.org/10.1198/016214501750332820