The Annals of Statistics
2021, Vol. 49, No. 5, 2755–2773
https://doi.org/10.1214/21-AOS2054

PREDICTION BOUNDS FOR HIGHER ORDER TOTAL VARIATION REGULARIZED LEAST SQUARES

BY FRANCESCO ORTELLI* AND SARA VAN DE GEER†

*Seminar for Statistics, ETH Zürich, *fortelli@ethz.ch; †vsara@ethz.ch*

We establish adaptive results for trend filtering: least squares estimation with a penalty on the total variation of $(k-1)$th order differences. Our approach is based on combining a general oracle inequality for the ℓ_1-penalized least squares estimator with "interpolating vectors" to upper bound the "effective sparsity." This allows one to show that the ℓ_1-penalty on the kth order differences leads to an estimator that can adapt to the number of jumps in the $(k-1)$th order differences of the underlying signal or an approximation thereof. We show the result for $k \in \{1, 2, 3, 4\}$ and indicate how it could be derived for general $k \in \mathbb{N}$.

1. Introduction. Total variation penalties have been introduced by Rudin, Osher and Fatemi (1992) and Steidl, Didas and Neumann (2006). The present paper builds further on the theory as developed in Tibshirani (2014), Wang et al. (2016), and Guntuboyina et al. (2020). We show, for $k \in \{1, 2, 3, 4\}$, a method for proving that the kth order total variation regularized least squares estimator adapts to the number of jumps in the $(k-1)$th order differences and indicate how this method could be generalized to any $k \in \mathbb{N}$. Inspired by Candès and Fernandez-Granda (2014), our main tool is a well-chosen vector interpolating the signs of the jumps.

The estimation method we will study is known as trend filtering. See Tibshirani (2020) for a comprehensive overview and connections. Trend filtering is a special case of the lasso (Tibshirani (1996)): it is least squares estimation with an ℓ_1-penalty on a subset of the coefficients. For trend filtering, the minimization problem can also be formulated as a so-called analysis problem (Elad, Milanfar and Rubinstein (2007)) with the analysis matrix D being the kth order differences operator (see equation (2)). Our main result, given in Theorem 1.1, is based on an oracle inequality for the general analysis problem with arbitrary analysis matrix $D \in \mathbb{R}^{m \times n}$, as given in Theorem 2.2. The latter is a modification of results in Dalalyan, Hebiri and Lederer (2017): we generalize their projection arguments by allowing for adding "mock" variables to the active set. We furthermore use an improved version of their "compatibility constant" (see Remark 2.3) and—up to scaling—refer to its reciprocal as "effective sparsity"; see Definition 2.1. The effective sparsity for the lasso problem is the number of active parameters (the sparsity) discounted by a factor due to correlations between variables. This discounting factor is called the compatibility constant (see Remark 2.3). In our situation, the effective sparsity can be dealt with invoking what we call an "interpolating vector" (see Definition 2.3), which can be seen as a quantified noisy version of the so-called dual certificate used in basis pursuit. See Remark 2.4 for more details.

Consider an n-dimensional Gaussian vector of independent observations $Y \sim \mathcal{N}_n(f^0, I)$ with unknown mean vector $f^0 \in \mathbb{R}^n$, and with known variance $\text{var}(Y_i) = 1$, $i = 1, \ldots, n$ (see Remark 2.3 for the case of unknown variance). All our results also hold for $Y - f^0$ having independent sub-Gaussian entries with known sub-Gaussian parameter 1. More details on the

Received July 2020; revised January 2021.
MSC2020 subject classifications. Primary 62J05; secondary 62J99.
Key words and phrases. Oracle inequality, projection, compatibility, lasso, analysis, total variation regularization, minimax, Moore–Penrose pseudo inverse.

sub-Gaussian case can be found in Remark A.1 in the Supplementary Material (Ortelli and van de Geer (2021)).

Let $D \in \mathbb{R}^{m\times n}$ be a given matrix. The analysis estimator is

$$\hat{f} := \arg\min_{f\in\mathbb{R}^n} \{\|Y - f\|_n^2 + 2\lambda\|Df\|_1\}, \tag{1}$$

where we invoke the (abuse of) notation $\|v\|_n^2 := \sum_{i=1}^n v_i^2/n$, $v \in \mathbb{R}^n$. We call $D \in \mathbb{R}^{m\times n}$ the analysis matrix. The general aim is to show that $\hat{f}$ is close to the mean $f^0 := \mathbb{E}Y$ of Y, or to some approximation $f \in \mathbb{R}^n$ thereof that has $\|Df\|_0$ "small."

The trend filtering problem has as analysis matrix D the kth order differences operator $\Delta(k) \in \mathbb{R}^{(n-k)\times n}$, which is defined as

$$\Delta(k)_{ij} := \begin{cases} (-1)^l \binom{k}{l}, & j = i - l, l \in [0:k], i \in \mathcal{D}, \\ 0, & \text{else}, \end{cases} \tag{2}$$

where $\mathcal{D} = [k+1:n]$ and $k \in [1:n-1]$ is fixed. We alternatively call $\Delta(k)$ the kth order discrete derivative operator. Moreover, we apply the notation

$$[a:b] = \{j \in \mathbb{N} : a \le j \le b\}, \quad 0 \le a \le b < \infty.$$

The kth order differences operator $\Delta(k)$ can also be obtained by a recursive relation as the product of k first-order difference operators of suitable dimensions. The recursive relation is that to obtain kth order differences for $k \ge 2$ we take the first-order differences of the $(k-1)$th order differences.

Theorem 2.2 below presents results for the general analysis problem and we apply it in Theorem 1.1 to the trend filtering problem. This application means that we need to introduce a "dictionary" as described in Section 2.2, to bound the lengths of the dictionary vectors, and finally calculate an interpolating vector to obtain a bound for the effective sparsity. We do the calculations for $k \in \{1, 2, 3, 4\}$ and sketch the way to proceed for general $k \in \mathbb{N}$.

1.1. *Related work.* Total variation regularization and trend filtering have been studied from different angles in a variety of papers. The paper Mammen and van de Geer (1997) studies numerical adaptivity and rates of convergence. In Kim et al. (2009), it is shown that interior point methods work well for trend filtering. The paper Tibshirani (2014) clarifies connections with splines and also has minimax rates. In Wang et al. (2016), trend filtering on graphs is examined and it has theoretical error bounds in terms of the ℓ_1-norm $\|Df\|_1$. The paper Sadhanala and Tibshirani (2019) contains theory for additive models with trend filtering. The paper Padilla and Chatterjee (2020) extends to quantile regression the idea of trend filtering. In Sadhanala et al. (2017), trend filtering in higher dimensions is studied and minimax rates are proved. The paper Chatterjee and Goswami (2019) proposes a recursive partitioning scheme for higher dimensional trend filtering. Our work is closely related to the paper Guntuboyina et al. (2020), which concerns the constrained problem as well as the penalized problem. Our results for the penalized problem with $k \in \{2, 3, 4\}$ improve those in Guntuboyina et al. (2020), up to log terms. As a special case, we derive that under a "minimum length condition" saying that the distances between jumps of the $(k-1)$th discrete derivative are all of the same order, and under an appropriate condition on the tuning parameter λ, the prediction error of the penalized least squares estimator is of order $(s_0+1)\log(n/(s_0+1))\log n/n$ where s_0 is the number of jumps of $\Delta(k-1)f^0$ (see Corollary 1.2). For s_0 growing at least as $\log^{\frac{1}{2k-1}} n$, this is an improvement on Corollary 2.13 in Guntuboyina et al. (2020) for $k \ge 2$. There, for $\lambda/n^{k-1} \asymp \sqrt{\log(n/(s_0+1))/n}$, the rate is

shown to be $(s_0+1)^{2k}\log(n/(s_0+1))/n$. In fact, we show a more general result where f^0 may be replaced by a sparse approximation. For $k=1$, we show the result with a superfluous log-factor: it is known that in that case the rate of convergence for the prediction error is of order $(s_0+1)\log(n/(s_0+1))/n$; see Guntuboyina et al. (2020) and its references. Our extra log-factor is due to the use of projection arguments instead of more refined empirical process theory. In van de Geer (2020), it is shown that the log-factor for $k=1$ can be removed when invoking entropy arguments, while keeping the approach via interpolating vectors and effective sparsity.

The approach with interpolating vectors is in our view quite natural and lets itself be extended to other problems. We discuss this briefly in the concluding section, Section 4.

1.2. *Organization of the paper.* In the next subsection, Section 1.3, we present in Theorem 1.1 an adaptive result for trend filtering, where adaptivity means that the presented bound for the prediction error can be smaller when f^0 can be well approximated by a vector with fewer jumps in its $(k-1)$th discrete derivative. Section 2 presents in Theorem 2.2 adaptive and nonadaptive bounds for the general analysis problem, which will be our starting point for proving Theorem 1.1. We introduce effective sparsity and interpolating vectors in Definitions 2.1 and 2.3.

Section 3 applies the general result of Theorem 2.2 to the case $D=\Delta(k)$. We then need to introduce a projected dictionary for trend filtering, which is done in Section 3.1. With this we arrive at nonadaptive, almost minimax rates in Theorem 3.2. In Section 3.3, we construct interpolating vectors and bounds for the effective sparsity for the case $k\in\{1,2,3,4\}$ and also sketch how this can be done for general k. With these results in hand, we finish in Section 3.4 the proof of the adaptive bounds for trend filtering with $k\in\{1,2,3,4\}$. Section 4 concludes the paper.

The Supplementary Material (Ortelli and van de Geer (2021)) contains a proof of Theorem 2.2. Its arguments are to a large extent in Dalalyan, Hebiri and Lederer (2017) and Ortelli and van de Geer (2020b), but there are modifications. The Supplementary Material also has the proofs for Section 3.1 and 3.3.

1.3. *Main result for trend filtering.* For $D=\Delta(k)$ and $\mathcal{D}=[k+1:n]$, we let $S=\{t_1,\ldots,t_s\}\subseteq\mathcal{D}$, $k+1\leq t_1<\cdots<t_s\leq n$ and let $t_0:=k$ and $t_{s+1}:=n+1$. We define $n_i=t_i-t_{i-1}$, $i\in[1:s+1]$ and $n_{\max}:=\max_{1\leq i\leq s+1}n_i$. Moreover, for $f\in\mathbb{R}^n$ we write $(\Delta(k)f)_{-S}:=\{(\Delta(k)f)_j:j\in\mathcal{D}\backslash S\}$.

In Theorem 1.1 below, the set S is fixed but arbitrary. The theorem presents an oracle inequality that allows for a trade-off between approximation error and estimation error by choosing S and f appropriately, depending on the unknown f^0. However, the tuning parameter will then depend on s. Remark 1.5 reverses this viewpoint.

Write for $u>0$,

$$\lambda_0(u):=\sqrt{\frac{2\log(2(n-k-s))+2u}{n}}.$$

THEOREM 1.1 (Adaptive rates for $k=1,2,3,4$). *Let $k\in\{1,2,3,4\}$. There exists constants c_k and C_k depending only on k such that the following holds.*

Let $u>0$ be arbitrary and choose the tuning parameter λ satisfying

$$\lambda\geq c_k n^{k-1}\left(\frac{n_{\max}}{2n}\right)^{\frac{2k-1}{2}}\lambda_0(u).$$

Let $f\in\mathbb{R}^n$ be arbitrary and define the signs

$$q_{t_i}:=\operatorname{sign}(Df)_{t_i},\quad i=[1:s].$$

Write $S^{\pm} := \{i \in [2:s] : q_{t_i} q_{t_{i-1}} = -1\} \cup \{1, s+1\}$. *Assume* $n_i \geq k(k+2)$ *for all* $i \in S^{\pm}$. *Finally, let* $v > 0$ *be arbitrary. Then with probability at least* $1 - e^{-u} - e^{-v}$ *we have*

$$\|\hat{f} - f^0\|_n^2 \leq \underbrace{\|f - f^0\|_n^2}_{\text{"approximation error"}} + 4\lambda \|(\Delta(k) f)_{-S}\|_1 + \underbrace{\left(\sqrt{\frac{k(s+1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma_S\right)^2}_{\text{"estimation error"}},$$

where

$$\Gamma_S^2 = n C_k \left(\sum_{i \in S^{\pm}} \frac{1 + \log n_i}{n_i^{2k-1}} + \sum_{i \in S \setminus S^{\pm}} \frac{1 + \log n_i}{n_{\max}^{2k-1}} \right). \tag{3}$$

To prove this result, we will invoke Theorem 2.2. This requires providing a dictionary and bounding the effective sparsity given by Definition 2.1. In Section 3.4, we then put the pieces together.

REMARK 1.1. The quantity Γ_S^2 in the above theorem is a bound for the effective sparsity.

REMARK 1.2. One may take $c_1 = c_2 = 2$. For $\min_{i \in S^{\pm}} n_i \to \infty$, asymptotic expressions for c_3 and c_4 can be taken to be $c_3 \to 19/2$ and by numerical computation, $c_4 \to 2 \times 6^{7/2}/(18.62) \approx 56.83$. See Section 3.3.1.

REMARK 1.3. Our method of proof is along the lines of Dalalyan, Hebiri and Lederer (2017). In Ortelli and van de Geer (2020b), it is shown that one can also use this method for the square-root lasso. This means that as a corollary of Ortelli and van de Geer (2020b), our result also hold for "square-root" trend filtering, with a choice of the tuning parameter that does not depend on the variance of the noise.

REMARK 1.4. In Section 3.3.1, we indicate how the bound of Theorem 1.1 could be established for general $k \in \mathbb{N}$.

We formulate a corollary for the case where the distances between jumps are all of the same order as the maximal distance $n_{\max}$. To facilitate the statement, we give an asymptotic formulation. For sequences $\{a_n\}$ and $\{b_n\}$ in $(0, \infty)$, we use the notation $a_n = \mathcal{O}(b_n)$ if $\limsup_{n \to \infty} a_n / b_n < \infty$ and $a_n \asymp b_n$ (or equivalently $a_n = \Theta(b_n)$) if also $b_n / a_n = \mathcal{O}(1)$. For a sequence of random variables $\{Z_n\}$, we write $Z_n = \mathcal{O}_{\mathbb{P}}(1)$ if $\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P}(|Z_n| > M) = 0$.

COROLLARY 1.2. *Fix* $k \in \{1, 2, 3, 4\}$. *Let* $u \asymp v \asymp \log n$. *Choose* S *such that*

$$\min_{i \in [1:s+1]} n_i \asymp n_{\max}.$$

Then we can choose λ *of order*

$$\lambda \asymp n^{k-1} \left(\frac{1}{s+1} \right)^{\frac{2k-1}{2}} \sqrt{\frac{\log n}{n}}$$

and for all $f \in \mathbb{R}^n$, *under this choice, with probability* $1 - \Theta(1/n)$ *it holds that*

$$\|\hat{f} - f^0\|_n^2 \leq \|f - f^0\|_n^2 + 4\lambda \|(\Delta(k) f)_{-S}\|_1 + \mathcal{O}\left(\frac{s+1}{n} \log\left(\frac{n}{s+1} \right) \log n \right).$$

REMARK 1.5. Theorem 1.1 holds for an arbitrary active set S and gives a theoretical justification for choosing a smaller tuning parameter λ than the universal choice $\lambda/n^{k-1} \asymp \lambda_0(u)$. The tuning parameter λ depends on the maximal distance $n_{\max}$ between jumps in the active set S. One can therefore incorporate eventual prior knowledge about the true active set in the tuning via $n_{\max}$.

In practice, we can also think the other way around. Given a choice of λ (which implicitly means a choice of $n_{\max}$), the theorem will hold for a restricted group of active sets: the ones characterized by

$$n_{\max} \leq 2\left(\frac{\sqrt{n}\lambda}{c_k \lambda_0(u)}\right)^{\frac{2}{2k-1}}.$$

One can even set $n_{\max} = n$ in the tuning parameter, and thus choose $\lambda/n^{k-1} \geq c_k 2^{-\frac{2k-1}{2}} \lambda_0(u)$ independently of S. The upper bound of Theorem 1.1 can then accommodate all S but the rate is worse.

We face a tradeoff in the choice of λ. The choice of λ (and thus of $n_{\max}$) determines the set of active sets over which the upper bound could be optimized over, but also influences the rate of the estimation error. Choosing a smaller λ, that is, a smaller $n_{\max}$—results in fewer admitted active sets S but potentially a faster rate in the estimation error.

Alternatively, the tuning parameter could be selected by sample splitting.

REMARK 1.6. Theorem 1.1 improves on Corollary 2.13 by Guntuboyina et al. (2020) by replacing $(s_0+1)^{2k}$ with $(s_0+1)\log n$. This is possible because of the increased flexibility in the choice of the tuning parameter λ. Moreover, we allow for a sparse approximation f of f^0 and present a sharp oracle inequality. The paper Guntuboyina et al. (2020), also studies the constrained problem where they arrive at rates which are up to log-terms comparable to ours for the regularized problem when taking $f = f^0$.

Suppose we are in the setting of Corollary 1.2. Corollary 2.13 by Guntuboyina et al. (2020) requires $\lambda/n^{k-1} \asymp \sqrt{\log(n/(s+1))/n}$ to obtain the rate

$$\frac{(s_0+1)^{2k}}{n} \log\left(\frac{n}{s_0+1}\right).$$

If we choose $\lambda/n^{k-1} \asymp \lambda_0(\log n) \asymp \log n$ in Theorem 1.1, we find the rate

$$\frac{(s+1)^{2k}}{n} \log\left(\frac{n}{s+1}\right) \log n$$

and we retrieve with $f = f^0$ and $S = S^0$ the rate by Guntuboyina et al. (2020) up to a log term. But the requirement on our tuning parameter is more flexible. By allowing for a smaller tuning parameter, we replace $(s+1)^{2k}$ by $(s+1)\log n$. Our improved rate has still an additional logarithmic term and is therefore at least as good as the one by Guntuboyina et al. (2020) only if s_0 grows at least as $\log^{\frac{1}{2k-1}} n$.

It in fact suffices to choose $\lambda_0(u)$ of order $\sqrt{\log\log n/n} + \sqrt{\log s/n}$ instead of $\sqrt{\log n/n}$ This follows from the fact that the basis functions used to construct the trend filtering estimator (cf. Section 3.1) form a VC class and from using bounds for weighted empirical processes.

To remove the extra log-term in our result altogether, the paper van de Geer (2020) uses a strategy to transform the Euclidean norm to a weighted Euclidean norm and then applies entropy bounds. This allows one to move the log term coming from $\lambda_0(u)$ to the term involving the dimension of the space we project on. The paper van de Geer (2020) applies this technique for the case $k = 1$. It exploits the fact that for a space of functions with total variation

bounded by a constant the entropy is of the same order for all L_2-norms. We are not sure whether this is the case for $k > 1$. The proof technique in van de Geer (2020) still shares the main idea using effective sparsity, but is definitely more involved and leads to many large and unspecified constants (coming from the entropy bounds, Dudley's entropy integral and from bounding weighted empirical processes).

2. Adaptive bounds for the general analysis estimator. Recall the analysis problem

$$\hat{f} := \arg\min_{f \in \mathbb{R}^n} \{\|Y - f\|_n^2 + 2\lambda \|Df\|_1\}, \tag{4}$$

where $D \in \mathbb{R}^{m \times n}$ is a given analysis matrix, $\lambda > 0$ is a tuning parameter and $\|v\|_n^2 := \|v\|_2^2/n$, $v \in \mathbb{R}^n$.

To be able to state Theorem 2.2—a modification of results in Dalalyan, Hebiri and Lederer (2017) (see Remark 2.3)—we introduce some notation in Section 2.1, and then describe the dictionary (Section 2.2) and the effective sparsity (Section 2.3). Theorem 2.2 can then be found in Section 2.4. To apply it, one needs to upper bound the effective sparsity. This is done in Lemma 2.4, which invokes interpolating vectors as defined in Definition 2.3 of Section 2.5. Theorem 2.2 and Lemma 2.4 combined serve as starting point for proving the result for trend filtering in Theorem 1.1.

2.1. *Some notation.* The rows of the analysis matrix $D \in \mathbb{R}^{m \times n}$ are indexed by a set $\mathcal{D}$ of size $|\mathcal{D}| = m$. We consider a set $S \subseteq \mathcal{D}$, which is arbitrary and can be chosen as the active set of an "oracle" that trades off "approximation error" and "estimation error" (see Theorem 2.2). The size of S is denoted by $s := |S|$. We define for a vector $b_{\mathcal{D}} \in \mathbb{R}^m$ with index set $\mathcal{D}$,

$$b_S := \{b_j\}_{j \in S}, \qquad b_{-S} := \{b_j\}_{j \in \mathcal{D} \setminus S}.$$

We let $\mathcal{N}_{-S} := \{f \in \mathbb{R}^n : (Df)_{-S} = 0\} = \{f \in \mathbb{R}^n : (Df)_j = 0 \ \forall j \in \mathcal{D} \setminus S\}$ and write $r_S := \dim(\mathcal{N}_{-S})$. As benchmark for our result, consider the active set $S_0 := \{j : (Df^0)_j \neq 0\}$. If S_0 were known, the least squares estimator

$$\hat{f}_{\mathrm{LSE}} := \arg\min_{f \in \mathcal{N}_{-S_0}} \|Y - f\|_n^2$$

would satisfy, for all $v > 0$, with probability at least $1 - e^{-v}$,

$$\|\hat{f}_{\mathrm{LSE}} - f^0\|_n \leq \sqrt{\frac{r_{S_0}}{n}} + \sqrt{\frac{2v}{n}}.$$

This follows from a concentration bound for chi-squared random variables; see Lemma 1 in Laurent and Massart (2000). An aim is to show that the estimator $\hat{f}$ converges with the same rate $\sqrt{r_{S_0}/n}$, modulo log-factors. In fact, we aim at showing this type of result with f^0 potentially replaced by a sparse approximation. We hope to be able to choose the active set S of a sparse approximation such that r_S is small. On the other hand, as we will see, the distance of the "nonactive" variables to the linear space $\mathcal{N}_{-S}$ will play an important role: the smaller this distance, the less noise is left to be overruled by the penalty. Therefore, we allow for the possibility to extend $\mathcal{N}_{-S}$ to a larger linear space $\bar{\mathcal{N}}_{-S} \supseteq \mathcal{N}_{-S}$. This can be done for instance by adding some "mock" active variables to the active set. We let $\bar{r}_S = \dim(\bar{\mathcal{N}}_{-S})$. Thus, $\bar{r}_S \geq r_S$ but in our application to trend filtering we will choose them of the same order.

For $\mathcal{U}$ and $\mathcal{V}$ being two linear subspaces of $\mathbb{R}^n$ spanned by $\{u_i\}$ and $\{v_j\}$, we define the direct product of $\mathcal{U}$ and $\mathcal{V}$ as the linear space spanned by $\{u_i\} \cup \{v_j\}$.

2.2. *The dictionary.* Given the linear space $\bar{\mathcal{N}}_{-S} \supseteq \mathcal{N}_{-S}$ we can decompose a vector $f \in \mathbb{R}^n$ into its projection $f_{\bar{\mathcal{N}}_{-S}}$ onto $\bar{\mathcal{N}}_{-S}$ and its projection onto the orthocomplement $\bar{\mathcal{N}}_{-S}^{\perp}$, which we call its antiprojection:

$$f = f_{\bar{\mathcal{N}}_{-S}} + f_{\bar{\mathcal{N}}_{-S}^{\perp}}.$$

The antiprojection is the part we want to overrule by the penalty. For this purpose, we define a dictionary $\Psi^{-S} := \{\psi_j^{-S}\}_{j \in \mathcal{D}\setminus S} \in \mathbb{R}^{n \times (m-s)}$ such that for all $f \in \mathbb{R}^n$ and for $b_{-S} = (Df)_{-S}$,

$$f_{\bar{\mathcal{N}}_{-S}^{\perp}} = \Psi^{-S} b_{-S}.$$

In general, there can be several choices for Ψ^{-S}. In the application to trend filtering that we consider in this paper, Ψ^{-S} will be uniquely defined. (When $\bar{\mathcal{N}}_S = \mathcal{N}_{-S}$, it holds that $\Psi^{-S} = D'_{-S}(D_{-S}D'_{-S})^{-1}$ with D_{-S} being the matrix D with the rows indexed by S removed.)

2.3. *Effective sparsity.* The effective sparsity will be invoked to establish adaptive bounds.

Fix some $u > 0$. Its value will occur in the confidence level of the inequalities in Theorem 2.2. We define

$$\lambda_0(u) := \sqrt{\frac{2\log(2(m-s)) + 2u)}{n}}. \tag{5}$$

In what follows, we assume that the tuning parameter λ satisfies

$$\lambda \geq \max_{j \in \mathcal{D}\setminus S} \|\psi_j^{-S}\|_n \lambda_0(u). \tag{6}$$

For a vector w_{-S} with $0 \leq w_j \leq 1$ for all $j \in \mathcal{D}\setminus S$, we write

$$(1 - w_{-S})(Df)_{-S} := \{(1 - w_j)(Df)_j\}_{j \in \mathcal{D}\setminus S}.$$

DEFINITION 2.1 (Effective sparsity). Let $q_S \in \{-1, +1\}^s$ be a sign vector. The noiseless effective sparsity is

$$\Gamma^2(S, q_S) := \big(\max\{q_S'(Df)_S - \|(Df)_{-S}\|_1 : \|f\|_n = 1\}\big)^2.$$

The noisy effective sparsity is

$$\Gamma^2(S, q_S, w_{-S}) := \big(\max\{q_S'(Df)_S - \|(1 - w_{-S})(Df)_{-S}\|_1 : \|f\|_n = 1\}\big)^2,$$

where

$$w_j = \|\psi_j^{-S}\|_n \lambda_0(u)/\lambda, \quad j \in \mathcal{D}\setminus S,$$

with λ satisfying (6).

REMARK 2.1. The effective sparsity can be interpreted as the effective number of parameters one has to estimate. With the universal choice of the tuning parameter $\lambda/n^{k-1} \asymp \lambda_0(u)$, the fast rate is of order $\Gamma^2(S, q_S, w_{-S}) \log n/n$.

For the trend filtering problem, we will derive in Section 3.3 bounds on the noisy effective sparsity that, when $\min_{i \in [s+1]} n_i \asymp n_{\max}$, scale as

$$\Gamma^2(S, q_S, w_{-S}) = \mathcal{O}\big((s+1)^{2k} \log\big(n/(s+1)\big)\big), \quad k \in \{1, 2, 3, 4\}.$$

The log term is due to the noise. As shown in Corollary 1.2, we can choose the tuning parameter smaller than the universal choice $\lambda/n^{k-1} \asymp \lambda_0(u)$ thanks to the projection arguments by Dalalyan, Hebiri and Lederer (2017) in the background. Thus, we can obtain the rate $(s+1)/n$ up to logarithmic terms.

REMARK 2.2. On the slightly negative side, when applying Theorem 2.2 one may need to choose λ strictly larger than (but of the same order as) required in (6) in order to have a "well-behaved" effective sparsity. On the positive side, depending on the situation, one may improve upon $\lambda_0(u)$ in (5) using bounds for weighted empirical processes.

2.4. *Main result for the general analysis problem.* Recall that the set S is arbitrary. In the following theorem, the set S and also its vector $f \in \mathbb{R}^n$ can be chosen to optimize the bounds by trading off approximation error and estimation error. The theorem provides adaptive bounds (oracle inequalities) since the trade-off depends on the unknown signal f^0.

THEOREM 2.2. *Let $u > 0$, $v > 0$ and let the tuning parameter λ satisfy (6). Then $\forall f \in \mathbb{R}^n$ the following bounds hold*:

- *a nonadaptive bound*: *with probability at least $1 - e^{-u} - e^{-v}$,*

$$\|\hat{f} - f^0\|_n^2 \le \|f - f^0\|_n^2 + 4\lambda\|Df\|_1 + \left(\sqrt{\frac{\bar{r}_S}{n}} + \sqrt{\frac{2v}{n}}\right)^2;$$

- *and an adaptive bound*: *with probability at least $1 - e^{-u} - e^{-v}$,*

$$\|\hat{f} - f^0\|_n^2 \le \|f - f^0\|_n^2 + 4\lambda\|(Df)_{-S}\|_1 + \left(\sqrt{\frac{\bar{r}_S}{n}} + \sqrt{\frac{2v}{n}} + \lambda\Gamma(S, q_S, w_{-S})\right)^2,$$

where $q_S = \operatorname{sign}(Df)_S$.

PROOF OF THEOREM 2.2. See Appendix A in the Supplementary Material (Ortelli and van de Geer (2021)). □

REMARK 2.3. Theorem 2.2 is a modification of the findings by Dalalyan, Hebiri and Lederer (2017) who study the lasso problem. Theorem 2.2 is in terms of the analysis problem, as in Ortelli and van de Geer (2020b). We furthermore allow for augmentation of $\mathcal{N}_{-S}$. Moreover, Dalalyan, Hebiri and Lederer (2017) and Ortelli and van de Geer (2020b) replace $\Gamma(S, q_S, w_{-S})$ by the larger quantity $\max\{\|(Df)_S\|_1 - \|(1 - w_{-S})(Df)_{-S}\|_1 : \|f\|_n = 1\} =: \sqrt{r_S}/\kappa(S, w_{-S})$ where $\kappa^2(S, w_{-S})$ is called the "compatibility constant." The paper Ortelli and van de Geer (2020b) derives oracle results for the square-root analysis problem, which is the analysis version of the square-root lasso introduced by Belloni, Chernozhukov and Wang (2011). Joining these findings allows to derive a square-root version of Theorem 2.2, which can be applied to the case of unknown noise variance (see also Remark 1.3).

2.5. *Interpolating vectors.* To upper bound the effective sparsity, one may invoke interpolating vectors.

DEFINITION 2.3 (Interpolating vector). Let $q_S \in \{-1, +1\}^s$ be a sign vector. We call the completed vector $q \in \mathbb{R}^m$ with index set $\mathcal{D}$ an interpolating vector (that interpolates the given signs at S).

LEMMA 2.4. *Let $q_S \in \{-1, +1\}^s$ be a sign vector. The noiseless effective sparsity $\Gamma^2(S, q_S)$ satisfies*

$$\Gamma^2(S, q_S) \le \inf\{n\|D'q\|_2^2 : q \text{ interpolating}, |q_j| \le 1 \ \forall j \in \mathcal{D}\backslash S\}.$$

The noisy effective sparsity $\Gamma^2(S, q_S, w_{-S})$ *satisfies*

$$\Gamma^2(S, q_S, w_{-S}) \leq \inf\{n\|D'q\|_2^2 : q \text{ interpolating}, |q_j| \leq 1 - w_j \ \forall j \in \mathcal{D}\backslash S\},$$

where

$$w_j = \|\psi_j^{-S}\|_n \lambda_0(u)/\lambda, \quad j \in \mathcal{D}\backslash S,$$

with λ *satisfying* (6).

PROOF OF LEMMA 2.4. We only prove the statement of the lemma for the noisy case as the argument is the same for the noiseless case. For any vector q_{-S} with $|q_j| \leq 1 - w_j$ for all $j \in \mathcal{D}\backslash S$, it is true that for all f,

$$\|(1 - w_{-S})(Df)_{-S}\|_1 \geq q_{-S}(Df)_{-S}.$$

Therefore, for all f,

$$q_S'(Df)_S - \|(1 - w_{-S})(Df)_{-S}\|_1 \leq q'Df \leq \sqrt{n}\|D'q\|_2\|f\|_n. \qquad \square$$

REMARK 2.4. To bound the effective sparsity invoking interpolating vectors, as the above lemma does, we were inspired by the dual certificates as applied in Candès and Fernandez-Granda (2014). These have also been used in other works. The paper Candès and Fernandez-Granda (2014) considers the superresolution problem and develops an interpolating function to establish exact recovery in the noiseless problem. The requirement on this interpolating function is that it is in the range of the transpose of the design matrix. Dual certificates can be found in earlier work as well; see, for example, Candès and Plan (2011). The latter applies a "near" dual certificate to deal with noisy measurements. However, their "nearness" appears to be restricted to a special setting. In Tang, Bhaskar and Recht (2015), the approach is related to ours but very much tied down to the noisy superresolution problem. We are not aware of any work where an explicit connection is made between dual certificates and compatibility constants, the latter being related to the reciprocal of effective sparsity; see Remark 2.3. The relation between dual certificates and interpolating vectors on the one hand and compatibility on the other hand, appears to have been hidden in the literature. Moreover, the notion of compatibility has developed itself over the years. In, for example, Boyer, De Castro and Salmon (2017), it is shown that compatibility conditions do not hold for superresolution, but they show it for an older version of compatibility, not for the newer version based on $\kappa^2(S, w_{-S})$.

3. Application of Theorem 2.2 when $D = \Delta(k)$. In order to apply Theorem 2.2 with $D = \Delta(k)$, we need to establish a bound for the length of the columns of an appropriate dictionary Ψ^{-S}. This is done in Section 3.1. Then we can apply the first part of Theorem 2.2 and this will, as we will see in Section 3.2, result in the minimax rate up to log-terms. Next, for the application of the second part of Theorem 2.2 to obtain adaptive results, we upper bound the effective sparsity using an appropriate interpolating vector. This is done in Section 3.3. We then have all the material to establish Theorem 1.1, as summarized in Section 3.4.

3.1. *The dictionary when* $D = \Delta(k)$. We start with some remarks, whose purpose is mainly to introduce some further notation. Note that by definition, $\mathcal{N}_\mathcal{D} := \{f \in \mathbb{R}^n : (Df)_j = 0 \ \forall j \in \mathcal{D}\}$. For vectors $f_{\mathcal{N}_\mathcal{D}}^\perp$, a dictionary is denoted by $\Psi^\mathcal{D}$. If D has full row rank (which will be the case for $D = \Delta(k)$ see Wang, Smola and Tibshirani (2014)) it holds that $\Psi^\mathcal{D} = D'(DD')^{-1}$. This is the Moore–Penrose pseudo inverse D^+ of D. Let now $\Psi = \{\psi_j\}_{j=1}^n \in \mathbb{R}^{n\times n}$ be a "complete" dictionary, which means that we can write each $f \in \mathbb{R}^n$

as $f = \Psi b$, where $b_{\mathcal{D}} = Df$. Then obviously $\Psi^{\mathcal{D}} = \{(\psi_j)_{\mathcal{N}_{\mathcal{D}}^{\perp}}\}_{j\in\mathcal{D}}$ is formed by the projections of the dictionary vectors ψ_j with index in $\mathcal{D}$ on the ortho-complement of the space spanned by $\{\psi_{j'}\}_{j'\notin\mathcal{D}}$. Moreover, the dictionary $\Psi^{-S} = \{\psi_j^{-S}\}_{j\in\mathcal{D}\setminus S}$ for vectors $f_{\mathcal{N}_{-S}^{\perp}}$ has $\psi_j^{-S} = (\psi_j)_{\mathcal{N}_{-S}^{\perp}} = (\psi_j^{\mathcal{D}})_{\mathcal{N}_{-S}^{\perp}}$.

As said, we may want to augment the space $\mathcal{N}_{-S}$ to a larger linear space $\bar{\mathcal{N}}_{-S}$ so that the antiprojections will have smaller length. This is done by taking the direct product of $\mathcal{N}_{-S}$ with a space spanned by additional linearly independent vectors $\{\phi_j\}$ with $\phi_j \notin \mathcal{N}_{-S}$ for all j. We call these additional vectors "mock" variables. One may pick them in the set $\{\psi_j\}_{j\in\mathcal{D}\setminus S}$ (or $\{\psi_j^{\mathcal{D}}\}_{j\in\mathcal{D}\setminus S}$) in which case we call them "mock" active variables.

We now first present upper bounds for $\{\|\psi_j\|_2^2\}_{j\in\mathcal{D}\setminus S}$ for the case $D = \Delta(k)$ for general k. In the Supplementary Material (Ortelli and van de Geer (2021)), we give exact expressions for $k \in \{1, 2, 3\}$ to illustrate that the corresponding bounds are sharp.

3.1.1. *The dictionary for general k*: *Upper bounds*. We take $\bar{\mathcal{N}}_{-S}$ as the direct product of $\mathcal{N}_{-S}$ and the space spanned by $\{\psi_{t_i+1}, \ldots, \psi_{t_i+k-1}\}_{i=1}^{s}$ (assuming $t_s + k - 1 \le n$). In this way, we disconnect the system into $s + 1$ components each having the same structure. The matrix D with the rows indexed by $S \cup \{t_i + 1, \ldots, t_i + k - 1\}_{i=1}^{s}$ removed is a block matrix. The space $\bar{\mathcal{N}}_{-S}$ has dimension $\bar{r}_S = r_S + s(k-1) = k(s+1)$. One may apply a reformulation for the subintervals $\{[t_{i-1} + 1 : t_i - 1]\}_{i=1}^{s+1}$, to arrive at upper bounds for the lengths of the columns of Ψ^{-S} from upper bounds for the lengths of the columns of $\Delta(k)^+$.

LEMMA 3.1 (An upper bound for the length of the columns of $\Delta(k)^+$). *We have that for* $j \in [k+1 : n]$

$$\|\psi_j^{\mathcal{D}}\|_2^2 \le \min\big((j-k)^{2k-1}, (n+1-j)^{2k-1}\big).$$

PROOF OF LEMMA 3.1. See Appendix B in the Supplementary Material (Ortelli and van de Geer (2021)). □

It follows that

$$\|\psi_{t_{i-1}+j}^{-S}\|_2^2 \le \min\big(j^{2k-1}, (n_i - j)^{2k-1}\big), \quad j \in [1 : n_i - 1], i = [1 : s+1]. \tag{7}$$

Recall our abuse of notation $\|\cdot\|_n^2 = \|\cdot\|_2^2/n$. We get

$$\max_{j\in[1:n_i-1]} \|\psi_{t_{i-1}+j}^{-S}\|_n^2 \le \frac{(n_i/2)^{2k-1}}{n}, \quad i = [1 : s+1].$$

We conclude that the requirement (6) on the tuning parameter is met when

$$\lambda \ge n^{k-1}\left(\frac{n_{\max}}{2n}\right)^{\frac{2k-1}{2}} \sqrt{\frac{2\log(2(n-s-k)) + 2u}{n}}. \tag{8}$$

3.2. *An almost minimax rate*. Although minimax rates are not the main theme of this paper, we present a result in this direction because it comes almost for free.

THEOREM 3.2. *Let* $u > 0$, $v > 0$ *and let the tuning parameter* λ *satisfy* (8). *Then it holds that* $\forall f \in \mathbb{R}^n$, *with probability at least* $1 - e^{-u} - e^{-v}$,

$$\|\hat{f} - f^0\|_n^2 \le \|f - f^0\|_n^2 + 4\lambda\|Df\|_1 + \left(\sqrt{\frac{k(s+1)}{n}} + \sqrt{\frac{2v}{n}}\right)^2.$$

PROOF OF THEOREM 3.2. This follows from applying the first, that is, the nonadaptive result of Theorem 2.2, and invoking that its requirement (6) on the tuning parameter λ is met if we impose the bound given in (8). $\square$

COROLLARY 3.3. *Recall that requirement* (8) *on* λ *depends on* S *and the choice of* S *is free in Theorem* 3.2. *We can take* S *such that* $\min_{i\in[1:s+1]} n_i \asymp n_{\max}$ *in which case* $n_{\max}/n \asymp 1/s$. *Then we may take*

$$\lambda \asymp n^{k-1}\left(\frac{1}{s}\right)^{\frac{2k-1}{2}}\sqrt{\frac{\log n}{n}}.$$

Now s *is still a free parameter. Choosing* s *by a trade off, that is, in such a way that* $\lambda/n^{k-1} \asymp s/n$, *we get for* $n^{k-1}\|\Delta(k)f\|_1 \le 1$ (*say*)

$$\|\hat{f} - f^0\|_n^2 \le \|f - f^0\|_n^2 + \mathcal{O}_{\mathbb{P}}\big(n^{-\frac{2k}{2k+1}}\log^{\frac{1}{2k+1}} n\big).$$

For $f = f^0$, *this corresponds, up to the log-factor, to the minimax rate over* $\{f^0 : n^{k-1}\|\Delta(k)f^0\|_1 \le 1\}$ (*Donoho and Johnstone* (1998)).

REMARK 3.1. To obtain Corollary 3.3 from Theorem 3.2, we choose s not depending on $n^{k-1}\|\Delta(k)f\|_1$ by trading off $\lambda/n^{k-1} \asymp s/n$. Thus, the choice of the tuning parameter λ dictated by Corollary 3.3 does not depend on $n^{k-1}\|\Delta(k)f\|_1$.

Alternatively, we can choose s depending on $n^{k-1}\|\Delta(k)f\|_1$ by trading off $\lambda\|\Delta(k)f\|_1 \asymp s/n$. Then the choice of the tuning parameter λ does depend on $n^{k-1}\|\Delta(k)f\|_1$. If $n^{k-1}\|\Delta(k)f\|_1 = \mathcal{O}(1)$, the rates obtained are the same.

It is remarkable that f can be chosen arbitrarily. Therefore, the upper bound holds for the infimum over all f. If we tune the estimator depending on $n^{k-1}\|\Delta(k)f\|_1$, then the upper bound holds for the infimum over all f with the chosen value for $n^{k-1}\|\Delta(k)f\|_1$.

By applying the idea of the square root lasso (Belloni, Chernozhukov and Wang (2011)) to analysis estimators as in Ortelli and van de Geer (2020b), we can tune the square root version of trend filtering independently of the noise variance σ^2.

3.3. *Interpolating vectors and effective sparsity for* $D = \Delta(k)$. Observe that for $D = \Delta(k)$ it holds that $(D'q)_{k+j} = ((-1)^k\Delta(k)q)_j$ for $j \in [1 : n - 2k]$ (this is in the background of partial integration). The above observation leads in the noiseless case to taking q as piecewise kth degree polynomial interpolation. To avoid boundary effects, one may use an interpolation including the points $t_0 := k$ and $t_{s+1} := n + 1$, with $q_{t_0} = q_{t_{s+1}} = 0$. Moreover, still in the noiseless case, if there is no sign change (i.e., $q_{t_{i-1}}q_{t_i} = 1$) one can simply take $q_{t_{i-1}+j} = q_{t_i}$ for $j \in [1 : n_i - 1]$.

In the noisy case, one can use a similar interpolation except near the active points in S, where we need to change to powers of $(2k-1)/2$ instead of k. This is due to the requirement $|q_j| \le 1 - w_j$ for all $j \in \mathcal{D}$ from Lemma 2.4 and to the form of the weights following from (7). It has an effect on the constants involved in the interpolating vector and, moreover, for $t_{i-1} + j$ near the active points, the absolute kth discrete derivative $|\Delta(k)q_{t_{i-1}+j}|$ behaves like $1/\sqrt{j}$. It follows from the next lemma that this leads to an additional logarithmic factor as compared to the noiseless case.

LEMMA 3.4. *Let for some* $d \in \mathbb{N}$, $d \ge 2k$,

$$\mathbf{q}_j := (j)^{\frac{2k-1}{2}}, \quad j = 0, \dots, d.$$

Then for some constant $\tilde{C}_k$,

$$\|\Delta(k)'\mathbf{q}\|_2^2 \le \tilde{C}_k^2(1 + \log d).$$

PROOF OF LEMMA 3.4. See Appendix C in the Supplementary Material (Ortelli and van de Geer (2021)). □

For each given k the interpolating vector, we suggest below can be computed by solving a system of linear equations. We construct an interpolating vector giving place to a suitable bound on the effective sparsity by joining sufficiently many polynomial pieces in a smooth enough way. The required smoothness depends on the order of the differences considered in the trend filtering problem. For kth order differences, we match the polynomial pieces and their first $k-1$ discrete derivatives. Matching discrete derivatives gives a certain number of equations that the coefficients of the polynomial pieces of the interpolating vector have to satisfy. Both the number of equations and the number of coefficients depend on the number of pieces into which we split an interval, that is, on the number of different polynomial pieces we allow. The number of splits is then obtained by equating the number of equations with the number of coefficients. Because we work in the discrete setting, these splits need to contain sufficiently many points: at least k each. Indeed, matching the $k-1$ discrete derivatives of two polynomials at a given point corresponds to matching the two polynomials at the given point and at the $k-1$ previous ones.

Once we have computed an interpolating vector q by joining polynomial pieces by derivatives matching, we still have to check that it satisfies the requirements to be an interpolating vector. That is, it must hold $q_j \le 1 - w_j = 1 - \|\psi_j^{-S}\|_n \lambda_0(u)/\lambda$ for all $j \in \mathcal{D}\backslash S$ and $|q_j| \le 1$ for all $j \in \mathcal{D}$. Sufficient conditions are that the resulting q is monotone and λ is chosen large enough. Indeed, if $\lambda/\lambda_0(u)$ is large, the weights w become small and the requirements on the interpolating vector q become weaker. This can be read from the formula for w in Definition 2.1 and Lemma 2.4.

It is not clear whether the interpolation is monotone between two active points. We verify the monotonicity for $k = \{1, 2, 3, 4\}$ in Sections 3.3.2, 3.3.3, 3.3.4 and 3.3.5, respectively. In other words, Section 3.3.1 presents the general idea, and the four following subsections work out the details for $k \in \{1, 2, 3, 4\}$.

3.3.1. *Construction of an interpolating vector.* Define

$$S^{\pm} := \{i \in [2:s] : q_{t_i} q_{t_i - 1} = -1\} \cup \{1, s+1\}$$

and let $t_0 = k$ and $t_{s+1} = n+1$. We call $[t_0 : t_1]$ the left boundary interval and $[t_s : t_{s+1}]$ the right boundary interval. We assume that

$$n_i \ge k(k+2) \quad \forall i \in S^{\pm}. \tag{9}$$

For $i \in S^{\pm}$, we split $[t_{i-1} : t_i]$ into $k+2$ subintervals of equal (Lebesgue) size when k is even, and into $k+1$ subintervals when k is odd. We call these subintervals the local subintervals. By (9), we are assured that each local subinterval has at least k elements. We call the left (right) subinterval of $[t_{i-1} : t_i]$ the left (right) local boundary interval. The other subintervals of the split will be called the local interior intervals. We will define q_j for each local subinterval and join them by discrete derivatives matching, the latter having the following meaning. Let $p_1(j)$ and $p_2(j)$ be two functions of $j \in [k+1 : n]$. We then say that their $(k-1)$th order discrete derivatives match at the point $j_0 \in [2k+1 : n]$ if $p_1(j) = p_2(j)$ for $j \in [j_0 - k + 1 : j_0]$.

- *The continuous version of the interpolating vector*. A continuous interpolation $\mathrm{q} : [0, 1] \to [-1, 1]$ with $\mathrm{q}(0) = 1$ and $\mathrm{q}(1) = -1$ can be constructed as follows. We choose q antisymmetric around $x = 1/2$, that is, given $\mathrm{q}(x)$ for $x \in [0, 1/2]$ we let $\mathrm{q}(x) := -\mathrm{q}(1-x)$ for $x \in [1/2, 1]$. We split $[0, 1]$ into N intervals of equal size where $N = k+2$ if k is even,

and $N = k + 1$ if k is odd. Call these subintervals $\{[x_{l-1}, x_l]\}_{l=1}^{N}$ (thus $x_0 = 0$, $x_N = 1$ and $x_{N/2} = 1/2$). For $x \in [x_0, x_1]$, we let

$$\mathrm{q}(x) := 1 - \mathrm{a}_0 x^{\frac{2k-1}{2}},$$

where the constant $\mathrm{a}_0 > 0$ is to be determined. For $x \in [x_{N-1}, x_N]$, we then have

$$\mathrm{q}(x) = -1 + \mathrm{a}_0 (1 - x)^{\frac{2k-1}{2}}.$$

For $x \in [x_{N/2-1}, x_{N/2+1}]$, we let $\mathrm{q}(x) = \mathrm{a}_L (1/2 - x)^L + \cdots + \mathrm{a}_1 (1/2 - x)$ be a linear combination of odd powers of $(1/2 - x)$ where $L = k - 1$ if k is even and $L = k$ if k is odd. By the antisymmetry, it remains do define $\mathrm{q}(x)$ for $x \in [x_1, x_{N/2-1}]$. For $x \in [x_{l-1}, x_l]$ with $l \in \{2, \ldots, N/2 - 1\}$, we let $\mathrm{q}(x) := \mathrm{b}_{l,k} x^k + \cdots + \mathrm{b}_{l,1} x + \mathrm{b}_{l,0}$ be a polynomial of degree k. We choose the coefficients $\{\{\mathrm{a}_j\}, \{\mathrm{b}_{l,j}\}\}$ by derivatives matching: solving a linear system with $k(k/2)$ equations and $k(k/2)$ unknowns when k is even, and with $k(k-1)/2$ equations and $k(k-1)/2$ unknowns when k is odd. The resulting function $\mathrm{q} : [0, 1] \to \mathbb{R}$ is interpolating between $+1$ and -1 and it is continuous with $k - 1$ continuous derivatives, such that the kth left derivative is piecewise constant except on the left boundary interval where it behaves like $-1/\sqrt{x}$ and the right boundary interval where it behaves like $(-1)^k/\sqrt{1-x}$. For a given k, the coefficients $\{\{\mathrm{a}_j\}, \{\mathrm{b}_{l,j}\}\}$ can be given explicitly and it can then be checked whether q is a decreasing function on the interval $[0, 1]$ (or stays away from ± 1). We did this for $k \in \{1, 2, 3, 4\}$ below.

- *The case of a sign change*. If $q_{t_i} q_{t_{i-1}} = -1$, we apply a discrete version of the continuous function q described above. One way to do this is using the map $t_{i-1} + j \mapsto q_{t_{i-1}} \mathrm{q}(j/n_i)$ for $j \in [1 : n_i - 1]$. Alternatively, one may apply discrete derivatives matching. We take $q_{t_{i-1}+j}$ antisymmetric around the midpoint of $[t_{i-1} : t_i]$. We choose $q_{t_{i-1}} q_{t_{i-1}+j} := 1 - a_0 (j/n_i)^{\frac{2k-1}{2}}$ for j in the left local boundary interval (and thus $q_{t_{i-1}} q_{t_i - j} = -1 + a_0 (1 - j/n_i)^{\frac{2k-1}{2}}$ at the right local boundary interval) where $a_0 > 0$ depends on k and n_i. At the two local interior intervals around the midpoint of $[t_{i-1}, t_i]$, we take $q_{t_{i-1}+j}$ a linear combination of odd powers $l \le k$ of $(j - n_i/2)$. At the other interior intervals, we let $q_{t_{i-1}+j}$ be a polynomial of degree k. Then we choose the coefficient a_0 and the coefficients of the polynomials by discrete derivatives matching. For $\min_{i \in S^\pm} n_i \to \infty$, the coefficient a_0 converges to its continuous counterpart a_0. We conclude $a_0 \asymp 1$. The same is true for the other coefficients in the interpolation.
- *The boundary intervals*. We set $q_{t_0} = 0$, and $\{q_{t_0+j}\}_{j=1}^{n_1 - 1}$ an interpolating vector constructed as for the case of a sign change, except that we now interpolate between 0 and ± 1 instead of between 1 and -1. A similar construction is made for the right boundary interval $[t_s : t_{s+1}]$ where we set $q_{t_{s+1}} = 0$.
- *The case of no sign change*. When $q_{t_i} q_{t_{i-1}} = 1$, we take

$$q_{t_i} q_{t_{i-1}+j} := 1 - \left(\frac{4j(n_i - j)}{n_i n_{\max}} \right)^{\frac{2k-1}{2}}, \quad j = [1 : n_i - 1].$$

- *Joining the intervals* $\{[t_{i-1} : t_i]\}_{i=1}^{s+1}$. By the above construction, the first order differences for $j \in [t_{i-1} : t_{i-1} + k]$ or $j \in [t_i - k : t_i]$ are all of order $n_i^{-(2k-1)/2}$ and in fact of order $n_{\max}^{-(2k-1)/2}$ if there is no sign change. This means we can glue the interpolations for the intervals $\{[t_{i-1}, t_i]\}_{i=1}^{s+1}$ together and have the kth discrete derivatives matching up to a finite (depending on k) number of terms of order $n_i^{-(2k-1)/2}$ (or even $n_{\max}^{-(2k-1)/2}$).
- *The kth derivative of q*. The following bound holds: for some constant C_k,

$$n \left\| \Delta(k)' q \right\|_2^2 \le C_k n \left[\sum_{i \in S^\pm} \frac{1 + \log n_i}{n_i^{2k+1}} + \sum_{i \notin S^\pm} \frac{1 + \log n_i}{n_{\max}^{2k+1}} \right].$$

This follows from the construction of q and from Lemma 3.4.

- *The requirement* $q_j \le 1 - w_j$, $j \in \mathcal{D}\backslash S$. Recall the definition of the weights $w_j = \|\psi_j^{-S}\|_n \lambda_0(u)/\lambda$ for all $j \in \mathcal{D}\backslash S$; cf. Definition 2.1 and Lemma 2.4. For $i \notin S^{\pm}$, we have $|q_{t_{i-1}+j}| \le 1 - w_{t_{i-1}+j}$, $j \in [1 : n_i - 1]$ by construction as long as λ satisfies (8). To conclude the same for $i \in S^{\pm}$, we strengthen the requirement (8) to: for an appropriate constant $c_k \ge 1$ depending only on k,

$$\lambda \ge c_k n^{k-1} \left(\frac{n_{\max}}{2n}\right)^{\frac{2k-1}{2}} \sqrt{\frac{2\log(2(n-s-k)) + 2u)}{n}}. \tag{10}$$

It can be shown that if the interpolations for $i \in S^{\pm}$ are monotone, then one can take as $\min_{i \in S^{\pm}} n_i \to \infty$,

$$c_k \to \begin{cases} 2(k+2)^{\frac{2k-1}{2}}/\mathrm{a}_0, & k \text{ even}, \\ 2(k+1)^{\frac{2k-1}{2}}/\mathrm{a}_0, & k \text{ odd}. \end{cases} \tag{11}$$

Equation (11) results by solving, for the boundary intervals, $\mathrm{q}(1/(k+2)) = 1 - \mathrm{a}_0(k+2)^{-\frac{2k-1}{2}}/2 \le 1 - \max_j w_j$ for k even and $\mathrm{q}(1/(k+1)) = 1 - \mathrm{a}_0(k+1)^{-\frac{2k-1}{2}}/2 \le 1 - \max_j w_j$ for k odd, where w depends on λ. The factor $1/2$ comes from the fact that we consider the boundary intervals, where the interpolation happens between 1 and 0 or -1 and 0 (and not between 1 and -1).

It is not a priori clear to us that the interpolations $i \in S^{\pm}$ are monotone. We check this for $k = \{1, 2, 3, 4\}$ in the next 4 subsections.

3.3.2. *Interpolating vector and effective sparsity for* $k = 1$. The case $k = 1$ has been well studied; see Guntuboyina et al. (2020) and its references. We include it here to highlight the additional argument needed when $k > 1$. In the noiseless case and when $k = 1$, we take a linear interpolation of $(q_1 := 0, q'_S, q_{n+1} := 0)$. At a sign change: $q_{t_{i-1}} q_{t_i} = -1$, we take a linear interpolation between plus and minus one over an interval of length n_i. The slope in this interval will then be $2/n_i$, which gives a contribution $4/n_i$ to the bound for the effective sparsity. Similar observations can be made for the boundary interval $[t_0 : t_1]$ where we face a boundary effect due to partial integration because $q_2 = 1/n_1 \ne 0$. For the right boundary interval $[t_s : t_{s+1}]$, we see the same boundary effect. So for $k = 1$,

$$\Gamma^2(S, q_S) \le \frac{n}{n_1^2} + \frac{n}{n_1} + \sum_{q_{t_i} q_{t_{i-1}} = -1} \frac{4n}{n_i} + \frac{n}{n_{s+1}} + \frac{n}{n_{s+1}^2}.$$

REMARK 3.2. In presence of a "staircase pattern"—consecutive entries of $(Df)_S$ having the same sign—the interpolating vector q for the noiseless case can be chosen to be constant between consecutive entries of q_S having the same sign. Therefore, staircase patterns seem to favour prediction, while for $f = f^0$ they are known to negatively affect sign consistency of the first-order differences. A sufficient and almost necessary condition for sign consistency of the first- order differences in the case of the fused lasso is the irrepresentable condition by Zhao and Yu (2006). Qian and Jia (2016) prove that in presence of staircase patterns in $(Df^0)_{S_0}$ the irrepresentable condition holds if and only if the jumps involved in the staircase patterns occur at consecutive entries.

In the noisy case, and when $i \in S^{\pm}$ we use a scaled discrete variant of

$$\mathrm{q}(x) := \begin{cases} +1 - \sqrt{2x}, & 0 \le x \le 1/2, \\ -1 + \sqrt{2(1-x)}, & 1/2 \le x \le 1. \end{cases}$$

When $q_{t_{i-1}} q_{t_i} = -1$, we let

$$q_{t_{i-1}+j} q_{t_{i-1}} := \mathrm{q}(j/n_i) = \begin{cases} +1 - \sqrt{2j/n_i}, & j \in [1 : n_i/2], \\ -1 + \sqrt{2(n_i - j)/n_i}, & j \in [n_i/2 : n_i]. \end{cases} \tag{12}$$

At the boundary intervals, say the left boundary interval, we let $q_1 := 0$ and

$$q_{1+j} q_{t_1} := \begin{cases} \sqrt{j/(2n_1)}, & j \in [1 : n_1/2], \\ 1 - \sqrt{(n_1 - j)/(2n_1)}, & j \in [n_1/2 : n_1]. \end{cases} \tag{13}$$

If $q_{t_i} q_{t_{i-1}} = 1$, we take for $j \in [1 : n_i - 1]$

$$q_{t_{i-1}+j} q_{t_{i-1}} := 1 - \sqrt{4j(n_i - j)/(n_i n_{\max})}. \tag{14}$$

In other words, at locations t_i, with $i \in [2 : s]$, where the signs do not change (i.e., $q_{t_i} = q_{t_{i-1}}$) one may choose "less steep" interpolations.

With this choice for q, we get by straightforward calculations: for λ satisfying (10) with $c_1 = 1$, and for a universal constant C_1

$$\begin{aligned} \Gamma^2(S, q_S, w_{-S}) &\le n \big\| \Delta(1)' q \big\|_2^2 \\ &\le C_1 n \Bigg[\frac{1 + \log n_1}{n_1} + \sum_{q_{t_i} q_{t_{i-1}} = -1} \frac{1 + \log n_i}{n_i} \\ &\quad + \sum_{q_{t_i} q_{t_{i-1}} = 1} \frac{1 + \log n_i}{n_{\max}} + \frac{1 + \log n_{s+1}}{n_{s+1}} \Bigg]. \end{aligned}$$

3.3.3. *Interpolating vector and effective sparsity for $k = 2$.* The splitting scheme for the noisy case of Section 3.3.1 applied to $k = 2$ gives as continuous interpolation

$$\mathrm{q}(x) := \begin{cases} +1 - \dfrac{2}{5}(4x)^{3/2}, & 0 \le x \le \dfrac{1}{4}, \\ \dfrac{3}{5} 4 \Big(\dfrac{1}{2} - x \Big), & \dfrac{1}{4} \le x \le \dfrac{3}{4}, \\ -1 + \dfrac{2}{5} 4^{3/2} (1 - x)^{3/2}, & \dfrac{3}{4} \le x \le 1. \end{cases}$$

Else, we may use a simpler alternative solution, namely

$$\mathrm{q}_{\mathrm{alt}}(x) := \begin{cases} +1 - (2x)^{3/2}, & 0 \le x \le \dfrac{1}{2}, \\ -1 + \big(2(1 - x)\big)^{3/2}, & \dfrac{1}{2} \le x \le 1. \end{cases}$$

Then $\mathrm{q}_{\mathrm{alt}}$ is decreasing, $\mathrm{q}_{\mathrm{alt}}$ and $\mathrm{q}'_{\mathrm{alt}}$ are continuous and

$$\mathrm{q}''_{\mathrm{alt}}(x) = \begin{cases} -\dfrac{3}{\sqrt{2x}}, & 0 < x < \dfrac{1}{2}, \\ +\dfrac{3}{\sqrt{2(1 - x)}}, & \dfrac{1}{2} \le x < 1. \end{cases}$$

Consider now a sign change: $q_{t_i} q_{t_{i-1}} = -1$. We may assume without loss of generality that n_i is even. (If n_i is odd, we take $q_{t_{i-1}+1} = q_{t_{i-1}}$, which is possible because $t_{i-1} + 1$ is set to be a "mock" active variable and replace n_i by $n_i - 1$.) For $j \in [1 : n_i/2]$, define

$$q_{t_{i-1}+j} q_{t_{i-1}} := \mathrm{q}_{\mathrm{alt}}(j/n_i) = \begin{cases} +1 - (2j/n_i)^{3/2}, & j \in [1 : n_i/2], \\ -1 + \big(2(n_i - j)/n_i\big)^{3/2}, & j \in [n_i/2 : n_i]. \end{cases} \tag{15}$$

At the boundary intervals, say the left boundary interval, we let $q_2 = 0$. Again we may without loss of generality, assume n_1 is even (otherwise we let $q_3 = 0$ and replace n_1 by $n_1 - 1$). Then let

$$q_{2+j}q_{t_1} := \begin{cases} \sqrt{2}(j/n_1)^{3/2}, & j \in [1 : n_1/2], \\ 1 - \sqrt{2}\big((n_1 - j)/n_1\big)^{3/2}, & j \in [n_1/2 : n_1]. \end{cases} \tag{16}$$

Finally, if there is no sign change: $q_{t_i}q_{t_{i-1}} = 1$, we let

$$q_{t_{i-1}+j}q_{t_{i-1}} := 1 - \left(\frac{4j(n_i - j)}{n_i n_{\max}}\right)^{3/2}. \tag{17}$$

Thus when λ satisfies (10) for an appropriate constant c_2, then for a constant C_2 the bound (3) is true for the effective sparsity.

3.3.4. *Interpolating vector and effective sparsity for $k = 3$.* For the noisy case, we invoke a scaled and discrete variant of

$$\mathrm{q}(x) := \begin{cases} +1 - \dfrac{4}{19}(4x)^{\frac{5}{2}}, & 0 \le x \le \dfrac{1}{4}, \\ -\dfrac{5}{38}4^3\left(\dfrac{1}{2} - x\right)^3 + \dfrac{35}{38}4\left(\dfrac{1}{2} - x\right), & \dfrac{1}{4} \le x \le \dfrac{3}{4}, \\ -1 + \dfrac{4}{19}4^{5/2}(1 - x)^{\frac{5}{2}}, & \dfrac{3}{4} \le x \le 1, \end{cases} \tag{18}$$

which can be derived using the construction for the continuous version of the interpolating vector for general k, given in Section 3.3.1. Note that q is decreasing, q, q′ and q″ are continuous.

If $n_i/4 \in \mathbb{N}$, the rescaled and discrete variant when $q_{t_i} = -1$, $q_{t_{i-1}} = 1$ is

$$q_{t_i+j} := \begin{cases} 1 - \bar{a}_0(4j/n_i)^{5/2}, & 1 \le j \le n_i/4, \\ -\bar{a}_3 4^3\big((n_i/2 - j)/n_i\big)^3 + \bar{a}_1 4(n_i/2 - j)/n_i, & n_i/4 \le j \le 3n_i/4, \\ -1 + \bar{a}_0 4^{5/2}\big((n_i - j)/n_i\big)^{5/2}, & 3n_i/4 \le i \le n_i, \end{cases} \tag{19}$$

where $\bar{a}_0$, $\bar{a}_1$ and $\bar{a}_3$ can be calculated using the following lemma with $d = n_i/4$. (In the notation of Section 3.3.1, $a_0 = 4^{3/2}\bar{a}_0$, $a_1 = 4\bar{a}_1$ and $a_3 = 4^3\bar{a}_3$.)

LEMMA 3.5. *Let $d \in \mathbb{N}$ and define*

$$\alpha_1 := \frac{[\Delta(d+1)^{5/2}]}{d^{3/2}} = \frac{(d+1)^{5/2} - d^{5/2}}{d^{3/2}},$$

$$\gamma_1 := \frac{[\Delta d^3]}{d^2} := \frac{d^3 - (d-1)^3}{d^2},$$

$$\alpha_2 := \frac{[\Delta(2)(d+2)^{5/2}]}{d^{1/2}} := \frac{[\Delta(d+2)^{5/2}] - [\Delta(d+1)^{5/2}]}{d^{1/2}},$$

$$\gamma_2 := \frac{[\Delta(2)d^3]}{d} := \frac{[\Delta d^3] - [\Delta(d-1)^3]}{d}.$$

Let

$$\bar{a}_0 := \frac{\gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)},$$

$$\bar{a}_3 := \frac{\alpha_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)},$$
$$\bar{a}_1 := \frac{\gamma_1\alpha_2 + \alpha_1\gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)},$$

and for $j \in \{d, d+1, d+2\}$

$$\mathbf{q}_j := 1 - \bar{a}_0 j^{5/2}/d^{5/2},$$
$$\mathbf{p}_j := -\bar{a}_3(2d-j)^3/d^3 + \bar{a}_1(2d-j)/d.$$

Then

$$\Delta(l)\mathbf{q}_{d+l} = \Delta(l)\mathbf{p}_{d+l}, \quad l \in \{0, 1, 2\}.$$

PROOF OF LEMMA 3.5. See Appendix C in the Supplementary Material (Ortelli and van de Geer (2021)). □

The values of the parameters $\bar{a}_0$, $\bar{a}_1$ and $\bar{a}_3$ in the above lemma depend on d, but one easily checks that for $d \to \infty$: $\alpha_1 \approx 5/2$, $\gamma_1 \approx 3$, $\alpha_2 \approx 15/4$ and $\gamma_2 \approx 6$. Hence $\bar{a}_0 \approx \frac{4}{19}$, $\bar{a}_3 \approx 5/38$ and $\bar{a}_1 \approx 35/38$ as in (18). If $n_i/4 \notin \mathbb{N}$, we have similar calculations: the discrete derivatives are then to match at say $\lfloor n_i/4 \rfloor$ and $\lceil 3n_i/4 \rceil$. (By the same arguments as for $k = 2$, one may without loss of generality assume that n_i is even.)

For the boundary intervals, we have similar expressions and when $q_{t_{i-1}} q_{t_i} = 1$ we take q_j, $j \in [t_{i-1} : t_i]$, as for the general k case. This gives when λ satisfies (10) for some appropriate constant c_3, then for a constant C_3 the bound (3) for the effective sparsity.

3.3.5. *Interpolating vector and effective sparsity for* $k = 4$. For $k = 4$ and $i \in S^{\pm}$, we take a scaled and discrete version of the function $\mathrm{q} : [0, 1/2] \to [0, 1]$ defined (up to rounding errors) as

$$\mathrm{q}(x) := \begin{cases} 1 - (18.62)x^{7/2}, & 0 \le x \le \dfrac{1}{6}, \\ (44.34)x^4 - (46.19)x^3 + (10.16)x^2 - (1.10)x + 1.05, & \dfrac{1}{6} \le x \le \dfrac{1}{3}, \\ -(12.93)(1/2 - x)^3 + (4.23)(1/2 - x), & \dfrac{1}{3} \le x \le \dfrac{1}{2}. \end{cases}$$

The function q, illustrated in Figure 1, is decreasing with q, q', q'' q''' continuous. It was calculated by solving 8 equations with 8 unknowns, following the description in Section 3.3.1. We can now reason as in Section 3.3.4 to obtain for λ satisfying (8) the bound (3) for the effective sparsity where c_k and C_k are appropriate constants.

3.4. *Proof of Theorem* 1.1. Taking $\bar{\mathcal{N}}_{-S}$ as the direct product of $\mathcal{N}_{-S}$ and an appropriate space spanned by $(k-1)s$ additional variables, we derived in Section 3.1.1 a bound for the length of the columns of the dictionary Ψ^{-S}. With these, we saw that the requirement (6) for λ can is true when (8) holds. Then in Sections 3.3.2, 3.3.3, 3.3.4 and 3.3.5, we derived a bound for the effective sparsity $\Gamma^2(S, q_S, w_{-S})$ when (8) is strengthened to (10). Theorem 1.1 thus follows from the adaptive bound in Theorem 2.2 for the general analysis problem.

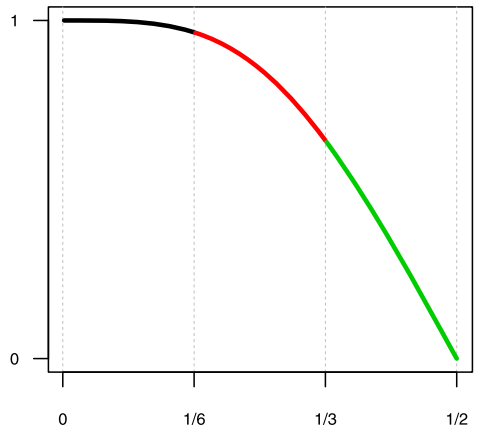

FIG. 1. *The function* $\mathrm{q} : [0, 1/2] \to [0, 1]$ *for* $k = 4$.

4. Conclusion. The sharp oracle inequalities with fast rates show that for $k \in \{1, 2, 3, 4\}$ the estimator adapts to the unknown number of jumps in the $(k-1)$th discrete derivative and provide finite-sample prediction bounds. In particular, these show that the prediction error of the total variation regularized estimator is upper bounded by the optimal trade-off between approximation error and estimation error. The key tool for providing these results is bounding the effective sparsity using interpolating vectors. However, we are not able to prove that the approach we use to find an interpolating vector for $k \in \{1, 2, 3, 4\}$ gives a suitable interpolating vector for general k. Thus, although for each given finite k we can check by computer whether our construction gives an interpolating vector; the problem remains open for general k.

The approach we use allows extensions to other problems as well, for instance, higher dimensional extensions. See Ortelli and van de Geer (2020a) where the Vitali variation serves as regularizer. For total variation on graphs, one may apply the fact that the dictionary can be formed by counting the number of times an edge is used when traveling from a given node to all other nodes. This can then be done on the subgraphs formed by removing the active edges. For graphs with cycles there are several paths from one node to another. One may then choose those that allow for a smooth interpolating vector. Finally, the approach can be extended to estimation problems with ℓ_1-penalty on the discrete derivative but loss functions other than least squares (see van de Geer (2020) for the case of logistic loss with total variation penalty on the canonical parameter).

Acknowledgments. We thank the Associate Editor and the referees for their very helpful remarks.

Funding. We acknowledge support for this project from the the Swiss National Science Foundation (SNF Grant 200020_169011).

SUPPLEMENTARY MATERIAL

Supplement to "Prediction bounds for higher order total variation regularized least squares" (DOI: 10.1214/21-AOS2054SUPP; .pdf). In Appendix A, we prove Theorem 2.2. In Appendix B, we provide proofs for Section 3.1 and in Appendix C for Section 3.3.

REFERENCES

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324 https://doi.org/10.1093/biomet/asr043

BOYER, C., DE CASTRO, Y. and SALMON, J. (2017). Adapting to unknown noise level in sparse deconvolution. *Inf. Inference* **6** 310–348. MR3764527 https://doi.org/10.1093/imaiai/iaw024

CANDÈS, E. J. and FERNANDEZ-GRANDA, C. (2014). Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.* **67** 906–956. MR3193963 https://doi.org/10.1002/cpa.21455

CANDÈS, E. J. and PLAN, Y. (2011). A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory* **57** 7235–7254. MR2883653 https://doi.org/10.1109/TIT.2011.2161794

CHATTERJEE, S. and GOSWAMI, S. (2019). Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees. Preprint. Available at arXiv:1911.11562.

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

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

ELAD, M., MILANFAR, P. and RUBINSTEIN, R. (2007). Analysis versus synthesis in signal priors. *Inverse Probl.* **23** 947–968. MR2329926 https://doi.org/10.1088/0266-5611/23/3/007

GUNTUBOYINA, A., LIEU, D., CHATTERJEE, S. and SEN, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** 205–229. MR4065159 https://doi.org/10.1214/18-AOS1799

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

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785 https://doi.org/10.1214/aos/1015957395

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

ORTELLI, F. and VAN DE GEER, S. (2020a). Adaptive rates for total variation image denoising. *J. Mach. Learn. Res.* **21** 1–38.

ORTELLI, F. and VAN DE GEER, S. (2020b). Oracle inequalities for square root analysis estimators with application to total variation penalties. *Inf. Inference* iaaa002.

ORTELLI, F. and VAN DE GEER, S. (2021). Supplement to "Prediction bounds for higher order total variation regularized least squares." https://doi.org/10.1214/21-AOS2054SUPP

PADILLA, O. H. M. and CHATTERJEE, S. (2020). Adaptive quantile trend filtering. Preprint. Available at arXiv:2007.07472.

QIAN, J. and JIA, J. (2016). On stepwise pattern recovery of the fused Lasso. *Comput. Statist. Data Anal.* **94** 221–237. MR3412821 https://doi.org/10.1016/j.csda.2015.08.013

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

SADHANALA, V. and TIBSHIRANI, R. J. (2019). Additive models with trend filtering. *Ann. Statist.* **47** 3032–3068. MR4025734 https://doi.org/10.1214/19-AOS1833

SADHANALA, V., WANG, Y.-X., SHARPNACK, J. and TIBSHIRANI, R. (2017). Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems* 5800–5810.

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

TANG, G., BHASKAR, B. N. and RECHT, B. (2015). Near minimax line spectral estimation. *IEEE Trans. Inf. Theory* **61** 499–512. MR3299978 https://doi.org/10.1109/TIT.2014.2368122

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

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

TIBSHIRANI, R. (2020). Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. Preprint. Available at arXiv:2003.03886.

VAN DE GEER, S. (2020). Logistic regression with total variation regularization. *Trans. A. Razmadze Math. Inst.* **174** 217–233. MR4150558

WANG, Y.-X., SMOLA, A. and TIBSHIRANI, R. (2014). The falling factorial basis and its statistical applications. In *International Conference on Machine Learning* 730–738.

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, 41. MR3543511

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449