Tensor denoising with trend filtering

Journal Article

Author(s):
Ortelli, Francesco ; van de Geer, Sara

Publication date:
2022

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000525559

Originally published in:
Mathematical Statistics and Learning 4, https://doi.org/10.4171/MSL/26

Funding acknowledgement:
- Topics in High-Dimensional Statistics ()

Math. Stat. Learn. 4 (2021), 87–142
DOI 10.4171/MSL/26

Tensor denoising with trend filtering

Francesco Ortelli and Sara van de Geer

Abstract. We extend the notion of trend filtering to tensors by considering the kth-order Vitali variation – a discretized version of the integral of the absolute value of the kth-order total derivative. We prove adaptive ℓ^0-rates and not-so-slow ℓ^1-rates for tensor denoising with trend filtering.

For $k = \{1, 2, 3, 4\}$ we prove that the d-dimensional margin of a d-dimensional tensor can be estimated at the ℓ^0-rate n^{-1}, up to logarithmic terms, if the underlying tensor is a product of $(k-1)$th-order polynomials on a constant number of hyperrectangles. For general k we prove the ℓ^1-rate of estimation $n^{-\frac{H(d)+2k-1}{2H(d)+2k-1}}$, up to logarithmic terms, where $H(d)$ is the dth harmonic number.

Thanks to an ANOVA-type of decomposition we can apply these results to the lower dimensional margins of the tensor to prove bounds for denoising the whole tensor. Our tools are interpolating tensors to bound the effective sparsity for ℓ^0-rates, mesh grids for ℓ^1-rates and, in the background, the projection arguments by Dalalyan, Hebiri, and Lederer (2017).

1. Introduction

Let $f^0 \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a d-dimensional tensor with $n = n_1 \cdot \ldots \cdot n_d$ entries. We want to prove error bounds for tensor denoising, which is the task of recovering f^0 from its noisy version $Y = f^0 + \varepsilon$, where ε has i.i.d. Gaussian entries with mean 0 and variance σ^2.

We show that we can estimate the underlying tensor f^0 in an adaptive manner with a regularized least-squares signal approximator. As regularizer we propose the Vitali variation of the $(k-1)$th-order total differences of the candidate estimator for $k \geq 1$. We call this regularizer the "kth-order Vitali total variation". We use the abbreviation TV for "total variation". This approach extends the idea of "trend filtering" [12, 28] to tensors.

We expose the notion of TV regularization, review the literature on adaptive results for TV regularization, explain the concept of adaptation for structured problems, introduce an ANOVA-type of decomposition of a tensor, outline our contributions and finally present the organization of the paper.

2020 Mathematics Subject Classification. 62J07.
Keywords. Tensor denoising, total variation, Vitali variation, trend filtering, oracle inequalities.

1.1. TV regularization

A regularized (least-squares) signal approximator is an estimator $\widehat{f}$ defined as

$$\widehat{f} := \arg\min_{\mathrm{f} \in \mathbb{R}^{n_1 \times \cdots \times n_d}} \left\{ \|Y - \mathrm{f}\|_2^2 / n + 2\lambda \, \mathrm{pen}(\mathrm{f}) \right\},$$

where $\|\cdot\|_2^2$ denotes the sum of the squared entries of its argument, $\lambda > 0$ is a tuning parameter and pen(f) is a regularization penalty.

When $\mathrm{pen}(\mathrm{f}) = \|D\mathrm{f}\|_1$ for a linear operator D and for $\|\cdot\|_1$ denoting the sum of the absolute values of the entries of its argument, the regularized signal approximator is called "ℓ^1-analysis estimator" or simply "analysis estimator" [5]. If the linear operator D is a difference operator, then $\mathrm{pen}(\mathrm{f}) = \|D\mathrm{f}\|_1$ is usually called TV of f and the estimator $\widehat{f}$ is called TV regularized estimator. Different choices of the difference operator D are possible, resulting in different notions of TV.

For a continuous image defined on $(x_1, \ldots, x_d) \in [0, 1]^d$, one can choose D as a discretized version of either the total kth-order derivative operator $\prod_{i=1}^d \partial^k / (\partial x_i)^k$ or of the sum of kth-order partial derivative operators $\sum_{i=1}^d \partial^k / (\partial x_i)^k$.

1.2. Literature review: adaptive results for TV regularization

For $d = 1$ partial and total derivatives coincide. With D being the first order difference matrix, the TV regularized estimator is also known under the name "fused Lasso" [8, 27]. Adaptivity of the fused Lasso has been proved by Dalalyan et al. [4], Lin et al. [13], Ortelli and van de Geer [16], and Guntuboyina et al. [10].

The "edge Lasso" extends the fused lasso to graphs and is studied by Sharpnack et al. [24], and Hütter and Rigollet [11]. Ortelli and van de Geer [16, 18] prove adaptivity of the edge Lasso on tree graphs and cycle graphs, respectively.

The idea of the fused Lasso can also be extended to the penalization of higher-order differences. This extension is called "trend filtering" [12, 25, 28]. Adaptivity of trend filtering is established in [10, 19]. Wang et al. [32] consider trend filtering on graphs, Sadhanala et al. [20, 22] in higher-dimensional situations and [21] for additive models.

Here, we consider the case of D being a discretization of $\prod_{i=1}^d \partial^k / (\partial x_i)^k$. We call the corresponding notion of TV "kth-order Vitali TV". In the literature, signal approximators regularized with the Vitali TV are studied by Mammen and van de Geer [14], Ortelli and van de Geer [17], and Fang et al. [6]. Ortelli and van de Geer [17] prove adaptivity for $d = 2$ and $k = 1$. Fang et al. [6] show adaptivity for $d = 2$ and $k = 1$ using as regularizer the Hardy–Krause variation, which is the sum of the Vitali TV of a matrix and of its margins. In this paper we will prove adaptivity of tensor denoising with kth-order Vitali TV regularization for $k \in \{1, 2, 3, 4\}$

and general dimension $d \geq 1$. The results obtained for $k \in \{1, 2, 3, 4\}$ and $d = 1$ in [19] and for $k = 1$ and $d = 2$ in [17] will then be retrieved as special cases.

For $k = 1$, signal approximators regularized with D being a discretization of the partial derivative operator $\sum_{i=1}^{d} \partial^k/(\partial x_i)^k$ are studied by Hütter and Rigollet [11], and Sadhanala et al. [22] for general d. For $d = 2$, Chatterjee and Goswami [3] show the fast rate $n^{-3/4}$ for estimating axis-aligned rectangles. Sadhanala et al. [22], and Sadhanala et al. [20] call the estimator for general k Kronecker trend filtering.

1.3. Adaptation for structured problems

The analysis estimator $\widehat{f}$ can be recast in a constructive formulation as "synthesis estimator". One can find dictionary tensors $\{\phi_j \in \mathbb{R}^{n_1 \times \cdots \times n_d}\}_{j \in [p]}$, such that

$$\widehat{f} = \sum_{j=1}^{p} \widehat{\beta}_j \phi_j, \quad \text{where } \widehat{\beta} := \arg\min_{b \in \mathbb{R}^p} \Big\{ \|Y - \sum_{j=1}^{p} b_j \phi_j\|_2^2/n + 2\lambda \sum_{j \notin U} |b_j| \Big\},$$

and $U \subseteq \{1, \ldots, p\}$ is a set of indices, cf. [5]. The Lasso estimator [1, 26, 29] is an instance of synthesis estimator. The dictionary $\{\phi_j\}_{j \in [p]}$ and the set of unpenalized coefficients $U \subseteq [p]$ depend on D. We can see that D imposes structure on the estimator: it determines the dictionary with which the estimator is constructed. For instance, in the case of the 1st-order Vitali TV, the dictionary $\{\phi_j\}_{j \in [p]}$ consists of tensors being constant on hyperrectangles. Therefore, the estimator $\widehat{f}$ is constant on few hyperrectangular pieces.

Our goal is to prove adaptation of the estimator $\widehat{f}$ to the underlying signal f^0, when $\|D\mathrm{f}\|_1$ is the kth-order Vitali TV.

Adaptation is a consequence of a high-probability upper bound on the mean squared error (MSE) in the form of the oracle inequality

$$\|\widehat{f} - f^0\|_2^2/n \leq \|g - f^0\|_2^2/n + \mathrm{rem}(D, g, S), \tag{1.1}$$

where $g \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is an arbitrary tensor, S is an arbitrary set of indices of Dg and $\mathrm{rem}(D, g, S)$ is a remainder term. A result of the form of (1.1) establishes the adaptation of the estimator $\widehat{f}$, provided that the remainder term $\mathrm{rem}(D, g = f^0, S = S^0)$ converges to zero, where S^0 is the set of the indices of the nonzero coefficients of Df^0. The cardinality $s^0 := |S^0|$ of S^0 is called the "sparsity" of f^0 with respect to D.

We can optimize the upper bound in (1.1) over g and S. However, the optimizers g^* and S^* will depend on f^0 – which is unobserved. Hence the name "oracle" for the pair $(g^*(f^0), S^*(f^0))$ and the name "oracle inequality" for results as (1.1).

Such a result is considered to be adaptive, since different underlying true tensors f^0 will possibly give place to different oracles $(g^*(f^0), S^*(f^0))$ and to different values for the upper bound.

Results as (1.1) are only useful if it can be proved that $\mathrm{rem}(D, f^0, S^0)$ converges to zero. Typically

$$\mathrm{rem}(D, f^0, S^0) = \mathcal{O}\big(\lambda^2 \Gamma_D^2(S^0)\big),$$

where $\Gamma_D^2(S^0)$ is called "effective sparsity" and depends both on D and S^0. Proving adaptivity therefore translates into proving a bound for the effective sparsity: a task which depends on the structure imposed by D. To bound the effective sparsity for tensor denoising with trend filtering we use an interpolating tensor, in analogy to the interpolating vector and the interpolating matrix by Ortelli and van de Geer [17, 19].

Adaptive results as (1.1) are a consequence of a careful choice of λ. The general theory for the Lasso [1, 29] suggests the choice $\lambda \asymp \lambda_0 \asymp \sqrt{\log(n)/n}$, where λ_0 is called the "universal choice". The universal choice ensures that all the noise is overruled. However, Dalalyan et al. [4] show that also the smaller choice $\lambda \asymp \tilde{\gamma}\lambda_0$ is possible, where $\tilde{\gamma} > 0$ is a scaling factor which accounts for the correlation in the dictionary $\{\phi_j\}_{j \in [p]}$ induced by D. The projection arguments by Dalalyan et al. [4] in the background of our results allow us to choose the tuning parameter of smaller order than the universal choice λ_0.

Projection arguments have been discussed in the literature. We do not report them here but refer instead to Theorem 3 in [4], Lemma B2 and Lemma C2 in [18], Lemma 13 in [17], and to [31].

1.4. ANOVA decomposition

In the continuous case, the "nullspace" of the kth-order derivative operator along one coordinate is made of constant, linear, ..., $(k-1)$th-order monomial functions. The nullspace of the total derivative operator in d-dimensions is made of d-dimensional functions which are linear, ..., $(k-1)$th-order monomial along at least one coordinate. In the discrete case when $n_1 \asymp \cdots \asymp n_d$ the linear space spanned by such tensors is $n^{1-1/d}$-dimensional.

We will decompose a tensor $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ into a sum of mutually orthogonal tensors. Each of these mutually orthogonal tensors will be constant or linear or ... or $(k-1)$th-order monomial along a set of l coordinates, for $l \in [0:d]$. This construction will be carried out for all possible sets of coordinates in $[d]$. Tensors being constant or linear or ... or $(k-1)$th-order monomial along $d-l$ coordinates will be called l-dimensional margins.

We will adaptively estimate l-dimensional margins with l-dimensional Vitali TV regularized estimators, for $l \in [d]$. The 0-dimensional margins will be estimated by ordinary least squares at a rate n^{-1}. By estimating all the margins adaptively we will be able to prove adaptivity of the denoising of the whole tensor via Vitali TV regularization.

1.5. Contributions

Previously, we have derived tools like interpolating vectors and matching derivatives to prove adaptivity for trend filtering ($d = 1$ and $k \in \{1, 2, 3, 4\}$, see [19]). In [17] we have come up with tools to extend our results for adaptation of the fused Lasso ($d = 1$ and $k = 1$) to the two-dimensional case of image denoising ($d = 2$ and $k = 1$). Here, we show in the first place how to combine and extend the tools from image denoising and one-dimensional trend filtering to handle trend filtering for $k \in \{1, 2, 3, 4\}$ and for general dimension d. Establishing adaptivity requires a so-called "bound on the antiprojections". We prove a formula giving the bounds on the antiprojections for general k and d. We then propose an ANOVA decomposition to ensure that all the margins of a d-dimensional tensor can be estimated adaptively.

Lastly, we prove slow rates for tensor denoising with trend filtering. We extend the idea of mesh grid by Ortelli and van de Geer [17] to general d and general k. We then prove a bound on the antiprojections with the help of the mesh grid holding for all d and all k.

The integration of the arguments by Ortelli and van de Geer [19] with the ones by Ortelli and van de Geer [17], the general bounds on the antiprojections and the ANOVA decomposition allow us to present general risk bounds for tensor denoising with trend filtering.

1.6. Organization of the paper

In Section 2 we expose the required notation, the model and define the trend filtering estimator for the d-dimensional margin.

In Section 3 we list our contributions and give a preview of the results: adaptive ℓ^0-rates and not-so-slow ℓ^1-rates.

In Section 4 we derive the synthesis form of the trend filtering estimator for the d-dimensional margin.

Proving the main result on adaptivity for tensor denoising with trend filtering is the topic of Section 5.

In Section 6 we apply a general result on slow ℓ^1-rates for analysis estimators to tensor denoising with trend filtering.

In Section 7 we show the ANOVA decomposition of a tensor and define the estimators for lower-dimensional margins.

In Section 8 we apply the results on adaptivity and on not-so-slow slow ℓ^1-rates to the estimators for the lower-dimensional margins defined in Section 7. This will establish rates for the estimation of the whole tensor.

Section 9 concludes the paper.

2. Model, notation and estimator

We consider the model

$$Y = f^0 + \varepsilon,$$

where $Y, f^0, \varepsilon \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ are d-dimensional tensors and ε has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries with known variance $\sigma \in (0, \infty)$. For the case of unknown variance we refer to [18], who show how to estimate f^0 and σ at the same time.

The goal is to estimate f^0 given its noisy observations Y. We consider a signal approximator regularized with the Vitali TV.

2.1. Signals supported on d-dimensional tensors

For two integers $i \leq j$ we define

$$[i : j] := \{i, \ldots, j\}.$$

Moreover, if $i = 1$ we write $[j] := [1 : j]$.

Let $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a d-dimensional tensor with $n := n_1 \cdot \ldots \cdot n_d$ entries. For indices $(j_1, \ldots, j_d) \in [n_1] \times \cdots \times [n_d]$ we refer to the corresponding entry of f by $f_{j_1, \ldots, j_d}$ using indices or by $f(j_1, \ldots, j_d)$ using arguments.

For $(j'_1, \ldots, j'_d), (j''_1, \ldots, j''_d) \in [n_1] \times \cdots \times [n_d]$ we use the notation

$$\sum_{j'_1, \ldots, j'_d}^{j''_1, \ldots, j''_d} f_{j_1, \ldots, j_d} := \sum_{j_d = j'_d}^{j''_d} \cdots \sum_{j_1 = j'_1}^{j''_1} f_{j_1, \ldots, j_d}.$$

Similarly, we write

$$\{f_{j_1, \ldots, j_d}\}_{j'_1, \ldots, j'_d}^{j''_1, \ldots, j''_d} := \{f_{j_1, \ldots, j_d}\}_{(j_1, \ldots, j_d) = (j'_1, \ldots, j'_d)}^{(j''_1, \ldots, j''_d)}.$$

By $\|f\|_2 := (\sum_{1, \ldots, 1}^{n_1, \ldots, n_d} f^2_{j_1, \ldots, j_d})^{1/2}$ we denote the Frobenius norm of f. Moreover, we define $\|f\|_1 := \sum_{1, \ldots, 1}^{n_1, \ldots, n_d} |f_{j_1, \ldots, j_d}|$ as the sum of the absolute values of the entries of f.

2.1.1. Tensors with product structure. We now let $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a d-dimensional tensor with $n := n_1 \cdot \ldots \cdot n_d$ entries. Define the set of indices I of the entries of f as $I := [n_1] \times \cdots \times [n_d]$.

We say that f has product structure if there are vectors $\{f_j\}_{j \in [d]}$ such that

$$f(j_1, \ldots, j_d) = f_1(j_1) \cdot \ldots \cdot f_d(j_d), \quad \forall (j_1, \ldots, j_d) \in I.$$

We then write $f = f_1 \times \cdots \times f_d$.

Let f and g be tensors with product structure. We consider the entry-wise multiplication $(f \odot g)_{j_1,\ldots,j_d} = f_{j_1,\ldots,j_d} g_{j_1,\ldots,j_d}$, $(j_1, \ldots, j_d) \in I$. It holds that

$$(f \odot g)_{j_1,\ldots,j_d} = \prod_{l=1}^{d} f_l(j_l) g_l(j_l), \quad \forall (j_1, \ldots, j_d) \in I.$$

2.1.2. Orthogonality between tensors. The operation $\sum_{1,\ldots,1}^{n_1,\ldots,n_d} (f \odot g)_{j_1,\ldots,j_d}$ is the equivalent of the scalar product for tensors. We say that the tensors f and g are orthogonal if

$$\sum_{1,\ldots,1}^{n_1,\ldots,n_d} (f \odot g)_{j_1,\ldots,j_d} = 0.$$

If f and g have product structure and f_i and g_i are orthogonal to each other for at least one coordinate $i \in [d]$, then f and g are orthogonal too.

2.1.3. Linear subspaces and orthogonal projections. Let $\mathcal{W}$ be a linear subspace of $\mathbb{R}^{n_1 \times \cdots \times n_d}$ and let $\mathcal{W}^\perp$ be its orthogonal complement. We denote by

$$\mathrm{I}: \mathbb{R}^{n_1 \times \cdots \times n_d} \mapsto \mathbb{R}^{n_1 \times \cdots \times n_d}$$

the identity operator, i.e., $\mathrm{I} f = f$. By $\mathrm{P}_\mathcal{W}$ we denote the orthogonal projection operator onto $\mathcal{W}$ and by $\mathrm{A}_\mathcal{W} := \mathrm{I} - \mathrm{P}_\mathcal{W} = \mathrm{P}_{\mathcal{W}^\perp}$ the corresponding orthogonal antiprojection operator. For a tensor $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ we write $f_\mathcal{W} := \mathrm{P}_\mathcal{W}$ and $f_{\mathcal{W}^\perp} := f - f_\mathcal{W}$.

For a linear operator Δ, let $\mathcal{N}(\Delta)$ denote its nullspace.

2.2. Estimator

Let k be an integer in $\{1, \ldots, \min_{i \in [d]} n_i - 1\}$. Let D_i^k be the kth-order difference operator along the ith coordinate, defined as

$$(D_i^k f)(j_1, \ldots, j_i, \ldots, j_d) := n_i^{k-1} \sum_{l=0}^{k} (-1)^l \binom{k}{l} f(j_1, \ldots, j_i - l, \ldots, j_d),$$

for

$$(j_1, \ldots, j_{i-1}, j_i, j_{i+1}, \ldots, j_d) \in [n_1] \times \cdots \times [n_{i-1}] \times [k+1 : n_i] \times [n_{i+1}] \times \cdots \times [n_d].$$

Definition 2.1 (Total kth-order difference operator). The total kth-order difference operator D^k is defined as

$$D^k := \prod_{i=1}^{d} D_i^k.$$

The total kth-order difference operator D^k can be seen as a discretized version of $\prod_{i=1}^d \partial^k/(\partial x_i)^k$. It is important to note that the definition of D^k implicitly includes a factor n^{k-1} that stems from the discretization.

The Vitali TV of a tensor $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is defined as the sum of the absolute values of its total kth-order differences.

Definition 2.2 (kth-order Vitali TV). The kth-order Vitali TV $\text{TV}_k(f)$ of a d-dimensional tensor $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is defined as

$$\text{TV}_k(f) := \|D^k f\|_1.$$

The kth-order Vitali TV has the canonical scaling $\text{TV}_k(f) = \mathcal{O}(1)$ due to the normalization by the factor n^{k-1} in the definition of D^k. We refer to [23] for more about canonical scalings.

We define the nullspace $\mathcal{N}_k$ of D^k as

$$\mathcal{N}_k := \{f \in \mathbb{R}^{n_1 \times \cdots \times n_d} : D^k f = 0\}$$

and its orthogonal complement as $\mathcal{N}_k^\perp$. We call $f_{\mathcal{N}_k^\perp}$ the d-dimensional margin of a tensor $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$.

Definition 2.3 (kth-order Vitali trend filtering estimator). The kth-order Vitali trend filtering estimator $\hat{f}_{\mathcal{N}_k^\perp}$ for the d-dimensional margin $f^0_{\mathcal{N}_k^\perp}$ is defined as

$$\hat{f}_{\mathcal{N}_k^\perp} := \arg\min_{\mathrm{f} \in \mathbb{R}^{n_1 \times \cdots \times n_d}} \left\{ \|(Y - \mathrm{f})_{\mathcal{N}_k^\perp}\|_2^2/n + 2\lambda \,\text{TV}_k(\mathrm{f}) \right\},$$

where $\lambda > 0$ is a tuning parameter.

2.3. Active sets

Let $S \subseteq [3 : n_1 - 1] \times \cdots \times [3 : n_d - 1]$ be a subset of the indices of $D^k f$ for some tensor $f \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. We write $s := |S|$ and $S = \{t_1, \ldots, t_s\}$, where $t_m = (t_{1,m}, \ldots, t_{d,m})$. We call $\{t_m\}_{m=1}^s$ the jump locations.

Moreover, we define

$$a_S := \{a_{j_1,\ldots,j_d},\ (j_1, \ldots, j_d) \in S\} \quad \text{and} \quad a_{-S} := \{a_{j_1,\ldots,j_d},\ (j_1, \ldots, j_d) \notin S\}.$$

We will use the same notation a_S for the tensor which shares its entries with a for $(j_1, \ldots, j_d) \in S$ and has all its other entries equal to zero. Similarly, we will also denote by a_{-S} a tensor that shares its entries with a for $(j_1, \ldots, j_d) \notin S$ and has its other entries equal to zero.

3. Contributions

We make the following contributions:

- We extend the idea of trend filtering to d-dimensional settings via the Vitali variation and total discrete derivatives.
- We prove adaptive ℓ^0-rates for tensor denoising with trend filtering for $k \in \{1,2,3,4\}$; see Theorem 3.1, a simplified version of Theorem 5.2. The rates for $d = 1$ and $k \in \{1, 2, 3, 4\}$, and for $d = 2$ and $k = 1$ are known. The rates for the other cases are new contributions. We also expose some sufficient conditions to find adaptive bounds for general k. For each given k one can check by computer whether the conditions hold but the problem of showing that they hold for general k remains open.
- We prove slow ℓ^1-rates for tensor denoising with trend filtering, which turn out to be "not-so-slow", see Theorem 3.2. Here too, the rates for $d = 2$ and $k \geq 2$ and for $d \geq 3$ are new contributions. It is still an open problem whether these rates correspond for $d \geq 2$ to minimax rates (modulo log terms).
- We extend the idea of ANOVA decomposition from 1st-order differences to kth-order differences in d dimensions. By means of this ANOVA decomposition we can apply the results for the d-dimensional margin to lower dimensional margins. We obtain rates for the estimation of the whole tensor by trend filtering.
- Our results allow to recover previous results for trend filtering and image denoising [17, 19] as special cases.

3.1. Preview of the results

We consider tensors in $\mathbb{R}^{n_1 \times \cdots \times n_d}$ such that $n_1 = \cdots = n_d$.

Let

$$\lambda_0(t) := \sigma \sqrt{\frac{2 \log(2n) + 2t}{n}}, \quad t > 0.$$

We call $\lambda_0(t)$ the "universal choice" of the tuning parameter. The universal choice $\lambda = \lambda_0(t)$ guarantees that all the noise is overruled. However, our results also allow for a smaller choice than the universal choice, due to the projection arguments by Dalalyan et al. [4] in the background.

Theorem 3.1 (Adaptivity of Vitali trend filtering, simplified). *Fix* $k \in \{1,2,3,4\}$. *Let* $g \in \mathbb{R}^{n^{1/d} \times \cdots \times n^{1/d}}$ *be arbitrary. Let* $S \subseteq \times_{i \in [d]} [k+2 : n^{1/d} - 1]$ *be an arbitrary set of size* $s := |S|$ *defining a regular grid of cardinality* $s^{1/d} \times \cdots \times s^{1/d}$ *parallel to the coordinate axes. For a large enough constant* $C > 0$ *only dependent on* k*, choose*

$$\lambda \geq C\, d^{3/2} \frac{\lambda_0(\log(2n))}{s^{(2k-1)/2d}}.$$

Then, with probability at least $1 - 1/n$*, it holds that*

$$\|(\hat{f} - f^0)_{\mathcal{N}_k^\perp}\|_2^2/n \leq \|g - f^0_{\mathcal{N}_k^\perp}\|_2^2/n + 4\lambda\|(D^k g)_{-S}\|_1 + \mathcal{O}\left(\lambda^2 \frac{s^{2k}\log(n/s)}{n}\right).$$

Proof. See Section 5.7 for the proof of the more general Theorem 5.2. ■

Some examples of the exponent of s in the rate of Theorem 3.1 for $d \in \{1,2,3\}$ and $k \in \{1,2,3,4\}$ are exposed in Table 1.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$d = 1$	1	1	1	1
$d = 2$	3/2	5/2	7/2	9/2
$d = 3$	5/3	3	13/3	17/3
d general	$2 - 1/d$	$4 - 3/d$	$6 - 5/d$	$8 - 7/d$

Table 1. Some examples of the exponent of s in the rate of Theorem 3.1 for the choice $\lambda \asymp s^{-(2k-1)/2d}\lambda_0(\log(2n))$.

If in Theorem 3.1 we set $g = f^0_{\mathcal{N}_k^\perp}$ and choose the tuning parameter

$$\lambda \asymp s^{-\frac{2k-1}{2d}} \lambda_0(\log(2n))$$

depending on the (typically unknown) true active set S_0, we obtain the rate

$$\mathcal{O}\left(\frac{s_0^{2k-((2k-1)/d)} \log(n/s)\log n}{n}\right).$$

If in Theorem 3.1 we set $g = f^0_{\mathcal{N}_k^\perp}$ and we choose the tuning parameter $\lambda \asymp \lambda_0(\log(2n))$ in a completely data-driven way not depending on the (typically unknown) true active set S_0, we obtain the rate

$$\mathcal{O}\left(\frac{s_0^{2k}\log(n/s)\log n}{n}\right).$$

We now fix $k \in [1 : \min_{i \in [d]} n_i - 1]$. For $d \in \mathbb{N}$ define the dth harmonic number $H(d)$ as $H(d) := \sum_{i=1}^{d} 1/i$.

Theorem 3.2 (Not-so-slow ℓ^1-rate for trend filtering). *Let $g \in \mathbb{R}^{n^{1/d} \times \cdots \times n^{1/d}}$ be an arbitrary tensor.*

Dependence of λ on g allowed. *Choose*

$$\lambda \asymp n^{-\frac{H(d)+2k-1}{2H(d)+2k-1}} \log^{\frac{H(d)}{2H(d)+2k-1}}(n) \|D^k g\|_1^{-\frac{2k-1}{2H(d)+2k-1}}.$$

Then with probability at least $1 - \Theta(1/n)$ it holds that

$$\|(\hat{f} - f^0)_{\mathcal{N}_k^\perp}\|_2^2/n \leq \|g - f^0_{\mathcal{N}_k^\perp}\|_2^2/n + \mathcal{O}\Big(n^{-\frac{H(d)+2k-1}{2H(d)+2k-1}} \log^{\frac{H(d)}{2H(d)+2k-1}}(n) \|D^k g\|_1^{\frac{2H(d)}{2H(d)+2k-1}}\Big).$$

Dependence of λ on g not allowed. *Choose*

$$\lambda \asymp n^{-\frac{H(d)+2k-1}{2H(d)+2k-1}} \log^{\frac{H(d)}{2H(d)+2k-1}}(n).$$

Then with probability at least $1 - \Theta(1/n)$ it holds that

$$\|(\hat{f} - f^0)_{\mathcal{N}^\perp}\|_2^2/n \leq \|g - f^0_{\mathcal{N}_k^\perp}\|_2^2/n + \mathcal{O}\Big(n^{-\frac{H(d)+2k-1}{2H(d)+2k-1}} \log^{\frac{H(d)}{2H(d)+2k-1}}(n)\big(1 + \|D^k g\|_1\big)\Big).$$

Proof. See Section 6.3. ∎

Some examples of the exponent of n in the rate of Theorem 3.2 for $d \in \{1, 2, 3\}$ and $k \in \{1, 2, 3\}$ are exposed in Table 2.

	$k = 1$	$k = 2$	$k = 3$	k general
$d = 1$	$-2/3$	$-4/5$	$-6/7$	$-2k/(2k+1)$
$d = 2$	$-5/8$	$-3/4$	$-13/16$	$-(4k+1)/(4k+4)$
$d = 3$	$-17/28$	$-29/40$	$-41/52$	$-(12k+5)/(12k+16)$

Table 2. Some examples of the exponent of n in the rate of Theorem 3.2.

We call the rates of Theorem 3.2 "not-so-slow" because they turn out to be faster than the slow rate $n^{-1/2}$ for a Lasso problem where no specific structure is imposed.

3.2. Comparison with the literature

We compare Vitali trend filtering (VTF) with two other approaches for extending trend filtering to higher dimensions: graph trend filtering (GTF) by Wang et al. [32] and Kronecker trend filtering (KTF) by Sadhanala et al. [22]. GTF is studied in [22] for grid graphs of general dimension, while KTF is further studied in [20], who show a phase transition for minimax rates achieved by KTF (see Table 2.1 therein).

While the VTF penalty $\|D^k \mathrm{f}\|_1$ involves a discretization of total derivatives of order k, the KTF penalty $\sum_{i=1}^d \|D_i^k \mathrm{f}\|_1$ involves a discretization of partial derivatives of order k. The GTF penalty coincides with the KTF penalty for $k = 1$ but differs at the boundaries for $k \geq 2$, as explained in [22]. It can be observed that the analysis operators of the VTF penalty and of the KTF penalty have different properties.

First, when $n_1 \asymp \cdots \asymp n_d$ it holds that

$$\dim\big(\{f : \|D^k f\|_1\} = 0\big) = n^{1-1/d},$$

while

$$\dim\left(\left\{f : \sum_{i=1}^d \|D_i^k f\|_1\right\} = 0\right) = k^d.$$

The nullspace of the KTF penalty can be estimated by least squares at the rate n^{-1} while the least squares rate for the nullspace of the VTF penalty would be $n^{-1/d}$. This motivates the penalized estimators we propose in Section 7 for the margins.

Second, the analysis operator involved in the VTF penalty is of full rank, while the one for the KTF penalty is not. The estimates produced by VTF are thus products of polynomials of order at most $k-1$ on hyperrectangular pieces as shown in Section 4. The estimates produced by KTF are also products of polynomials of order at most $k-1$, but on pieces that do not need to be hyperrectangular.

VTF on one side, and KTF and GTF on the other side are not directly comparable unless $d = 1$. However, the estimates produced by VTF can be compared with the estimates produced by other methods: the multivariate adaptive regression splines (MARS) by Friedman [7], the dyadic classification and regression trees (dyadic CART) and the optimal regression trees (ORT) proposed by Chatterjee and Goswami [2]. In particular, Chatterjee and Goswami [2] show that ORT achieves the rate z_0/n for $d = 2$ and $z_0^{(d+1)/3}/n$ for $d \geq 3$ for estimating tensors that are products of monomials on any hyperrectangular partition of the tensor into z_0 pieces. The degree of the monomials is constrained to sum up at most to k on the pieces of the partition. Note that Chatterjee and Goswami [2] do not consider an ANOVA type of decomposition of the tensor into margins, so that z_0 has in general a different meaning than s_0 and can be large if compared to the sparsity of the different margins considered here.

Moreover, Chatterjee and Goswami [2] also show that the dyadic CART can attain the same fast and slow rates as the optimally tuned trend filtering estimator for $d = 1$.

Trend filtering ($d = 1$ and $k \geq 1$) has previously been studied by Guntuboyina et al. [10], yet with other proof techniques than the ones by Ortelli and van de Geer [19], which we extend here. VTF for $d = 2$ and $k = 1$ has been studied by [17]. An analogous estimator, yet in constrained form and considering the Hardy–Krause variation, has been studied by Fang et al. [6]. For the estimator constrained with the Hardy–Krause variation, Fang et al. [6] prove the same slow rate $n^{-2/3}$ as for the one-dimensional situation and the fast rate n^{-1} for estimating images with only one rectangular piece. For $f \in \mathbb{R}^{n_1 \times n_2}$ the Hardy–Krause variation is

$$\mathrm{TV}_2(f) + \mathrm{TV}_1(f(1,\cdot)) + \mathrm{TV}_1(f(\cdot,1)) + |f(1,1)|$$

and differs from our approach in the treatment of the lower-dimensional margins.

Even if the model classes of the ORT estimator by Chatterjee and Goswami [2] and of the estimator constrained with the Hardy–Krause variation by Fang et al. [6] differ from our model class, it is an interesting question whether VTF can attain the fast rate s/n and dimension-independent slow rates for estimating the d-dimensional margin of a tensor.

3.3. Optimality

To answer the question whether our rates are optimal or not, we refer to two publications. For fast rates involving the effective sparsity we refer to [30]. For slow rates we refer to [31].

3.3.1. Fast rates. The paper by van de Geer [30] shows a lower bound on the mean squared error of the Lasso in the noisy case scaling as $\lambda^2\Gamma^2$. It also shows that the bound on the effective sparsity for $d = 1$ and $k = 1$ is tight in the noiseless case. The result can be easily extended to show that also the bound on the effective sparsity for general d and $k \in \{1, 2, 3, 4\}$ is tight in the noiseless case up to constants, see Appendix E. We therefore conjecture that the only way to improve on the error bound would be by enabling a smaller choice of the tuning parameter λ. This could be achieved if we were able to find a smaller estimate for the inverse scaling factor $\tilde{\gamma}$.

3.3.2. Slow rates. The paper by van de Geer and Hinz [31] shows that the theory used to prove slow rates in this paper is optimal, up to logarithmic terms. In particular, it shows that bounds on the entropy of a class of functions imply tight bounds on the inverse scaling factor $\tilde{\gamma}$ up to log terms.

Since entropy bounds on the class of functions we consider are not known, our estimate of the inverse scaling factor $\tilde{\gamma}$ may not be tight and therefore the rate could be improved. To provide bounds on the class of functions of bounded Vitali variation is a future research question in the field of approximation theory. This could help us answer the question whether the optimal rate for the estimation of the d-dimensional margin of a tensor depends on d or not. For a standard Hölder class with smoothness k the minimax rate would be $n^{-2k/(2k+d)}$, while for the Hölder class with smoothness dk the minimax rate would be $n^{-2k/(2k+1)}$ for all $d \geq 1$. The rate provided by Theorem 3.2 lies in between these two rates. Intuitively, requiring bounded kth-order Vitali variation of the d-dimensional margin is a weaker constraint than the requirement of the Hölder class with smoothness dk, since only derivatives of order k are needed in all directions. However, since the class of functions with bounded Vitali total variation is not a standard class it is not yet clear which minimax rate to expect.

The approach we present here penalizes differences of the same order along all coordinates and can thus be seen as an isotropic approach. However, one could adapt the framework we present to penalize differences of possibly different orders along different coordinates in an anisotropic approach. This would allow to handle more general smoothness classes too.

4. Synthesis form

According to Definition 2.3, the trend filtering estimator is an analysis estimator. In this section we want to rewrite it in a constructive form, that is, in synthesis form. We show that the trend filtering estimator can be constructed as a linear combination of tensors with product structure, where the factors are truncated monomials of order $k-1$. We call the collection of such tensors the "dictionary".

We first define the dictionary and then show that it is the right dictionary to construct the trend filtering estimator.

We start with the one-dimensional case. We then obtain the d-dimensional dictionary from the one-dimensional dictionary by constructing tensors with product structure.

4.1. Dictionary for $d = 1$

Let $\phi_j^1 := \{1_{\{j' \geq j\}}\}_{j' \in [n]}$, $j \in [n]$. The vectors $\{\phi_j^1\}_{j \in [n]}$ are linearly independent and piecewise constant.

For $2 \leq k \leq n-1$, define recursively

$$\phi_j^k := \begin{cases} \phi_j^j, & j \in [k-1], \\ \sum_{l \geq j} \phi_l^{k-1}/n, & j \in [k:n]. \end{cases}$$

We call the collection $\Phi^k = \{\phi_j^k\}_{j \in [n]}$ the "original" dictionary.

The dictionary Φ^k is a collection of n linearly independent discrete (truncated) monomials: the first k are monomials of order $0, 1, \ldots, k-1$, while the last $n-k$ are truncated monomials of order $k-1$.

We now define a partially orthonormalized version of the dictionary Φ^k, $k \in [n-1]$.

Definition 4.1 (Partially orthonormalized dictionary in one dimension). The (partially orthonormalized) dictionary $\widetilde{\Phi}^k = \{\widetilde{\phi}_j^k\}_{j\in[n]}$ is defined as

$$\widetilde{\phi}_j^k := \begin{cases} \sqrt{n}\mathrm{A}_{\{\phi_l^l, l\in[j-1]\}}\phi_j^j / \|\mathrm{A}_{\{\phi_l^l, l\in[j-1]\}}\phi_j^j\|_2, & j \in [k], \\ \mathrm{A}_{\{\phi_l^l, l\in[k]\}}\phi_j^k, & j \in [k+1:n]. \end{cases}$$

For $k \in [n-1]$, $\widetilde{\Phi}^k = \{\widetilde{\phi}_j^k\}_{j\in[n]}$ is again a collection of n linearly independent vectors, where $\widetilde{\phi}_1^k, \ldots, \widetilde{\phi}_k^k, \{\widetilde{\phi}_j^k\}_{j\in[k+1:n]}$ are mutually orthogonal. Moreover,

$$\|\widetilde{\phi}_j^k\|_2^2 = n, \quad j \in [k].$$

Lemma 4.2 (Relation between dictionary and difference operator). *Fix $k \in [n-1]$. It holds that*

$$D^k\phi_j^k = D^k\widetilde{\phi}_j^k = \begin{cases} 0, & j \in [k], \\ 1_{\{j\}}, & j \in [k+1:n]. \end{cases}$$

Proof. See Appendix B.1. ∎

As a consequence of Lemma 4.2, $\{\widetilde{\phi}_j^j\}_{j\in[k]}$ is an orthogonal basis for $\mathcal{N}_k$. Moreover, $\{\widetilde{\phi}_j^k\}_{j\in[k+1:n]}$ span $\mathcal{N}_k^\perp$.

By Lemma 4.2 combined with Lemma 2.2 in [15] about the Moore–Penrose pseudoinverse we obtain for the pseudoinverse $(D^k)^+$ that $(D^k)^+ = \{\widetilde{\phi}_j^k\}_{j\in[k+1:n]}$.

With the dictionary $\widetilde{\Phi}^k$ and some coefficients $\{\beta_j\}_{j=k+1}^n$ we can write a vector $f_{\mathcal{N}_k^\perp} \in \mathcal{N}_k^\perp$ as $f_{\mathcal{N}_k^\perp} = (D^k)^+\beta$. Then $\beta = D^k f_{\mathcal{N}_k^\perp}$.

For $d = 1$, we therefore obtain the following synthesis form of the estimator $\widehat{f}_{\mathcal{N}_k^\perp}$:

$$\widehat{f}_{\mathcal{N}_k^\perp} = \sum_{j=k+1}^{n} \widetilde{\phi}_j^k \widehat{\beta}_j,$$

where

$$\widehat{\beta} = \arg\min_{b\in\mathbb{R}^{n-k}} \Big\{ \|Y_{\mathcal{N}_k^\perp} - \sum_{j=k+1}^{n} b_j\widetilde{\phi}_j^k\|_2^2/n + 2\lambda\|b\|_1 \Big\}.$$

4.2. Dictionary for general d

Hereafter we fix $k \in [1 : \min_{l\in[d]} n_l - 1]$.

Definition 4.3 (Partially orthonormalized dictionary in d-dimensions). The dictionary $\{\widetilde{\phi}^k_{j_1,\dots,j_d} \in \mathbb{R}^{n_1\times\dots\times n_d}\}_{1,\dots,1}^{n_1,\dots,n_d}$ is defined as

$$\widetilde{\phi}^k_{j_1,\dots,j_d} = \widetilde{\phi}^k_{j_1} \times \dots \times \widetilde{\phi}^k_{j_d}, \quad (j_1, \dots, j_d) \in I.$$

The dictionary $\{\widetilde{\phi}^k_{j_1,\dots,j_d}\}_{1,\dots,1}^{n_1,\dots,n_d}$ is a collection of d-dimensional tensors with product structure. By Lemma 4.2 and the product structure,

$$\mathcal{N}_k^{\perp} = \text{span}\left(\{\widetilde{\phi}^k_{j_1,\dots,j_d}\}_{k+1,\dots,k+1}^{n_1,\dots,n_d}\right).$$

For a tensor of coefficients $\{\beta_{j_1,\dots,j_d}\}_{k+1,\dots,k+1}^{n_1,\dots,n_d}$, write

$$f_{\mathcal{N}_k^{\perp}} = \sum_{k+1,\dots,k+1}^{n_1,\dots,n_d} \beta_{j_1,\dots,j_d} \widetilde{\phi}^k_{j_1,\dots,j_d}.$$

Because of the product structure of $\widetilde{\phi}^k_{j_1,\dots,j_d}$ it holds that

$$D^k f_{\mathcal{N}_k^{\perp}} = \sum_{k+1,\dots,k+1}^{n_1,\dots,n_d} \beta_{j_1,\dots,j_d} \left(1_{\{j_1\}} \times \dots \times 1_{\{j_d\}}\right) = \beta.$$

From the fact that any candidate estimator has to belong to the space spanned by $Y_{\mathcal{N}_k^{\perp}}$, it follows that

$$\widehat{f}_{\mathcal{N}_k^{\perp}} = \sum_{k+1,\dots,k+1}^{n_1,\dots,n_d} \widehat{\beta}_{j_1,\dots,j_d} \widetilde{\phi}^k_{j_1,\dots,j_d},$$

where

$$\widehat{\beta} = \arg\min_{b\in\mathbb{R}^{(n_1-k)\times\dots\times(n_d-k)}} \left\{\|Y_{\mathcal{N}_k^{\perp}} - \sum_{k+1,\dots,k+1}^{n_1,\dots,n_d} b_{j_1,\dots,j_d} \widetilde{\phi}^k_{j_1,\dots,j_d}\|_2^2/n + 2\lambda\|b\|_1\right\}.$$

The synthesis form of the estimator $\widehat{f}_{\mathcal{N}_k^{\perp}}$ is useful in two ways. Firstly, to determine the structure of the estimator by specifying the dictionary used to construct it. In our case, $\widehat{f}_{\mathcal{N}_k^{\perp}}$ is a linear combination of d-dimensional products of $(k-1)$th-order polynomials. Secondly, the dictionary facilitates the approximation of some orthogonal projections in the proof of adaptive ℓ^0-rates and not-so-slow ℓ^1-rates.

5. Adaptivity

In this section we first expose some notation for our main result. After having exposed our main result, Theorem 5.2, we work out explicit expressions for the bound on the antiprojections $\tilde{v}$, the inverse scaling factor $\tilde{\gamma}$ and the noise weights v. Finally, we show a bound on the effective sparsity via a suitable interpolating tensor. In Section 5.7 we put the pieces together to prove Theorem 5.2.

Fix $k \in [1 : \min_{i\in[d]} n_i - 1]$ and an active set $S \subseteq \times_{i\in[d]}[k+2 : n_i - k]$.

To every jump location in S, we associate a hyperrectangle of k^d additional jump locations to obtain the enlarged active set $\tilde{S}$, defined as

$$\tilde{S} := \bigcup_{m=1}^{s} \left(\times_{i\in[d]}[t_{i,m} : t_{i,m} + k - 1]\right).$$

Definition 5.1 (Hyperrectangular tessellation). We call $\{R_m\}_{m=1}^s$ a hyperrectangular tessellation of $\times_{i\in[d]}[k+1 : n_i]$ if it satisfies the following conditions:

- each $R_m \subseteq \times_{i\in[d]}[k+1 : n_i]$ is a hyperrectangle for $m \in [s]$;
- $\cup_{m=1}^s R_m = \times_{i\in[d]}[k+1 : n_i]$;
- for all m and $m' \neq m$, the hyperrectangles R_m and $R_{m'}$ possibly share boundary points but not interior points;
- for all m, the points $\times_{i\in[d]}[t_{i,m} : t_{i,m} + k - 1]$ are interior points of R_m.

For a hyperrectangular tessellation $\{R_m\}_{m=1}^s$ we denote the vertices of the hyperrectangle R_m by $(t_{1,m}^{z_1}, \dots, t_{d,m}^{z_d})$, $(z_1, \dots, z_d) \in \{-, +\}^d$ for $m \in [s]$.

Moreover, we define the distances of the jump locations from the vertices of their respective hyperrectangle and the respective set of indices as

$$\begin{aligned}
d_{i,m}^- &:= (t_{i,m} - t_{i,m}^-), & R_{i,m}^- &:= [t_{i,m}^- : t_{i,m}],\\
d_{i,m}^0 &:= k, & R_{i,m}^0 &:= [t_{i,m} : t_{i,m} + k - 1],\\
d_{i,m}^+ &:= (t_{i,m}^+ - t_{i,m} - k + 1), & R_{i,m}^+ &:= [t_{i,m} + k - 1 : t_{i,m}^+],
\end{aligned}$$

for $i \in [d]$ and $m \in [s]$. Each hyperrectangle R_m of the hyperrectangular tessellation $\{R_m\}_{m\in[s]}$ can be partitioned into 3^d hyperrectangles. Define, for all $(z_1, \dots, z_d) \in \{-, 0, +\}^d$,

$$R_m^{z_1 \cdots z_d} := R_{1,m}^{z_1} \times \cdots \times R_{d,m}^{z_d}, \quad m \in [s].$$

For $m \in [s]$, let

$$d_m^{z_1 \cdots z_d} := d_{1,m}^{z_1} \cdot \ldots \cdot d_{d,m}^{z_d}, \quad \{z_1, \dots, z_d\} \in \{-, +\}^d.$$

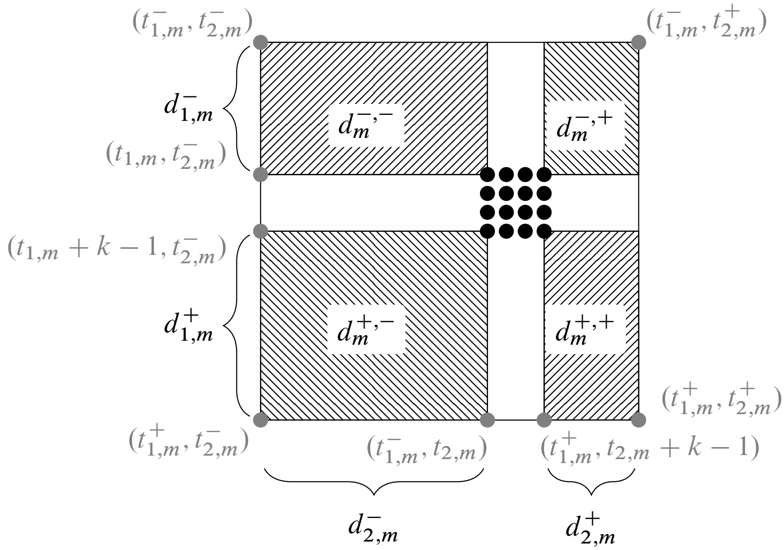


Figure 1. A rectangle of the tessellation $\{R_m\}_{m=1}^s$ for $d = 2$ and $k = 4$.

We define the maximal distance from an (enlarged) jump location to the boundary of the corresponding rectangular region along the coordinate $i \in [d]$ as

$$d_{i,\max}(S) := \max_{m \in [1:s]} \max\{d^-_{i,m}, d^+_{i,m}\}.$$

For $d = 2$ and $k = 4$, a rectangle of the tessellation is depicted in Figure 1.

5.1. Main result

We present our main result, that shows that Vitali trend filtering leads to an adaptive estimation of the d-dimensional margin $f^0_{\mathcal{N}_k^\perp}$ of f^0.

Theorem 5.2 (Adaptivity of Vitali trend filtering). *Fix $k \in \{1,2,3,4\}$ and choose $x, t > 0$. Let $g \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be arbitrary. Let S be an arbitrary subset of size $s := |S|$ of $\times_{i \in [d]}[k + 1 + (k + 2)k : n_i - k + 1 - (k + 2)k]$. For a large enough constant $C > 0$ that only depends on k, choose*

$$\lambda \geq C\, d \sqrt{\sum_{i=1}^{d} \left(\frac{d_{i,\max}(S)}{n_i}\right)^{2k-1}}\, \lambda_0(t).$$

Then, with probability at least $1 - e^{-x} - e^{-t}$, *it holds that*

$$\|(\hat{f} - f^0)_{\mathcal{N}_k^\perp}\|_2^2/n \leq \|g - f^0_{\mathcal{N}_k^\perp}\|_2^2/n + 4\lambda\|(D^k g)_{-S}\|_1 + \frac{2\sigma^2}{n}\left(\sqrt{x} + \sqrt{ks}\right)^2 + \mathcal{O}\left(\lambda^2\left(\sum_{i=1}^d \log\left(ed_{i,\max}(S)\right)\right)\sum_{m=1}^s \sum_{z\in\{-,+\}^d}\left(\frac{n}{d_m^z}\right)^{2k-1}\right).$$

In particular, the constraint on C *is*

$$C \geq \frac{k^{(2k-1)/2}}{a_0} \quad \textit{with } a_0 = \begin{cases} 1, & k = 1, \\ 8\sqrt{2}/7 \approx 1.62, & k = 2, \\ 144\sqrt{3}/76 \approx 3.28, & k = 3, \\ 10.10, & k = 4, \end{cases}$$

as $\min_{i\in[d]} \min_{m\in[s]} \min\{d_{i,m}^-, d_{i,m}^+\} \to \infty$.

Proof. See Section 5.7. ∎

By choosing $x \asymp t \asymp \log n$ in Theorem 5.2 and by constraining the active set S to be a regular grid we retrieve Theorem 3.1. In that case, since S is a regular grid, we can choose $\lambda \asymp s^{-(2k-1)/2d}\lambda_0(\log(2n))$ and the oracle inequality has the rate

$$\mathcal{O}\left(\frac{s^{2k-((2k-1)/d)}}{n}\log(n/s)\log n\right).$$

Remark 5.3 (The role of the hyperrectangular tessellation). Given an active set S, the choice of a hyperrectangular tessellation in Theorem 5.2 can be seen as arbitrary.

Remark 5.4 (Minimum length condition). In opposition to results for $d = 1$ by Guntuboyina et al. [9], Theorem 5.2 does not explicitly require a minimum length condition on the active set S. For $d = 1$, we can interpret the minimum length condition as a condition saying that $d_{1,1}^- \asymp d_{1,1}^+ \asymp \cdots \asymp d_{s,1}^- \asymp d_{s,1}^+$. If S approximately satisfies a minimum length condition, then the estimation error has a faster rate than if S does not satisfy the minimum length condition.

Theorem 5.2 is an oracle inequality and S can be interpreted as a free parameter. If the true active set S_0 of f^0 does not satisfy some kind of minimum length condition, then the upper bound of Theorem 5.2 still holds for a choice of S which could give place to a faster rate for the estimation error by being closer to satisfying the minimum length condition at the price of allowing for an approximation error of f^0.

Note that also the lower bound on the choice of the tuning parameter λ depends on the structure of the active set S and is smaller when the elements of S approximately satisfy a minimum length condition.

5.2. Some definitions

We introduce some quantities on which Theorem 5.2 relies: the bound on the antiprojections $\widetilde{v}$, the inverse scaling factor $\widetilde{\gamma}$, the noise weights v, a sign configuration q and the effective sparsity $\Gamma^2_{D^k}$.

Let $\widetilde{S}$ be the enlarged active set induced by some active set S. Let $\mathrm{P}_{\widetilde{S}}$ be the orthogonal projection operator on $\text{span}(\{\widetilde{\phi}^k_{j_1,\dots,j_d}\}_{(j_1,\dots,j_d)\in\widetilde{S}})$.

Definition 5.5 (Bound on the antiprojections). A bound on the antiprojections is a tensor $\widetilde{v} \in \mathbb{R}^{(n_1-k)\times\dots\times(n_d-k)}$ such that

$$\widetilde{v}_{j_1,\dots,j_d} \geq \|(\mathrm{I}-\mathrm{P}_{\widetilde{S}})\widetilde{\phi}^k_{j_1,\dots,j_d}\|_2/\sqrt{n}, \quad \forall (j_1,\dots,j_d) \in \times_{i\in[d]}[k+1:n_i].$$

Let $\widetilde{v}$ be a bound on the antiprojections.

Definition 5.6 (Inverse scaling factor). The inverse scaling factor $\widetilde{\gamma} \in \mathbb{R}$ is defined as $\widetilde{\gamma} := \|\widetilde{v}_{-\widetilde{S}}\|_\infty$.

Let $\widetilde{v}$ be a bound on the antiprojections and $\tilde{\gamma}$ the corresponding inverse scaling factor.

Definition 5.7 (Noise weights). The noise weights $v \in \mathbb{R}^{(n_1-k)\times\dots\times(n_d-k)}$ are defined as $v \geq \widetilde{v}/\widetilde{\gamma} \in [0,1]^{(n_1-k)\times\dots\times(n_d-k)}$.

We can now introduce the effective sparsity. The effective sparsity depends on a so-called "sign configuration", that is, on the sign pattern associated with the jump locations.

Definition 5.8 (Sign configuration). Let $q \in [-1,1]^{(n_1-k)\times\dots\times(n_d-k)}$ be such that

$$q_{j_1,\dots,j_d} \in \begin{cases} \{-1,+1\}, & (j_1,\dots,j_d) = t_m \in S, \\ \{q_{t_m}\}, & (j_1,\dots,j_d) \in \times_{i\in[d]}[t_{i,m} : t_{i,m}+k-1],\ m \in [s], \\ [-1,1], & (j_1,\dots,j_d) \notin \widetilde{S}. \end{cases}$$

We call $q_S \in \{-1,0,1\}^{(n_1-k)\times\dots\times(n_d-k)}$ a sign configuration.

The basic definition of effective sparsity depends on the sign configuration associated with S. One can however remove this dependence by defining the effective sparsity as the maximum over all sign configurations.

Definition 5.9 (Effective sparsity). Let an active set S, a sign configuration q_S and noise weights v be given. The effective sparsity $\Gamma^2_{D^k}(S, v_{-S}, q_S) \in \mathbb{R}$ is defined as

$$\Gamma_{D^k}(S, v_{-S}, q_S)$$
$$:= \max\left\{ \sum_{m=1}^{s} (q_S)_{t_m} (D^k f)_{t_m} - \|(1-v)_{-S} \odot (D^k f)_{-S}\|_1 : \|f\|_2^2/n = 1 \right\}.$$

Moreover, we write

$$\Gamma^2_{D^k}(S, v_{-S}) := \max_{q_S} \Gamma^2_{D^k}(S, v_{-S}, q_S).$$

By the adaptive bound of Theorem 2.2 in [19] (see also Theorem 2.1 in [18] and Theorem 16 in [17] modified with an enlarged active set), we know that bounding the effective sparsity is a sufficient condition for proving adaptation of $\widehat{f}_{\mathcal{N}_k^\perp}$.

5.3. Effective sparsity via interpolating tensors

To bound the effective sparsity we extend the technique by Ortelli and van de Geer [19] involving interpolating vectors to interpolating tensors, i.e., tensors that interpolate the signs of the jumps.

Definition 5.10 (Interpolating tensor). Let $q_S \in \{-1, 0, 1\}^{(n_1-k)\times\cdots\times(n_d-k)}$ be a sign configuration and $v \in [0,1]^{(n_1-k)\times\cdots\times(n_d-k)}$ be a tensor of noise weights. The tensor $w(q_S) \in \mathbb{R}^{(n_1-k)\times\cdots\times(n_d-k)}$ is called an interpolating tensor for the sign configuration q_S and the weights v if it has the following properties:

- $w_{j_1,\ldots,j_d}(q_S) = (q_S)_{t_m}$, $\forall (j_1, \ldots, j_d) \in \times_{i\in[d]}[t_{i,m} : t_{i,m} + k - 1]$, $\forall m \in [s]$,
- $|w_{j_1,\ldots,j_d}(q_S)| \le 1 - v_{j_1,\ldots,j_d}$, $\forall (j_1, \ldots, j_d) \in \left(\times_{i\in[d]}[k : n_i]\right) \setminus \widetilde{S}$.

With the help of an interpolating tensor we can bound the effective sparsity, as the following lemma shows ([19, Lemma 2.4] in tensor form).

Lemma 5.11 (Bounding the effective sparsity with an interpolating tensor). *We have*

$$\Gamma^2_{D^k}(S, v_{-S}, q_S) \le n \min_{w(q_S)} \|(D^k)' w(q_S)\|_2^2,$$

where the minimum is over all interpolating tensors $w(q_S)$ for the sign configuration q_S.

Proof. It holds that

$$\begin{aligned}
&\sum_{m=1}^{s} (q_S)_{t_m} (D^k f)_{t_m} - \|(1-v)_{-S} \odot (D^k f)_{-S}\|_1 \\
&\quad \le \sum_{m=1}^{s} (q_S)_{t_m} (D^k f)_{t_m} - \|w(q_S)_{-S} \odot (D^k f)_{-S}\|_1 \\
&\quad \le \sum_{2,\dots,2}^{n_1,\dots,n_d} w(q_S)_{j_1,\dots,j_d} (D^k f)_{j_1,\dots,j_d} \\
&\quad = \sum_{1,\dots,1}^{n_1,\dots,n_d} ((D^k)' w(q_S))_{j_1,\dots,j_d} f_{j_1,\dots,j_d} \\
&\quad \le \sqrt{n} \|(D^k)' w(q_S)\|_2 \|f\|_2 / \sqrt{n}.
\end{aligned}$$

■

5.4. Requirements on an interpolating tensor

Theorem 5.2 follows by a bound on the effective sparsity obtained by Lemma 5.11 with the help of an interpolating tensor. In the definition of an interpolating tensor (cf. Definition 5.10), there is a constraint posed by the noise weights v.

Therefore, we now calculate in Section 5.5 a bound on the antiprojections $\widetilde{v}$ to derive an appropriate inverse scaling factor $\widetilde{\gamma}$ and noise weights v. In this way we will make explicit the constraints that the interpolating tensor has to satisfy in the specific case of tensor denoising with trend filtering.

After that, we will show in Section 5.6 an explicit form for the interpolating tensor for $k \in \{1, 2, 3, 4\}$ and derive the corresponding bound on the effective sparsity.

That bound on the effective sparsity combined with the fact that the interpolating tensor used indeed is an interpolating tensor for trend filtering will allow us to derive Theorem 5.2 from Theorem A.1.

5.5. Antiprojections, inverse scaling factor and noise weights

We start by finding a bound on the antiprojections $\widetilde{v}$.

Define, for $m \in [s]$ and $i \in [d]$,

$$\widetilde{v}^2_{i,m}(j_i) = \begin{cases} \left(\dfrac{t_{i,m} - j_i}{n_i}\right)^{2k-1}, & j_i \in R^-_{i,m} = [t^-_{i,m} : t_{i,m}], \\ 0, & j_i \in R^0_{i,m} = [t_{i,m} : t_{i,m} + k - 1], \\ \left(\dfrac{j_i - t_{i,m} - k + 1}{n_i}\right)^{2k-1}, & j_i \in R^+_{i,m} = [t_{i,m} + k - 1 : t^+_{i,m}]. \end{cases}$$

Moreover, for $(j_1, \dots, j_d) \in R_m$ we define

$$\widetilde{v}_{j_1,\dots,j_d} := \sqrt{\sum_{i=1}^{d} \widetilde{v}_{i,m}^2(j_i)}\,.$$

Lemma 5.12 (A valid bound on the antiprojections). *For all $(j_1, \dots, j_d) \in R_m$ and for all $m \in [s]$ it holds that*

$$\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_{j_1,\dots,j_d}\|_2^2/n \le \widetilde{v}^2_{j_1,\dots,j_d},$$

i.e., the tensor $\widetilde{v} \in \mathbb{R}^{(n_1-k)\times\cdots\times(n_d-k)}$ is a valid bound on the antiprojections.

Proof. See Appendix C.1. ∎

Define, for $m \in [s]$ and $i \in [d]$,

$$v_{i,m}^2(j_i) = \begin{cases} \left(\dfrac{t_{i,m} - j_i}{d^-_{i,m}}\right)^{2k-1}, & j_i \in R^-_{i,m} = [t^-_{i,m} : t_{i,m}], \\ 0, & j_i \in R^0_{i,m} = [t_{i,m} : t_{i,m} + k - 1], \\ \left(\dfrac{j_i - t_{i,m} - k + 1}{d^+_{i,m}}\right)^{2k-1}, & j_i \in R^+_{i,m} = [t_{i,m} + k - 1 : t^+_{i,m}]. \end{cases}$$

For a constant $C = C(k) \ge 1$, we define for $(j_1, \dots, j_d) \in R_m$ and $m \in [s]$

$$v_{j_1,\dots,j_d} := \frac{1}{d} \sum_{i=1}^{d} \frac{v_{i,m}(j_i)}{C} \tag{5.1}$$

and

$$\widetilde{\gamma} = Cd \sqrt{\sum_{i=1}^{d} \left(\frac{d_{i,\max}(S)}{n_i}\right)^{2k-1}}\,.$$

Lemma 5.13 (Valid noise weights). *For all $m \in [s]$ and for all $(j_1, \dots, j_d) \in R_m$ it holds that*

$$\widetilde{v}_{j_1,\dots,j_d} \le v_{j_1,\dots,j_d} \widetilde{\gamma},$$

i.e., the tensor $v \in \mathbb{R}^{(n_1-k)\times\cdots\times(n_d-k)}$ in equation (5.1) defines valid noise weights.

Proof. See Appendix C.2. ∎

The constant $C \ge 1$ in equation (5.1) can be chosen arbitrarily. Choosing a larger C makes the noise weights smaller. As a result, the requirements imposed on the interpolating tensor by the noise weights become weaker.

5.6. Bound on the effective sparsity for trend filtering

We now define an interpolating tensor $w = w(q_S)$ for any sign configuration q_S.

For $(j_1, \dots, j_d) \in R_m, m \in [s]$ and the same constant $C = C(k) > 0$ as in the definition of the noise weights in equation (5.1), define the tensor

$$w_{j_1,\dots,j_d}(q_S) := \frac{1}{d} \sum_{i=1}^{d} \prod_{l=1}^{d} w_{l,i,m}(j_l), \tag{5.2}$$

where,

$$\begin{aligned}
w_{l,i,m}(j_l) &= q_{t_m}, & j_l &\in R^0_{l,m}, & l &\neq i,\\
w_{l,i,m}(j_l) &\in [0, q_{t_m}], & j_l &\in R^-_{l,m} \cup R^+_{l,m}, & l &\neq i,\\
w_{i,i,m}(j_i) &= q_{t_m}, & j_i &\in R^0_{i,m}, & l &= i,\\
|w_{i,i,m}(j_i)| &\leq 1 - v_{i,m}(j_i)/C, & j_i &\in R^-_{l,m} \cup R^+_{l,m}, & l &= i.
\end{aligned}$$

What differentiates the case $l = i$ is that $w_{i,i,m}$ has to satisfy the requirements imposed by the noise weights. For $l \neq i$ the only constraint imposed is that $|w_{l,i,m}| \leq 1$. The tensor w is a sum of terms with product structure if constrained to the set of indices of a hyperrectangle R_m.

We define $w^-_{l,i,m} := \{w_{l,i,m}(j_l)\}_{j_l \in R^-_{i,m}}$ and $w^+_{l,i,m} := \{w_{l,i,m}(j_l)\}_{j_l \in R^+_{i,m}}$.

Lemma 5.14 (A valid interpolating tensor). *For any given sign configuration q_S, the tensor $w = w(q_S)$ defined in equation (5.2) is a valid interpolating tensor.*

Proof. See Appendix C.3. ∎

5.6.1. Matching derivatives. We now want to find the explicit form of an appropriate interpolating tensor w, to apply in Lemma 5.11. We first consider continuous versions $\omega(x)$, respectively $\mathrm{w}(x)$, of the vectors $w^-_{i,i,m}$ and $w^+_{i,i,m}$, respectively $w^-_{l,i,m}$ and $w^+_{l,i,m}$ for $l \neq i$, on a mock interval $x \in [0, 1]$. We then set

$$\begin{aligned}
w^-_{i,i,m}(j_i) &:= \omega\left(\frac{t_{i,m} - j_i}{d^-_{i,m}}\right), & j_i &\in R^-_{i,m},\\
w^+_{i,i,m}(j_i) &:= \omega\left(\frac{j_i - t_{i,m} - k + 1}{d^+_{i,m}}\right), & j_i &\in R^+_{i,m},\\
w^-_{l,i,m}(j_l) &:= \mathrm{w}\left(\frac{t_{l,m} - j_l}{d^-_{l,m}}\right), & j_l &\in R^-_{l,m},\\
w^+_{l,i,m}(j_l) &:= \mathrm{w}\left(\frac{j_l - t_{l,m} - k + 1}{d^+_{l,m}}\right), & j_l &\in R^+_{l,m},
\end{aligned}$$

for $i \in [d], l \neq i, m \in [s]$.

We aim to find a form of ω and w giving place to continuous functions with $k-1$ continuous derivatives and piecewise constant kth derivative. Moreover, these functions have to be interpolating between the jump location ($x=0$) and the border ($x=1$). We guarantee that they interpolate the signs of the jumps by restricting to polynomials with

$$\begin{aligned}&\omega(0)=1, \quad \omega(1)=0,\\ &\mathrm{w}(0)=1, \quad \mathrm{w}(1)=0, \quad \mathrm{w}(x)=1-\mathrm{w}(1-x),\ x\in[0,1].\end{aligned}$$

The discretized version of these polynomials will vanish at the boundaries of the hyperrectangles while it will have the value 1 at the indices belonging to the enlarged active set $\widetilde{S}$, guaranteeing the interpolation of the signs of the jump locations. Moreover, we will have to choose the constant $C>0$ in equation (5.1) such that the noise weights are made small enough for the interpolating polynomial to satisfy the conditions of Lemma 5.14.

To obtain interpolating polynomials ω and w, we split the interval $[0,1]$ into an adequate number of subintervals. We then choose ω and w to be made of polynomial pieces of order at most k. The exception is the first subinterval $[0,x_1]$, $x_1\in(0,1]$ for ω, where we choose $\omega(x)=1-a_0x^{(2k-1)/2}$. We then find the explicit values of the coefficients of the polynomials by derivative matching, as in [19]. More details on derivative matching are given in the Appendix C.4.

To guarantee that ω and w can give place to interpolating tensors, one has to check that derivative matching renders a piecewise polynomial which is monotone. Monotonicity combined with the constraints $\omega(0)=\mathrm{w}(0)=1$ and $\omega(1)=\mathrm{w}(1)=0$ ensures that $|\omega(x)|\le 1$ and $|\mathrm{w}(x)|\le 1$.

Monotone interpolating polynomials ω and w and a large enough C in the tuning parameter are sufficient conditions for a valid interpolating tensor. In particular, given that ω is monotone, we require that

$$C\ge k^{\frac{2k-1}{2}}/a_0 \quad \text{as} \min_{i\in[d]}\ \min_{m\in[s]}\ \min\{d^-_{i,m},d^+_{i,m}\}\to\infty. \tag{5.3}$$

Note that for the construction of w, we do not have any constraint given by the antiprojections $\widetilde{v}$, the noise weights v and the inverse scaling factor $\widetilde{\gamma}$. Therefore, we can take the dependence on x^k instead of $x^{(2k-1)/2}$. This saves a logarithmic term, not visible in Lemma 5.15, which only contains the logarithmic terms stemming from ω. Indeed, as Lemma C.1 in Appendix C.5 shows, partial integration of a kth-order polynomial does not incur in log terms, while partial integration of $x^{(2k-1)/2}$ does so. We have to choose the worse dependence on $x^{(2k-1)/2}$ for ω though, because ω has to respect the constraints posed by the noise weights.

5.6.2. A bound on the effective sparsity. We now show a bound on the effective sparsity, using a "candidate" interpolating tensor generated from the discretizations of ω and w, whose construction has been exposed above. We call it "candidate" interpolating tensor because we have not yet shown that ω and w are monotone. For the moment we assume that matching derivatives renders monotone ω and w. We check the monotonicity for $k \in \{1, 2, 3, 4\}$ in the next section.

To make the notation and the computation steps lighter, we neglect the constants and use the order notation $\mathcal{O}$ instead.

Since the sign configuration q_S is typically unknown, we focus on finding an upper bound on the effective sparsity that does not depend on the sign configuration q_S. Thus, the bound also accommodates for the worst-case sign configuration.

Lemma 5.15 (Effective sparsity for trend filtering). *Take the interpolating vector w as defined in equation (5.2). Choose the vectors $w^-_{i,i,m}$ and $w^+_{i,i,m}$, respectively $w^-_{l,i,m}$ and $w^+_{l,i,m}$ for $l \neq i$, to be discretized versions of $\omega(x)$ and $\mathrm{w}(x)$ as in Section 5.6.1. Assume that $\omega(x)$ and $\mathrm{w}(x)$ obtained by derivative matching are monotone.*

For such an interpolating vector w, it holds that

$$\Gamma^2_D(S, v_{-S}) = \mathcal{O}\left(\left(\sum_{i=1}^{d} \log(ed_{i,\max}(S))\right) \sum_{m=1}^{s} \sum_{z\in\{-,+\}^d} \left(\frac{n}{d^z_m}\right)^{2k-1}\right).$$

Proof. See Appendix C.6. ∎

From Lemma C.1 and the matching of discrete derivatives, it follows that, if ω and w are monotone and C is chosen large enough

$$\Gamma^2_D(S, v_{-S}) = \mathcal{O}\left(\left(\sum_{i=1}^{d} \log(ed_{i,\max}(S))\right) \sum_{m=1}^{s} \sum_{z\in\{-,+\}^d} \left(\frac{n}{d^z_m}\right)^{2k-1}\right).$$

If the active set S defines a regular grid we therefore have a bound on the effective sparsity of order

$$\Gamma^2_D(S, v_{-S}) = \mathcal{O}\big(s^{2k} \log(n/s)\big).$$

It only remains to check the monotonicity of ω and w. We will do this for $k \in \{1, 2, 3, 4\}$. One can check monotonicity for higher values of k by solving (for instance at the computer) the appropriate system of equations and, say, graphically visualizing the result. We check monotonicity analytically for $k \in \{1, 2, 3\}$ and computationally for $k = 4$.

5.6.3. Interpolating tensor for $k = 1$. For $k = 1$

$$\omega(x) = 1 - \sqrt{x}, \quad x \in [0, 1],$$
$$\mathrm{w}(x) = 1 - x, \quad x \in [0, 1].$$

Both ω and w are monotone.

5.6.4. Interpolating tensor for $k = 2$. For $k = 2$

$$\omega(x) = \begin{cases} 1 - \frac{8\sqrt{2}}{7} x^{3/2}, & x \in [0, 1/2], \\ \frac{12}{7}(1-x)^2, & x \in [1/2, 1], \end{cases}$$
$$\mathrm{w}(x) = \begin{cases} 1 - \frac{8}{3} x^2, & x \in [0, 1/4], \\ \frac{4}{3}\left(\frac{1}{2} - x\right) + \frac{1}{2}, & x \in [1/4, 1/2]. \end{cases}$$

Both ω and w are monotone.

5.6.5. Interpolating tensor for $k = 3$. For $k = 3$

$$\omega(x) = \begin{cases} 1 - \frac{144\sqrt{3}}{76} x^{5/2}, & x \in [0, 1/3], \\ \frac{585}{76} x^3 - \frac{45}{4} x^2 + \frac{255}{76} x + \frac{145}{228}, & x \in [1/3, 2/3], \\ \frac{315}{76}(1-x)^3, & x \in [2/3, 1], \end{cases}$$
$$\mathrm{w}(x) = \begin{cases} 1 - \frac{16}{3} x^3, & x \in [0, 1/4], \\ -\frac{16}{3}\left(\frac{1}{2} - x\right)^3 + 2\left(\frac{1}{2} - x\right) + \frac{1}{2}, & x \in [1/4, 1/2]. \end{cases}$$

Both ω and w are monotone.

5.6.6. Interpolating tensor for $k = 4$. For $k = 4$

$$\omega(x) = \begin{cases} 1 - 7.29 x^{7/2}, & x \in [0, 1/4], \\ 27.39x^4 - 35.36x^3 + 12.26x^2 - 2.01x + 1.12, & x \in [1/4, 1/2], \\ -29.51x^4 + 78.43x^3 - 73.08x^2 + 26.44x - 2.43, & x \in [1/2, 3/4], \\ 10.10(1-x)^4, & x \in [3/4, 1], \end{cases}$$
$$\mathrm{w}(x) = \begin{cases} 1 - 16.2x^4, & x \in [0, 1/6], \\ 27x^4 - 28.8x^3 + 7.2x^2 - 0.8x + 1.03, & x \in [1/6, 1/3], \\ -7.2\left(\frac{1}{2} - x\right)^3 + 2.2\left(\frac{1}{2} - x\right) + \frac{1}{2}, & x \in [1/3, 1/2]. \end{cases}$$

Both ω and w are monotone.

5.7. Proof of Theorem 5.2

Theorem 5.2 follows by combining Theorem A.1 with a bound on the effective sparsity.

Lemma 5.15 uses Lemma 5.11 to give a bound on the effective sparsity holding for all sign configurations. This bound is based on a specific form of the interpolating tensor, obtained by derivative matching as explained in Section 5.6.1. The interpolating tensor obtained by derivative matching is valid if the monotonicity of ω and w is guaranteed. In Sections 5.6.3–5.6.6 we check that the interpolating tensors obtained by derivative matching for $k = \{1, 2, 3, 4\}$ satisfy the monotonicity requirement.

An interpolating vector also has to satisfy a constraint posed by the noise weights v and by the constant C. Lemma 5.12 gives a valid bound on the antiprojections. If we choose

$$\tilde{\gamma} = Cd\sqrt{\sum_{i=1}^{d}\left(\frac{d_{i,\max}(S)}{n_i}\right)^{2k-1}}$$

the noise weights given in equation (5.1) are valid noise weights, according to Lemma 5.13. By Lemma 5.14, an interpolating tensor of the form given in equation (5.2) is a valid interpolating tensor. The tensor obtained by the discretization of the result of derivative matching has such a form (as $\min_{m\in[s]}\min_{i\in[d]}\min\{d^-_{i,m}, d^+_{i,m}\} \to \infty$).

According to equation (5.3) in Section 5.6.1 one has to choose

$$C \geq k^{\frac{2k-1}{2}}/a_0 \quad \text{as} \min_{i\in[d]}\min_{m\in[s]}\min\{d^-_{i,m}, d^+_{i,m}\} \to \infty.$$

The values of a_0 are given in Sections 5.6.3–5.6.6.

Theorem 2.2 in [19], upon which Theorem A.1 is based, uses a bound on the increments of empirical process $\{\varepsilon' f, f \in \mathbb{R}^n\}$, where ε has i.i.d. entries. Theorem A.1 involves in the background an empirical process, whose increments are given by

$$\left\{\sum_{1,\dots,1}^{n_1,\dots,n_d}(\varepsilon_{\mathcal{N}_k^\perp}\odot f)_{j_1,\dots,j_d}, f \in \mathbb{R}^{n_1\times\cdots\times n_d}\right\}.$$

Note that the entries of $\varepsilon_{\mathcal{N}_k^\perp} = \mathrm{P}_{\mathcal{N}_k^\perp}\varepsilon$ are correlated. However, by the idempotence of orthogonal projections, we can work with uncorrelated errors and instead restrict to tensors $f_{\mathcal{N}_k^\perp} \in \mathcal{N}_k^\perp$. Indeed

$$\sum_{1,\dots,1}^{n_1,\dots,n_d}(\varepsilon_{\mathcal{N}_k^\perp}\odot f)_{j_1,\dots,j_d} = \sum_{1,\dots,1}^{n_1,\dots,n_d}(\varepsilon\odot f_{\mathcal{N}_k^\perp})_{j_1,\dots,j_d}.$$

This allows us to take over the arguments of Theorem 2.2 in [19]. ∎

Remark 5.16 (The influence of the dimensionality). If we choose $\lambda \asymp \tilde{\gamma}\lambda_0(t)$, the rate of the oracle inequality is

$$\tilde{\gamma}^2 \sum_{m=1}^{s} \sum_{z \in \{-,+\}^d} (n/d_m^z)^{2k-1}/n,$$

up to logarithmic factors. For simplicity, let S define a regular grid. Then the (hyper-) volume of one of the s hyperrectangles of the tessellation scales as $d_m^z \asymp n/s$. Hence, the scaling

$$\sum_{m=1}^{s} \sum_{z \in \{-,+\}^d} (n/d_m^z)^{2k-1} \asymp s^{2k}.$$

However, $\tilde{\gamma}$, the maximal length of an antiprojection, scales as $\tilde{\gamma} \asymp (s^{-1/d})^{(2k-1)/2}$, where $s^{-1/d} \asymp d_{i,\max}/n_i$ is proportional to the side length of a hyperrectangle of the tessellation. The influence of the dimensionality in the exponent of s is a consequence of the different scaling of volume and side length of a hyperrectangle in d-dimensions. The (hyper-)volume scales as s^{-1}, while the side length scales as $s^{-1/d}$. The reason for this discrepancy is that we are not able to find an upper bound for the noise weights proportional the volume of the hyperrectangles, i.e., to the product of side lengths. The bound we obtain involves rather the sum of side lengths.

6. Not-so-slow ℓ^1-rates

Theorem 3.2 about not-so-slow rates for trend filtering is based on Theorem A.2, where the choice of the active set S is arbitrary. The criterion guiding choice of S is to get an “as small as possible” value of the inverse scaling factor $\tilde{\gamma}$. Recall that the inverse scaling factor $\tilde{\gamma}$ is the maximal length of the antiprojection of a dictionary atom $\tilde{\phi}^k_{j_1,\dots,j_d}$ onto the set of dictionary atoms indexed by the active set S, that is

$$\tilde{\gamma} \geq \max_{(j_1,\dots,j_d) \in [n_1]\times\cdots\times[n_d]} \|\mathrm{A}_S \tilde{\phi}^k_{j_1,\dots,j_d}\|_2/\sqrt{n}.$$

The active set S could be chosen as a regular grid parallel to the coordinate axes. However, we will show that we can shorten the maximal length of the antiprojections by choosing an active set defining a so-called “mesh grid”, whose construction we illustrate hereafter.

6.1. Mesh grids

Let $\delta \in \mathbb{N}$. For a coordinate $i \in [d]$, we define the set of indices $Z_i(l)$ such that

$$Z_i(l) = \{\delta^{d/l} \text{ equispaced indices in } [n_i]\}, l \in [d]$$

and

$$Z_i(1) \supseteq Z_i(2) \supseteq \ldots \supseteq Z_i(d).$$

If, for any $l \in [d]$, n_i is not a multiple of $\delta^{\frac{d}{l}}$, we relax the requirement on the indices to be approximately equispaced, i.e., the distance between all the indices has to be asymptotically of the same order. For $i \in [d]$, we also define

$$\widetilde{Z}_i(l) = \bigcup_{h=0}^{k-1} \{Z_i(l) + h\}, \quad l \in [d].$$

Let now $(l_1, \ldots, l_d) \in [d]^d$ be a tuple of indices. We define the set

$$\mathcal{S} := \{(l_1, \ldots, l_d) \in [d]^d : |\{i \in [d] : l_i \leq z\}| \leq z, \ \forall z \in [d]\}.$$

Definition 6.1 (Mesh grid). A mesh grid S is defined as

$$S := \bigcup_{(l_1,\ldots,l_d)\in\mathcal{S}} \left(\times_{i\in[d]} Z_i(l_i)\right).$$

Figure 2a illustrates a mesh grid for $d = 2$.

We now want to enlarge a mesh grid S to allow us to handle kth-order trend filtering for $k > 1$.

Definition 6.2 (Enlarged mesh grid). An enlarged mesh grid $\widetilde{S}$ is defined as

$$\widetilde{S} := \bigcup_{(l_1,\ldots,l_d)\in\mathcal{S}} \left(\times_{i\in[d]} \widetilde{Z}_i(l_i)\right).$$

Figure 2b illustrates an enlarged mesh grid for $d = 2$ and $k = 2$.

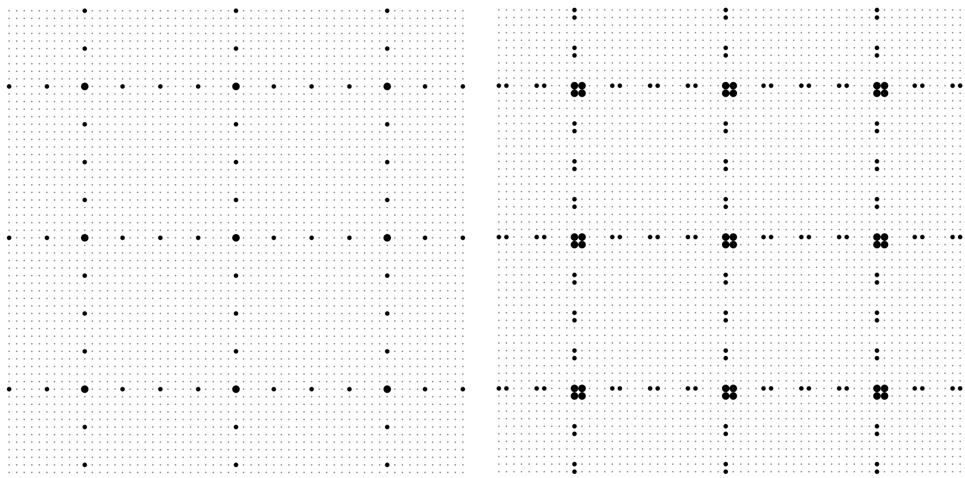
Let $s := |S|$ and $\widetilde{s} := |\widetilde{S}|$. It holds that

$$s \asymp \widetilde{s} \asymp \prod_{i=1}^{d} \delta^{\frac{d}{i}} \asymp \delta^{dH(d)},$$

where $H(d) = \sum_{i=1}^{d} 1/i$ is the dth harmonic number. Therefore, $\delta \asymp s^{\frac{1}{dH(d)}}$.

6.2. The inverse scaling factor when $\widetilde{S}$ is an enlarged mesh grid

We will now show that we can find a smaller bound on the inverse scaling factor if we choose $\widetilde{S}$ to be an enlarged mesh grid rather than an enlarged regular grid.



(a) **(b)**

Figure 2. **(a)** Mesh grid for $d = 2$; **(b)** Enlarged mesh grid for $d = 2$ and $k = 2$.

Lemma 6.3 (Inverse scaling factor when $\widetilde{S}$ is an enlarged mesh grid). *Let $n_1 \asymp \cdots \asymp n_d$ and $\widetilde{S}$ be an enlarged mesh grid. It holds that*

$$\widetilde{\gamma}(\widetilde{S}) = \mathcal{O}\big(s^{-\frac{2k-1}{2H(d)}}\big).$$

Proof. See Appendix D.1. ∎

6.3. Proof of Theorem 3.2

Theorem 3.2 follows from Theorem A.2. Theorem A.2 is allowed to have correlated errors for the same reasons as Theorem A.1 is, see the proof of Theorem 5.2 in Section 5.7.

In Theorem A.2 we set $x \asymp t \asymp \log n$. We can then choose the free parameters S and $g \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ independently of each other. If we allow λ to depend on g, we choose S to trade off the terms $\widetilde{\gamma}\lambda_0(\log n)\|D^k g\|_1 \asymp \widetilde{\gamma}\log^{1/2}(n)\|D^k g\|_1/n^{1/2}$ and s/n. Typically, we require S to have a regular structure and we obtain $\widetilde{\gamma} = \mathcal{O}(s^{-h})$, for some $h = h(d,k) \in \mathbb{R}$. The trade off is achieved with

$$s \asymp n^{\frac{1}{2(1+h)}} \log^{\frac{1}{2(1+h)}}(n)\|D^k f\|_1^{\frac{1}{1+h}}$$

and gives the rate

$$n^{-1+\frac{1}{2(1+h)}} \log^{\frac{1}{2(1+h)}}(n)\|D^k g\|_1^{\frac{1}{1+h}}.$$

Otherwise, we choose S to trade off $\tilde{\gamma}\lambda_0(\log n) \asymp \tilde{\gamma}\log^{1/2}(n)/n^{1/2}$ and s/n. If $\tilde{\gamma} = \mathcal{O}(s^{-h})$, the trade off is achieved with

$$s \asymp n^{\frac{1}{2(1+h)}} \log^{\frac{1}{2(1+h)}}(n)$$

and gives the rate

$$n^{-1+\frac{1}{2(1+h)}} \log^{\frac{1}{2(1+h)}}(n)\big(1 + \|D^k g\|_1\big).$$

We choose the active set to be an enlarged mesh grid $\tilde{S}$. Then, by Lemma 6.3, we can choose $\tilde{\gamma} = \mathcal{O}(s^{-(2k-1)/2H(d)})$ and the claim follows. ∎

Remark 6.4 (Mesh grids vs. regular grids). If we choose a regular grid as active set, according to Lemma 5.13 we obtain $\tilde{\gamma} \asymp s^{-(2k-1)/2d}$ and a slow rate

$$n^{-\frac{d+2k-1}{2d+2k-1}} \log^{\frac{d}{2d+2k-1}}(n),$$

which is slower than the rate obtained with an active set defining a mesh grid. Indeed, for all $d \geq 1$ it holds that $H(d) \leq d$.

In both cases, the slow rate for fixed k goes to $n^{-1/2}\log^{1/2}(n)$ as $d \to \infty$. If d is fixed, the slow rate goes to n^{-1} as $k \to \infty$.

7. Denoising lower-dimensional margins

In the previous sections we have shown how to estimate $f^0_{\mathcal{N}_k^\perp}$ by trend filtering and have established fast adaptive ℓ^0-rates and not-so-slow ℓ^1-rates. There is still an open question: how to estimate $f^0_{\mathcal{N}_k}$?

If $n_1 \asymp \cdots \asymp n_d$, the dimension of $\mathcal{N}_k$ is of order $n^{1-1/d}$. Estimating $f^0_{\mathcal{N}_k}$ by least squares would result in a rate of order $n^{-1/d}$ and therefore be limiting for $d \geq 2$.

The approach we take is to decompose $\mathcal{N}_k$ into lower dimensional mutually orthogonal linear spaces, the so-called marginal linear spaces, to which we can apply a lower dimensional version of trend filtering.

Let $\mathcal{P}[d]$ denote the power set of $[d] := \{1, \dots, d\}$. We consider sets of coordinate indices $M \subseteq [d]$.

The intuition behind the decomposition into margins is to partition the set of tensor indices into 2^d subsets as

$$\big([1:k] \cup [k+1:n_1]\big) \times \cdots \times \big([1:k] \cup [k+1:n_d]\big).$$

For $M \in \mathcal{P}[d]$ define the set of indices

$$I^k_M = \times_{i \in M}[k+1:n_i] \times_{i \notin M}[1:k].$$

We moreover define the linear spaces

$$\mathcal{M}(M) = \operatorname{span}\left\{\widetilde{\phi}^k_{k_1,\ldots,k_d}, (k_1,\ldots,k_d) \in I^k_M\right\}, \quad M \in \mathcal{P}[d].$$

Note that in one dimension, $\{\widetilde{\phi}^k_j\}_{j\in[k]}$ and $\{\widetilde{\phi}^k_j\}_{j\in[k+1:n]}$ are orthogonal to each other. Moreover, $M \triangle M' \neq \emptyset$, for $M \neq M' \in \mathcal{P}[d]$. Because of the product structure of the dictionary atoms spanning $\mathcal{M}$ this means that any $\mathcal{M}(M)$ and $\mathcal{M}(M')$ are mutually orthogonal, for $M \neq M'$.

The mutually orthogonal marginal linear subspaces $\{\mathcal{M}(M)\}_{M\in\mathcal{P}[d]}$ partition $\mathbb{R}^{n_1\times\cdots\times n_d}$. The dimension of $\mathcal{M}(M)$ is given by

$$k^{d-|M|} \prod_{i\in M} (n_i - k).$$

By the multi-binomial theorem it holds that

$$\prod_{i=1}^d n_i = \sum_{M\in\mathcal{P}[d]} k^{d-|M|} \prod_{i\in M} (n_i - k)$$

for $k \in [0 : \min_{l\in[d]} n_l - 1]$. This means that

$$\sum_{M\in\mathcal{P}[d]} \dim(\mathcal{M}(M)) = n$$

and because $\{\mathcal{M}(M)\}_{M\in\mathcal{P}[d]}$ are mutually orthogonal it follows that they also partition $\mathbb{R}^{n_1\times\cdots\times n_d}$.

We can further partition any $\mathcal{M}(M)$ into $k^{d-|M|}$ mutually orthogonal subspaces $\mathcal{M}(M,h)$, $h \in [1:k]^{d-|M|}$.

The partition results by defining the set of indices

$$I^k_{M,h} := \left(\times_{i\in M}[k+1:n_i]\right) \times \left(\times_{i\notin M}\{h_i\}\right)$$

and the linear subspaces

$$\mathcal{M}(M,h) := \operatorname{span}\left\{\widetilde{\phi}^k_{k_1,\ldots,k_d}, (k_1,\ldots,k_d) \in I^k_{M,h}\right\}.$$

Again, $\{\mathcal{M}(M,h)\}_{h\in[k]^{d-|M|},M\in\mathcal{P}[d]}$ are mutually orthogonal and partition $\mathbb{R}^{n_1\times\cdots\times n_d}$.

Definition 7.1 (ANOVA decomposition). The decomposition of a tensor f as

$$f = \sum_{M\in\mathcal{P}[d]} \sum_{h\in[1:k]^{d-|M|}} f_{\mathcal{M}(M,h)}$$

is called ANOVA decomposition.

By orthogonality we have that

$$\|f\|_2^2 = \sum_{M \in \mathcal{P}[d]} \sum_{h \in [1:k]^{d-|M|}} \|f_{\mathcal{M}(M,h)}\|_2^2.$$

7.1. Margins as lower dimensional objects

Our aim is to apply a lower dimensional version of trend filtering to estimate $f^0_{\mathcal{M}(M,h)}$, for $M \neq \emptyset$. For $M = \emptyset$ it holds that $|I^k_{M=\emptyset}| = k^d = \mathcal{O}(1)$. We will therefore estimate $f^0_{\mathcal{M}(\emptyset,h)}$ by the least squares estimate $Y_{\mathcal{M}(\emptyset,h)}$ at the parametric rate n^{-1}.

To apply a lower dimensional version of trend filtering to estimate $f^0_{\mathcal{M}(M,h)}$ we first need to reinterpret $f_{\mathcal{M}(M,h)}$ as a $|M|$-dimensional tensor. We then need to justify why we can apply Theorems A.1 and A.2 which require i.i.d. errors and are at the core of the adaptive rates by Theorem 5.2 and the not-so-slow rates by Theorem 3.2.

By writing

$$f_{\mathcal{M}(M,h)} = \bar{f}_{\mathcal{M}(M,h)} \times \left(\times_{i \notin M} \tilde{\phi}^k_{h_i}\right), \quad \bar{f}_{\mathcal{M}(M,h)} \in \mathbb{R}^{\times_{i \in M} n_i},$$

we can interpret $f_{\mathcal{M}(M,h)}$ as a $|M|$-dimensional object.

Similarly, we can write

$$Y_{\mathcal{M}(M,h)} = \bar{Y}_{\mathcal{M}(M,h)} \times \left(\times_{i \notin M} \tilde{\phi}^k_{h_i}\right), \quad \bar{Y}_{\mathcal{M}(M,h)} \in \mathbb{R}^{\times_{i \in M} n_i}.$$

Let $n_M := \prod_{i \in M} n_i$. Because of the (partial) product structure of $f_{\mathcal{M}(M,h)}$ and since $\|\tilde{\phi}^k_{h_i}\|_2^2 = n_i$, $h_i \in [k]$ (cf. Definition 4.1), it holds that

$$\|f_{\mathcal{M}(M,h)}\|_2^2/n = \|\bar{f}_{\mathcal{M}(M,h)}\|_2^2/n_M.$$

Thanks to the above equation and to the ANOVA decomposition we can add up the rates of estimation of the margins to estimate the whole tensor.

7.2. The estimator for the lower-dimensional margins

For $M \in \mathcal{P}[d] \setminus \emptyset$ define

$$D^k_M := n_M^{k-1} \prod_{i \in M} D^k_i.$$

To estimate the whole tensor, we consider the estimator

$$\hat{f} = \sum_{M \in \mathcal{P}[d]} \sum_{h \in [k]^{d-|M|}} \hat{f}_{\mathcal{M}(M,h)},$$

where

$$\hat{f}_{\mathcal{M}(M,h)} = \hat{\bar{f}}_{\mathcal{M}(M,h)} \times \left(\times_{i \notin M} \tilde{\phi}^k_{h_i}\right), \quad \hat{\bar{f}}_{\mathcal{M}(M,h)} \in \mathbb{R}^{\times_{i \in M} n_i}.$$

We define

$$\widehat{f}_{\mathcal{M}(\emptyset,h)} := \overline{Y}_{\emptyset,h} \times \left(\times_{i\in[d]}\widetilde{\phi}^k_{h_i}\right), \quad \forall h \in [k]^d$$

and

$$\widehat{\widehat{f}}_{\mathcal{M}(M,h)} := \underset{\overline{\mathrm{f}}_{\mathcal{M}(M,h)}\in\mathbb{R}^{\times_{i\in M} n_i}}{\arg\min} \left\{\|\overline{Y}_{\mathcal{M}(M,h)} - \overline{\mathrm{f}}_{\mathcal{M}(M,h)}\|_2^2/n_M + 2\lambda_{M,h}\|D^k_M \overline{\mathrm{f}}_{\mathcal{M}(M,h)}\|_1\right\},$$

where $\{\lambda_{M,h} > 0, h \in [k]^{d-|M|}, M \in \mathcal{P}[d] \setminus \emptyset\}$ are positive tuning parameters.

We call $\|D^k_M \overline{f}_{\mathcal{M}(M,h)}\|_1$ the kth-order $|M|$-dimensional Vitali total variation and $\widehat{\widehat{f}}_{\mathcal{M}(M,h)}$ the $|M|$-dimensional trend filtering estimator.

Remark 7.2 (We can apply Theorems A.1 and A.2). For $\overline{f} \in \mathbb{R}^{\times_{i\in M} n_i}$ it holds that

$$\begin{aligned}\overline{\varepsilon}_{\mathcal{M}(M,h)} \odot \overline{f} &= \overline{\varepsilon}_{\mathcal{M}(M,h)} \odot \overline{f}_{\mathcal{M}(M,h)} \\ &= \left(\sum_{M'\subseteq M, h'_M = h} \left(\overline{\varepsilon}_{\mathcal{M}(M',h')} \times \left(\times_{i\in M\setminus M'}\widetilde{\phi}^k_{h_i}\right)\right)\right) \odot \overline{f}_{\mathcal{M}(M,h)}.\end{aligned}$$

The n_M entries of the tensor

$$\sum_{M'\subseteq M, h'_M = h} \left(\overline{\varepsilon}_{\mathcal{M}(M',h')} \times \left(\times_{i\in M\setminus M'}\widetilde{\phi}^k_{h_i}\right)\right)$$

are the coefficients of the projection of ε onto the linear space spanned by $(\times_{i\notin M}\widetilde{\phi}^k_{h_i})\times(\times_{i\in M}\mathbb{R}^{n_i})$ and as such have i.i.d. $\mathcal{N}(0, \sigma^2 n_M/n)$-distributed entries. We can therefore apply Theorems A.1 and A.2 with noise variance $\sigma^2 n_M/n$.

Remark 7.3 (Synthesis form for the estimator of lower dimensional margins). The synthesis form of the estimator for the margins can be obtained in a similar way as for the d-dimensional margin (cf. Section 4).

8. Denoising the whole tensor

We now put together the results from Sections 5 and 6 with the ANOVA decomposition given in Section 7 to show rates for the estimation of the whole tensor.

In practice, when estimating different margins of the same tensor, one can tune the estimator for some margins to achieve slow rates and tune the estimator for other margins to achieve fast rates.

Consider an arbitrary partition of the set of margins

$$\{\mathcal{M}(M,h)\}_{h\in[1:k]^{d-|M|}, M\in\mathcal{P}[d]\setminus\emptyset}$$

into two sets $\mathcal{L}^0$ and $\mathcal{L}^1$. Then, by tuning the estimators for the margins in $\mathcal{L}^0$ to achieve the fast rates and the estimators for the margins in $\mathcal{L}^1$ to achieve the slow rates, we implicitly target tensors belonging to the class

$$\begin{aligned}\big\{f \in \mathbb{R}^{n_1,\dots,n_d} : & \|D^k_M \bar{f}_{\mathcal{M}(M,h)}\|_0 \le s_{M,h} \text{ for } \mathcal{M}(M,h) \in \mathcal{L}^0,\\ & \|D^k_M \bar{f}_{\mathcal{M}(M,h)}\|_1 \le C_{M,h} \text{ for } \mathcal{M}(M,h) \in \mathcal{L}^1\big\},\end{aligned}$$

where, for $\mathcal{M}(M,h) \in \mathcal{L}^0$, $s_{M,h}$ are integers and, for $\mathcal{M}(M,h) \in \mathcal{L}^1$, $C_{M,h} > 0$ are constants.

Fix now $k \in \{1,2,3,4\}$ and some $\mathcal{L}^0$ and $\mathcal{L}^1$. We restrict to tensors $n_1 \asymp \cdots \asymp n_d \asymp n^{1/d}$.

For $\mathcal{M}(M,h) \in \mathcal{L}^0$, we denote by $S_{M,h}$ a subset of $I^k_{M,h}$ satisfying the conditions for a hyperrectangular tessellation suitable for derivative matching. By $d^z_m(S_{M,h})$ we denote an analogue of the quantity d^z_m appearing in Theorem 5.2, but defined on a hyperrectangular tessellation of $I^k_{M,h}$ generated by the enlarged version $\widetilde{S}_{M,h}$ of the active set $S_{M,h}$.

For $\mathcal{M}(M,h) \in \mathcal{L}^0$, Let $C_{M,h} > 0$ be constants of order $\mathcal{O}(1)$.

Theorem 8.1 (Estimating a whole tensor by Vitali trend filtering). *Let $g \in \mathbb{R}^{n_1\times\cdots\times n_d}$ be arbitrary. For $\mathcal{M}(M,h) \in \mathcal{L}^0$ choose*

$$\lambda_{M,h} = \Omega\left(\sqrt{\sum_{i\in M}\left(\frac{d_{i,\max}(S_{M,h})}{n_i}\right)^{2k-1}\lambda_0(\log n)}\right).$$

Dependence of $\lambda_{M,h}$, $\mathcal{M}(M,h) \in \mathcal{L}^1$ on g allowed. *For $\mathcal{M}(M,h) \in \mathcal{L}^1$ choose*

$$\lambda_{M,h} \asymp n^{-\frac{H(|M|)+2k-1}{2H(|M|)+2k-1}} \log^{\frac{H(|M|)}{2H(|M|)+2k-1}}(n) \|D^k_M \bar{f}_{\mathcal{M}(M,h)}\|_1^{-\frac{2k-1}{2H(|M|)+2k-1}}.$$

Then with probability at least $1-\Theta(1/n)$ it holds that

$$\begin{aligned}\|\hat{f} - f^0\|_2^2/n \le \|g - f^0\|_2^2/n &+ \mathcal{O}\left(\sum_{\mathcal{M}(M,h)\in\mathcal{L}^0} \lambda_{M,h}\|(D^k_M \bar{g}_{\mathcal{M}(M,h)})_{-S_{M,h}}\|_1\right)\\ &+ \mathcal{O}\left(\sum_{\mathcal{M}(M,h)\in\mathcal{L}^0} \lambda^2_{M,h}\left(\sum_{i\in M}\log(ed_{i,\max}(S_{M,h}))\right)\sum_{m=1}^{s_{M,h}}\sum_{z\in\{-,+\}^{|M|}}\left(\frac{n_M}{d^z_m(S_{M,h})}\right)^{2k-1}\right)\\ &+ \mathcal{O}\left(\sum_{\mathcal{M}(M,h)\in\mathcal{L}^1} n^{-\frac{H(|M|)+2k-1}{2H(|M|)+2k-1}} \log^{\frac{H(|M|)}{2H(|M|)+2k-1}}(n)\|D^k_M \bar{g}_{\mathcal{M}(M,h)}\|_1^{\frac{2H(|M|)}{2H(|M|)+2k-1}}\right).\end{aligned}$$

Dependence of $\lambda_{M,h}$, $\mathcal{M}(M,h) \in \mathcal{L}^1$ on g not allowed. *For $\mathcal{M}(M,h) \in \mathcal{L}^1$ choose*

$$\lambda_{M,h} \asymp n^{-\frac{H(|M|)+2k-1}{2H(|M|)+2k-1}} \log^{\frac{H(|M|)}{2H(|M|)+2k-1}}(n).$$

Then with probability at least $1 - \Theta(1/n)$ *it holds that*

$$\|\hat{f} - f^0\|_2^2/n \le \|g - f^0\|_2^2/n + \mathcal{O}\Bigg(\sum_{\mathcal{M}(M,h) \in \mathcal{L}^0} \lambda_{M,h} \|(D_M^k \bar{g}_{\mathcal{M}(M,h)})_{-S_{M,h}}\|_1 \Bigg)$$

$$+ \mathcal{O}\Bigg(\sum_{\mathcal{M}(M,h) \in \mathcal{L}^0} \lambda_{M,h}^2 \Bigg(\sum_{i \in M} \log(e d_{i,\max}(S_{M,h})) \Bigg) \sum_{m=1}^{s_{M,h}} \sum_{z \in \{-,+\}^{|M|}} \left(\frac{n_M}{d_m^z(S_{M,h})} \right)^{2k-1} \Bigg)$$

$$+ \mathcal{O}\Bigg(\sum_{\mathcal{M}(M,h) \in \mathcal{L}^1} n^{-\frac{H(|M|)+2k-1}{2H(|M|)+2k-1}} \log^{\frac{H(|M|)}{2H(|M|)+2k-1}}(n) \big(1 + \|D_M^k \bar{g}_{\mathcal{M}(M,h)}\|_1\big) \Bigg).$$

Proof of Theorem 8.1. The result follows by the ANOVA decomposition. In total there are $(k+1)^d$ margins. As a consequence of the union bound, the result for the estimation of the whole tensor is attained with probability at least $1 - e^{-t} - e^{-x}$ if in the application of Theorems 5.2 and A.2 one chooses $x + d\log(k+1)$ and $t + d\log(k+1)$ instead of x and t for some $x, t > 0$. We choose $x \asymp t \asymp \log n$.

The bounds for the margins belonging to $\mathcal{L}^0$ follow directly from Theorem 5.2.

The bounds for the margins belonging to $\mathcal{L}^1$ follow from the application of Theorem A.2 to $\hat{\hat{f}}_{\mathcal{M}(M,h)}$ with $x + d\log(k+1)$ and $t + d\log(k+1)$. Let $\tilde{S}_{M,h}$ be an enlarged mesh grid. We have to trade off with respect to $\tilde{s}_{M,h} \asymp s_{M,h}$ the terms

$$\frac{n_M \sigma^2}{n} \frac{s_{M,h}}{n_M} \asymp \underbrace{\frac{1}{s^{\frac{2k-1}{H(|M|)}}}}_{\asymp \tilde{\gamma}} \underbrace{\sigma \sqrt{\frac{n_M}{n}} \sqrt{\frac{\log n}{n_M}}}_{\asymp \lambda_0(\log n)} \|D_M^k \bar{g}_{\mathcal{M}(M,h)}\|_1$$

or

$$\frac{n_M \sigma^2}{n} \frac{s_{M,h}}{n_M} \asymp \underbrace{\frac{1}{s^{\frac{2k-1}{H(|M|)}}}}_{\asymp \tilde{\gamma}} \underbrace{\sigma \sqrt{\frac{n_M}{n}} \sqrt{\frac{\log n}{n_M}}}_{\asymp \lambda_0(\log n)}.$$

We therefore obtain the rates

$$\mathcal{O}\Big(n^{-\frac{H(|M|)+2k-1}{2H(|M|)+2k-1}} \log^{\frac{H(|M|)}{2H(|M|)+2k-1}}(n) \|D_M^k \bar{g}_{\mathcal{M}(M,h)}\|_1^{\frac{2H(|M|)}{2H(|M|)+2k-1}} \Big)$$

or

$$\mathcal{O}\Big(n^{-\frac{H(|M|)+2k-1}{2H(|M|)+2k-1}} \log^{\frac{H(|M|)}{2H(|M|)+2k-1}}(n) (1 + \|D_M^k \bar{g}_{\mathcal{M}(M,h)}\|_1) \Big). \qquad \blacksquare$$

Remark 8.2 (The choice of k). If $\mathcal{L}^0 = \emptyset$ then one can choose $k \ge 1$. As soon as $\mathcal{L}^0 \ne \emptyset$, then one has to restrict to $k \in \{1, 2, 3, 4\}$. On the opposite side, when $\mathcal{L}^0 = \emptyset$, then one can drop the restriction $n_1 \asymp \cdots \asymp n_d \asymp n^{1/d}$.

The $\mathcal{O}(2^d)$ estimators we propose to estimate the margins are orthogonal to each other. Summing up these orthogonal estimators leads to a single optimization problem with $\mathcal{O}(2^d)$ tuning parameters. Can we reduce the number of the tuning parameters by choosing the same tuning parameter for estimating multiple margins? The answer depends on whether we consider slow or fast rates.

In general, the optimal choice of the tuning parameter depends on the inverse scaling factor $\widetilde{\gamma}$. The bounds we provide for $\widetilde{\gamma}$ depend on the active set $S_{M,h}$.

To obtain not-so-slow rates we choose $S_{M,h}$ to be a mesh grid and we obtain an estimate of the inverse scaling factor $\widetilde{\gamma}$ that depends on the dimension $|M|$ of the margin considered. Therefore we could penalize all the margins of the same dimension with the same tuning parameter and choose $\mathcal{O}(d)$ tuning parameters instead of $\mathcal{O}(2^d)$. Furthermore, if it were possible to refine the estimate of the inverse scaling factor $\widetilde{\gamma}$ such that it does not depend on the dimension $|M|$ anymore, we would only need to choose one tuning parameter.

To obtain fast rates, we allow for more general active sets $S_{M,h}$, such that the dependence of $\widetilde{\gamma}$ on $S_{M,h}$ is more intricate. The optimal tuning might be different for every one of the $\mathcal{O}(2^d)$ orthogonal estimators. If we aim for optimal tuning, then we need to choose $\mathcal{O}(2^d)$ tuning parameters. However, if one is fine with the suboptimal tuning which does not depend on $S_{M,h}$, one could use only one tuning parameter.

9. Conclusion

We have shown that imposing structure to denoise d-dimensional tensors leads to an adaptive reconstruction. The structure is imposed via penalties on the l-dimensional kth-order Vitali TV of the l-dimensional margins of the tensor, for $l \in [d]$. If the tensor is a product of polynomials on a constant number of hyperrectangles of any dimension $l \leq d$, then the MSE is bounded as

$$\|\widehat{f} - f^0\|_2^2/n = \mathcal{O}(\log^2 n/n),$$

with high probability. The true tensor f^0 can therefore be reconstructed at an almost parametric rate. The key aspects of our results are: the reformulation of the analysis estimator in synthesis form, the interpolating tensor to bound the effective sparsity and the ANOVA decomposition of a d-dimensional tensor. In the background of all our results there are the projection arguments by Dalalyan et al. [4] to bound the random part of the problem, which are fundamental to prove the adaptivity of $\widehat{f}$ to the underlying unobserved f^0.

Note that we only prove fast rates for Vitali trend filtering of order $k \in \{1, 2, 3, 4\}$. We are not able to prove that the approach we use to find an interpolating tensor for

$k \in \{1,2,3,4\}$ gives a suitable interpolating tensor for general k. Thus, although for each given finite k we can check by computer whether our construction gives a valid interpolating tensor, the problem remains open for general k.

Possible extensions of the Vitali trend filtering would be on one side to penalize total differences with different orders of differentiation along different coordinates – some sort of anisotropic version of the isotropic Vitali variation considered here – and on the other side to penalize the Vitali variation of different orders at the same time.

A. Oracle inequalities with fast and slow rates

In this section we report an oracle inequality with fast rates and one with slow rates. These oracle inequalities correspond to the adaptive and to the non-adaptive bound of Theorem 2.2 in [19], see also Theorems 2.1 and 2.2 in [18] and Theorems 16 and 17 in [17] adapted to have an enlarged active set.

Theorem A.1 (Oracle inequality with fast rates). *Let $g \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $S \subseteq \times_{i\in[d]} [k+2 : n_i - k]$ be arbitrary. For $x, t > 0$, choose $\lambda \geq \tilde{\gamma}\lambda_0(t)$. Then, with probability at least $1 - e^{-x} - e^{-t}$, it holds that*

$$\|(\hat{f} - f^0)_{\mathcal{N}_k^\perp}\|_2^2/n \leq \|g - f^0{}_{\mathcal{N}_k^\perp}\|_2^2/n + 4\lambda\|(D^k g)_{-S}\|_1 + \left(\sigma\sqrt{\frac{2x}{n}} + \sigma\sqrt{\frac{ks}{n}} + \lambda\Gamma_{D^k}(S, v_{-S}, q_S)\right)^2,$$

where $q_S = sign((D^k g)_S)$.

Theorem A.2 (Oracle inequality with slow rates). *Let $g \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $S \subseteq \times_{i\in[d]} [k+2 : n_i - k]$ be arbitrary. For $x, t > 0$, choose $\lambda \geq \tilde{\gamma}\lambda_0(t)$. Then, with probability at least $1 - e^{-x} - e^{-t}$, it holds that*

$$\|(\hat{f} - f^0)_{\mathcal{N}_k^\perp}\|_2^2/n \leq \|g - f^0_{\mathcal{N}_k^\perp}\|_2^2/n + 4\lambda\|D^k g\|_1 + \left(\sigma\sqrt{\frac{2x}{n}} + \sigma\sqrt{\frac{\tilde{s}}{n}}\right)^2.$$

B. Proofs of Section 4

B.1. Proof of Lemma 4.2

We prove Lemma 4.2 by induction.

Anchor: $k = 1$. Note that $\phi_1^1 = \tilde{\phi}_1^1$ and $\phi_j^1 - \tilde{\phi}_j^1 = \alpha\phi_1^1$ for some $\alpha \in \mathbb{R}$. Therefore, $D^1\phi_1^1 = D^1\tilde{\phi}_1^1 = 0$ and $D^1(\phi_j^1 - \tilde{\phi}_j^1) = 0$. It follows that

$$D^1\tilde{\phi}_j^1 = D^1\phi_j^1 = 1_{\{j' \geq j\}} - 1_{\{j' \geq j-1\}} = 1_{\{j\}}, \; j \in [2:n].$$

Step: $k-1$ implies k. For $j \in [k-1]$ it holds that

$$D^k \phi_j^k = D^k \tilde{\phi}_j^k = D^k \phi_j^{k-1} = D^k \tilde{\phi}_j^{k-1} = 0,$$

since by assumption $D^{k-1}\phi_j^{k-1} = D^{k-1}\tilde{\phi}_j^{k-1} = 0$ for $j \in [k-1]$. Moreover,

$$\begin{aligned} D^k \phi_j^k &= D^k \Big(\sum_{l \geq j} \phi_l^{k-1} \Big) / n = D^1 \Big(\sum_{l \geq j} D^{k-1} \phi_l^{k-1} \Big) \\ &= D^1 \{ 1_{\{j' \geq j\}} \}_{j' \in [k:n]} = \begin{cases} 0, & j = k, \\ 1_{\{j\}}, & j \in [k+1:n]. \end{cases} \end{aligned}$$

It also holds that

$$\phi_j^k - \tilde{\phi}_j^k = \sum_{l \in [k]} \alpha_l \phi_l^l, \quad j \in [k:n]$$

for some $\{\alpha_l \in \mathbb{R}\}_{l \in [k]}$, and therefore $D^k \phi_j^k = D^k \tilde{\phi}_j^k$, $j \in [k:n]$. ■

C. Proofs of Section 5

C.1. Proof of Lemma 5.12

To bound the antiprojections we can use the dictionary Φ^k instead of $\tilde{\Phi}^k$. Indeed, by Lemma 28 in [17], it holds that

$$\| \mathrm{A}_{\{\tilde{\phi}_t^k, t \in \tilde{S}\}} \tilde{\phi}_j^k \|_2^2 \leq \| \mathrm{A}_{\{\phi_t^k, t \in \tilde{S}\}} \phi_j^k \|_2^2, \quad j \in [k+1:n].$$

Bound on the antiprojections for $d = 1$. We first prove that, for $m = \in [s]$,

$$\| \mathrm{A}_{\tilde{S}} \tilde{\phi}_j^k \|_2^2 / n \leq \begin{cases} \left(\dfrac{t_m - j}{n} \right)^{2k-1}, & j \in R_m^- = [t_m^- : t_m], \\ 0, & j \in R_m^0 = [t_m : t_m + k - 1], \\ \left(\dfrac{j - t_m - k + 1}{n} \right)^{2k-1}, & j \in R_m^+ = [t_m + k - 1 : t_m^+]. \end{cases}$$

We then extend the reasoning to general dimension d.

For any $m \in [s]$, we fix $j \in R_m^-$ and approximate ϕ_j^k by $\phi_{t_m}^k, \dots, \phi_{t_m+k-1}^k$. By the definition of Φ^k we have that

$$\phi_j^k(j') = n^{-k+1}(j' - j + 1)^{k-1} 1_{\{j' \geq j\}}, \ j' \in [n].$$

Moreover, note that for $k' \in \{0, 1, \dots, k-1\}$

$$\sum_{l=0}^{k'}(-1)^l \binom{k'}{l} \phi^k_{t_m+l} = n^{-k'}\phi^{k-k'}_{t_m}$$
$$= n^{-k+1}\big\{(j' - t_m + 1)^{k-k'-1} 1_{\{j' \ge t_m\}}\big\}_{j' \in [n]}. \qquad \text{(C.1)}$$

We now decompose ϕ^k_j into a linear combination of $\phi^k_{t_m}, \dots, \phi^k_{t_m+k-1}$ and a remainder. The linear combination will approximate the projection of ϕ^k_j onto $\{\phi^k_j, j \in \widetilde{S}\}$, while the remainder will be an upper bound for the antiprojections.

For all $j' \in [n]$ it holds that

$$\phi^k_j(j') = n^{-k+1}(j' - j + 1)^{k-1}(1_{\{j \le j' \le t_m - 1\}} + 1_{\{j' \ge t_m\}}).$$

By the binomial theorem

$$\begin{aligned}
&(j' - t_m + 1 + t_m - j)^{k-1} 1_{\{j' \ge t_m\}} \\
&\quad = \sum_{l=0}^{k-1} \binom{k-1}{l} (t_m - j)^{k-l-1}(j' - t_m + 1)^l 1_{\{j' \ge t_m\}} \\
&\quad = \sum_{l=0}^{k-1} \binom{k-1}{l} (t_m - j)^{k-l-1} n^l \phi^{l+1}_{t_m}.
\end{aligned}$$

By equation (C.1) we know that $\{\phi^{l+1}_{t_m}\}_{l \in [0:k-1]} \in \operatorname{span}\big(\{\phi^k_{t_m+l}\}_{l \in [0:k-1]}\big)$.

Therefore, for $j \in R^-_m$,

$$\begin{aligned}
\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_j\|_2^2 &\le n^{-2k+2} \sum_{j'=j}^{t_m - 1} (j' - j + 1)^{2k-2} \\
&\le n^{-2k+2} \int_0^{t_m - j} (j')^{2k-2}\, dj' \le \frac{(t_m - j)^{2k-1}}{(2k-1)n^{2k-2}} \le n\left(\frac{t_m - j}{n}\right)^{2k-1}.
\end{aligned}$$

Note that the construction of the partially orthonormalized dictionary $\widetilde{\Phi}^k$ can of course also be made starting from the collection of functions $\{1_{\{j \le j'\}}\}_{j \in [n]}, j' \in [n]$ instead of $\{1_{\{j \ge j'\}}\}_{j \in [n]}, j' \in [n]$, cf. Definition 4.1. The resulting dictionaries $\widetilde{\Phi}^k$ coincide, up to permutation of the column indices. As a consequence, the calculation we showed to approximate $\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_j\|_2^2$ for $j \in R^-_m$ can be carried out with the dictionary $\widetilde{\Phi}^k$ based on $\{1_{\{j \le j'\}}\}_{j \in [n]}, j' \in [n]$ to obtain the approximation

$$\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_j\|_2^2 \le n\left(\frac{j - t_m - k + 1}{n}\right)^{2k-1}, \quad j \in R^+_m.$$

This consideration also applies in higher-dimensional situations.

Bound on the antiprojections for general dimension d. By the same reasons as above, we consider without loss of generality $(k_1, \dots, k_d) \in R_m^{-,\dots,-}$. We decompose $\phi^k_{k_1,\dots,k_d}$ as follows

$$\phi^k_{k_1,\dots,k_d}(j_1, \dots, j_d) = n^{-k+1} \prod_{i=1}^d \big(a_i(j_i) + b_i(j_i)\big),$$

for $j_i \in [n_i]$, $i \in [d]$, where

$$\begin{aligned} a_i &= a_i(j_i) = (j_i - k_i + 1)^{k-1} 1_{\{k_i \le j_i \le t_{i,m}-1\}}, \\ b_i &= b_i(j_i) = (j_i - k_i + 1)^{k-1} 1_{\{j_i \ge t_{i,m}\}}, \\ c_i &= c_i(j_i) = (j_i - k_i + 1)^{k-1} 1_{\{j_i \ge k_i\}} \ge a_i + b_i. \end{aligned}$$

Note that a_i, b_i depend on $t_{i,m}$, while c_i does not. Moreover, for all $(l_1, \dots, l_d) \in [0, k-1]^d$ it holds that $\times_{i\in[d]}\{t_{i,m+l_i}\} \in \widetilde{S}$. Thus, we approximate

$$\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_{k_1,\dots,k_d}\|_2^2 \le n^{-2k+2} \sum_{1,\dots,1}^{n_1,\dots,n_d} \left(\prod_{i=1}^d (a_i + b_i) - \prod_{i=1}^d b_i \right)^2,$$

since by equation (C.1) the contributions of $\prod_{i=1}^d b_i$ are spanned by ϕ^k_S. Note that $\prod_{i=1}^d (a_i + b_i) - \prod_{i=1}^d b_i$ is nonzero on

$$\big(\times_{i\in[d]}[k_i : n_i]\big) \setminus \big(\times_{i\in[d]}[t_{i,m} : n_i]\big) \subseteq \cup_{i\in[d]}\big([k_i : t_{i,m} - 1] \times \big(\times_{l\neq i}[1 : n_l]\big)\big).$$

Moreover, on $[k_i : t_{i,m} - 1] \times (\times_{l\neq i}[1 : n_l])$, it holds that

$$\prod_{i=1}^d (a_i + b_i) - \prod_{i=1}^d b_i \le a_i \prod_{l\neq i} c_l.$$

Therefore,

$$\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_{k_1,\dots,k_d}\|_2^2 \le n^{-2k+2} \sum_{i=1}^d \sum_{1,\dots,1}^{n_1,\dots,n_d} \left(a_i^2(j_1) \prod_{l\neq i} c_l^2(j_l) \right).$$

As in the one-dimensional case,

$$n_i^{-2k+2} \sum_{j_i=1}^{n_i} a_i^2(j_i) \le n_i \left(\frac{t_i - k_i}{n_i} \right)^{2k-1} \quad \text{and} \quad n_i^{-2k+2} \sum_{j_i=1}^{n_i} c_i^2(j_i) \le n_i.$$

It follows that

$$\|\mathrm{A}_{\widetilde{S}} \widetilde{\phi}^k_{k_1,\dots,k_d}\|_2^2 \le n \sum_{i=1}^d \left(\frac{t_i - k_i}{n_i} \right)^{2k-1}.$$

Note that as soon as $j_i \in R^0_{i,m}$ for some coordinate $i \in [d]$, then $a_i(j_i) = 0$ and the ith coordinate does not contribute to the antiprojections. The bounds for all other hyperrectangles R^z_m, $z \in \{-, 0, +\}^d$ follow by analogous calculations. □

C.2. Proof of Lemma 5.13

For any $m \in [s]$ and for any $(j_1, \dots, j_d) \in R_m$ it holds that

$$\begin{aligned}\sqrt{\sum_{i=1}^d \widetilde{v}^2_{i,m}(j_i)} &\le \sum_{i=1}^d \widetilde{v}_{i,m}(j_i) \le \sum_{i=1}^d v_{i,m}(j_i)\Big(\frac{\max\{d^-_{i,m}, d^+_{i,m}\}}{n_i}\Big)^{\frac{2k-1}{2}} \\ &\le \sum_{i=1}^d v_{i,m}(j_i)\sqrt{\sum_{l=1}^d \Big(\frac{\max\{d^-_{l,m}, d^+_{l,m}\}}{n_l}\Big)^{2k-1}} \le v_{j_1,\dots,j_d}\widetilde{\gamma}.\end{aligned}$$

■

C.3. Proof of Lemma 5.14

Fix $i \in [d]$ and $m \in [s]$. Say $q_{t_m} = 1$. Since $w_{i,l,m} \in [0,1]$, $l \neq i$, for any $j_i \in R^-_{i,m} \cup R^0_{i,m} \cup R^+_{i,m}$ it holds that

$$\prod_{l=1}^d w_{i,l,m}(j_l) \le \Big(1 - \frac{v_{i,m}(j_i)}{C}\Big)\prod_{l\neq i} w_{i,l,m}(j_l) \le \Big(1 - \frac{v_{i,m}(j_i)}{C}\Big).$$

Moreover, for any $(j_1, \dots, j_d) \in R_m$ it holds that

$$\begin{aligned}w_{j_1,\dots,j_d} &= \frac{1}{d}\sum_{i=1}^d \prod_{l=1}^d w_{i,l,m}(j_l) \\ &\le \frac{1}{d}\sum_{i=1}^d \Big(1 - \frac{v_{i,m}(j_i)}{C}\Big) = 1 - \sum_{i=1}^d \frac{v_{i,m}(j_i)}{dC} = 1 - v_{j_1,\dots,j_d}.\end{aligned}$$

Analogous expressions hold if $q_{t_m} = -1$. The claim follows by noting that the conditions of the definition of interpolating tensor (Definition 5.10) are satisfied for w. ■

C.4. Matching derivatives

To obtain continuous vectors with $k-1$ continuous derivatives and piecewise constant kth derivative, we split $[0,1]$ into N_ω, resp. N_{w}, intervals of equal length, where $N_{\mathrm{w}} = k+1$ if k is odd and $N_{\mathrm{w}} = k+2$ if k is even and $N_\omega = k$. We denote these intervals by $\{[x_{l-1}, x_l]\}_{l=1}^{N_{\{\omega,\mathrm{w}\}}}$ with $x_0 = 0$ and $x_{N_{\{\omega,\mathrm{w}\}}} = 1$. We choose

$$\omega(x) = \begin{cases} 1 - a_0 x^{\frac{2k-1}{2}}, & x \in [x_0, x_1], \\ b_{l,k}x^k + b_{l,k-1}x^{k-1} + \ldots + b_{l,1}x + b_{l,0}, & x \in [x_{l-1}, x_l], \\ & l \in [2 : k-1], \\ c_0(1-x)^k, & x \in [x_{k-1}, x_k]. \end{cases}$$

We moreover choose

$$\mathrm{w}(x) = \begin{cases} 1 - \mathrm{a}_0 x^k, & x \in [x_0, x_1], \\ \mathrm{b}_{l,k}x^k + \mathrm{b}_{l,k-1}x^{k-1} + \ldots + \mathrm{b}_{l,1}x + \mathrm{b}_{l,0}, & x \in [x_{l-1}, x_l], \\ & l \in [2 : N_{\mathrm{w}}/2 - 1], \\ \mathrm{a}_L(1/2 - x)^L + \ldots + \mathrm{a}_1(1/2 - x) + 1/2, & x \in [x_{N_{\mathrm{w}}/2-1}, x_{N_{\mathrm{w}}/2+1}], \end{cases}$$

where $L = k - 1$ if k is even and $L = k$ if k is odd.

We choose both the coefficients

$$(a_0, a_L, \ldots, a_1, \{b_{l,k}, \ldots, b_{l,0}\}_l, c_0) \quad \text{and} \quad (\mathrm{a}_0, \mathrm{a}_L, \ldots, \mathrm{a}_1, \{\mathrm{b}_{l,k}, \ldots, \mathrm{b}_{l,0}\}_l)$$

by derivative matching. We require the $k-1$ derivatives of the different pieces of the interpolating polynomials to match at the junctions between the intervals. This gives place to piecewise constant kth derivatives with the exception of the interval $[x_0, x_1]$, where $\omega^{(k)}(x) \asymp -1/\sqrt{x}$.

Matching derivatives for ω means solving a system of $k(k-1)$ equations and $k(k-1)$ unknowns. Matching derivatives for w means solving a system of $k(k/2)$ equations and $k(k/2)$ unknowns when k is even, and $k(k-1)/2$ equations and $k(k-1)/2$ unknowns when k is odd. We therefore do not need to do any derivative matching for $k = 1$, where we just take $\omega(x) = 1 - \sqrt{x}$ and $\mathrm{w}(x) = 1 - x$.

As an alternative to discretizing a continuous version of the interpolating polynomials, one can also proceed by matching discrete differences. The two approaches are equivalent when $\min_{i\in[d]} \min_{m\in[s]} \min\{d_{i,m}^-, d_{i,m}^+\} \to \infty$ as $n \to \infty$. Discrete derivative matching requires that the counterpart of each interval $[x_{l-1} : x_l]$ contains at least k points. We therefore require that

$$\min\{d_{i,m}^-, d_{i,m}^+\} \geq (k+2)k, \quad \forall i \in [d],\ \forall m \in [s].$$

We refer to [19] for details on discrete derivative matching.

C.5. Partial integration

Some consequences of the fact that both the resulting ω and w have piecewise constant kth derivatives with the exception of the interval $[0, x_1]$ where $\omega^{(k)}(x) \asymp -1/\sqrt{x}$ are shown in the next lemma, which is useful to compute the bound on the effective sparsity in Lemma 5.15.

Lemma C.1 (Discrete differences of some polynomials). *Let for some $d \in \mathbb{N}$, $d \geq 2k$,*

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

Then

$$n^{-2k+2} \| D^k \mathrm{q} \|_2^2 = \mathcal{O}\big(\log(ed)/d^{2k-1}\big).$$

Let for some $d \in \mathbb{N}$, $d \geq 2k$,

$$\mathrm{p}_j := (j/d)^k, \quad j = 0, \ldots, d.$$

Then

$$n^{-2k+2} \| D^k \mathrm{p} \|_2^2 = \mathcal{O}\big(1/d^{2k-1}\big).$$

Proof. We have for $j \geq k$

$$\begin{aligned} n^{-2k+2} (D^k \mathrm{q})_j &= \sum_{l=0}^{k} \binom{k}{l} (-1)^l \left(\frac{j-l}{d} \right)^{\frac{2k-1}{2}} \\ &= \left(\frac{j}{d} \right)^{\frac{2k-1}{2}} \left[\sum_{l=0}^{k} \binom{k}{l} (-1)^l \left(1 - \frac{l}{j} \right)^{\frac{2k-1}{2}} \right]. \end{aligned}$$

We do a $(k-1)$-term Taylor expansion of $x \mapsto (1-x)^{\frac{2k-1}{2}}$ around $x = 0$:

$$(1-x)^{\frac{2k-1}{2}} = \sum_{i=0}^{k-1} a_i x^i + \mathrm{rem}(x),$$

where $a_0 = 1, a_1 = -\frac{2k-1}{2}, \ldots, a_{k-1}$ are the coefficients of the Taylor expansion and where the remainder $\mathrm{rem}(x)$ satisfies

$$\sup_{0 \leq x \leq 1/2} |\mathrm{rem}(x)| = \mathcal{O}\big(|x|^k\big).$$

Thus,

$$\sum_{l=0}^{k} \binom{k}{l} (-1)^l \left(1 - \frac{l}{j} \right)^{\frac{2k-1}{2}} = \sum_{l=0}^{k} \binom{k}{l} (-1)^l \left(\sum_{i=0}^{k-1} a_i \left(\frac{l}{j} \right)^i + \mathrm{rem}\left(\frac{l}{j} \right) \right),$$

where

$$\sum_{l=0}^{k} \binom{k}{l} (-1)^l \sum_{i=0}^{k-1} a_i \left(\frac{l}{j} \right)^i = 0$$

since

$$\Big\{\sum_{i=0}^{k-1} a_i \Big(\frac{l}{j}\Big)^i\Big\}_{l=0}^{k}$$

is a polynomial of degree $k-1$ and hence its kth-order differences are zero. It follows that for $j \geq k$,

$$\Big|\sum_{l=0}^{k}\binom{k}{l}(-1)^l\Big(1-\frac{l}{j}\Big)^{\frac{2k-1}{2}}\Big| \leq \sum_{l=0}^{k}\binom{k}{l}\Big|\mathrm{rem}\Big(\frac{l}{j}\Big)\Big| = \mathcal{O}\Big(\frac{1}{j^k}\Big).$$

Then for $j \geq k$,

$$n^{-2k+2}(D^k\mathrm{q})_j = \mathcal{O}\Big(1/(j^{\frac{1}{2}}d^{\frac{2k-1}{2}})\Big).$$

So,

$$n^{-2k+2}\|D^k\mathrm{q}\|_2^2 = \mathcal{O} \leq \big(\log(ed)/d^{2k-1}\big).$$

For p the same arguments follow. We obtain that $(D^k\mathrm{p})_j = \mathcal{O}(1/d^k)$, and so

$$n^{-2k+2}\|D^k\mathrm{p}\|_2^2 = \mathcal{O}\big(1/d^{2k-1}\big). \qquad \blacksquare$$

C.6. Proof of Lemma 5.15

We prove a bound on the effective sparsity holding for every sign configuration. We eliminate the dependence on the sign configuration by decoupling partial integration on the whole interpolating tensor ($\|(D^k)'w\|_2^2$) into taking kth-order differences on the hyperrectangles $\{R_m\}_{m=1}^s$ ($\|D^k w(R_m)\|_2^2$, where $w(R_m)=\{w_{j_1,\dots,j_d}\}_{(j_1,\dots,j_d)\in R_m}$ denotes the restriction of the interpolating tensor w to the set of indices R_m).

To do this, we define the boundaries $B(R_m)$ of a rectangle R_m as

$$B(R_m) := R_m \setminus \times_{i\in[d]}[t_{i,m}^- + k : t_{i,m}^+ - k].$$

It holds that

$$n^{-2k+2}\|(D^k)'w\|_2^2 = \mathcal{O}\Bigg(\sum_{m=1}^{s}\big(n^{-2k+2}\|D^k w(R_m)\|_2^2 + \|w(B(R_m))\|_2^2\big)\Bigg).$$

By the definition of the interpolating tensor w it holds that

$$\begin{aligned}
n^{-2k+2}\|D^k w(R_m)\|_2^2 &= \mathcal{O}\Bigg(n^{-2k+2}\sum_{i=1}^{d}\|D^k \times_{l\in[d]} w_{l,i,m}\|_2^2\Bigg)\\
&= \mathcal{O}\Bigg(n^{-2k+2}\sum_{i=1}^{d}\prod_{l=1}^{d}\|D^k w_{l,i,m}\|_2^2\Bigg)
\end{aligned}$$

$$= \mathcal{O}\Bigg(\sum_{i=1}^{d} \prod_{l=1}^{d} \Bigg(n_l^{-2k+2} \| D^k w^-_{l,i,m} \|_2^2 + \sum_{j_l = t_{l,m}-k}^{t_{l,m}-1} \big(1 - w^-_{l,i,m}(j_l)\big)^2 \\ + n_l^{-2k+2} \| D^k w^+_{l,i,m} \|_2^2 + \sum_{j_l = t_{l,m}+k}^{t_{l,m}+2k-1} \big(1 - w^+_{l,i,m}(j_l)\big)^2 \Bigg)\Bigg),$$

where the sums stem from the differences involving the constant part of w on $R^0_{l,m}$. Because of the form chosen for ω and w, it holds that

$$\sum_{j_l = t_{l,m}-k}^{t_{l,m}-1} \big(1 - w^-_{l,i,m}(j_l)\big)^2 = \begin{cases} \mathcal{O}\big(\omega^2(1/d^-_{i,m})\big) \\ \mathcal{O}\big(\mathrm{w}^2(1/d^-_{l,m})\big) \end{cases} \\ = \begin{cases} \mathcal{O}\big(1/(d^-_{i,m})^{2k-1}\big), & l = i, \\ \mathcal{O}\big(1/(d^-_{l,m})^{2k}\big), & l \neq i. \end{cases}$$

A similar bound holds for $\sum_{j_l = t_{l,m}+k}^{t_{l,m}+2k-1} (1 - w^+_{l,i,m}(j_l))^2$. By Lemma C.1 it holds that

$$n_l^{-2k+2} \| D^k w^-_{l,i,m} \|_2^2 = \begin{cases} \mathcal{O}\big(\log(e d^-_{i,m})/(d^-_{i,m})^{2k-1}\big), & l = i, \\ \mathcal{O}\big(1/(d^-_{l,m})^{2k-1}\big), & l \neq i. \end{cases}$$

A similar bound holds for $n_l^{-2k+2} \| D^k w^+_{l,i,m} \|_2^2$.

We now just have to upper bound the contributions of the boundaries $B(R_m)$. For $k = 1$, $w(B(R_m)) = 0$, for all $m \in [s]$ and the boundaries do not contribute to the effective sparsity. For $k \geq 2$ it holds that

$$\sum_{B(R_m)} w^2_{j_1,\dots,j_d} = \mathcal{O}\Bigg(\sum_{i=1}^{d} \sum_{B(R_m)} \prod_{l=1}^{d} w^2_{l,i,m}(j_l) \Bigg) = \mathcal{O}\Bigg(\sum_{i=1}^{d} \sum_{z \in \{-,+\}^d} \frac{1}{(d^z_m)^{2k-1}} \Bigg)$$

since all the contributions on the boundaries have the same dependence on k and we can approximate the volume of the boundaries by the sum of the volume of the 2^d fractions $\{R^z_m\}_{z \in \{-,+\}^d}$ of the hyperrectangle.

It therefore holds that

$$n^{-2k+2} \| (D^k)' w \|_2^2 = \mathcal{O}\Bigg(\Bigg(\sum_{i=1}^{d} \log(e d_{i,\max}(S)) \Bigg) \sum_{m=1}^{s} \sum_{z \in \{-,+\}^d} \frac{1}{(d^z_m)^{2k-1}} \Bigg)$$

and the claim follows. ∎

D. Proofs of Section 6

D.1. Proof of Lemma 6.3

Setting. To calculate the inverse scaling factor when the active set is an enlarged mesh grid $\widetilde{S}$, we decompose a dictionary atom – which is a product of sums – into a sum of products. Some of the components will be spanned by the dictionary atoms indexed by the mesh grid. The remaining components will contribute to the antiprojection.

By [17, Lemma 28] we can look at the dictionary atoms $\phi^k_{j_1,\dots,j_d}$ instead of $\widetilde{\phi}^k_{j_1,\dots,j_d}$; see also the proof of Lemma 5.12 in Appendix C.1.

We therefore consider

$$\phi^k_{j_1,\dots,j_d} = \phi^k_{j_1} \times \cdots \times \phi^k_{j_d},$$

where for $i \in [d]$,

$$\phi^k_{j_i} = n_i^{-k+1}(j - j_i + 1)^{k+1} 1_{\{j \geq j_i\}}.$$

Projection of the mesh grid on single coordinates. Now choose $z_{i,l} \in Z_i(l)$ such that $j_i \leq z_{i,1} \leq \cdots \leq z_{i,d-1} \leq z_{i,d}$. By the definition of the mesh grid we can choose $z_{i,l} \in Z_i(l)$ such that

$$\begin{aligned} |j_i - z_{i,1}| &= \mathcal{O}\left(n_i / s^{\frac{1}{H(d)}}\right), \\ |z_{i,l} - z_{i,l-1}| &= \mathcal{O}\left(n_i / s^{\frac{1}{lH(d)}}\right), \quad l \in [2:d], \\ |z_{i,d}| &\leq n_i. \end{aligned}$$

The decomposition. We now decompose the factors into sums:

$$\phi^k_{j_i} = \sum_{l=0}^{d} u_{i,l},$$

where, for $j \in [n_i]$,

$$\begin{aligned} u_{i,0} &:= 1_{\{j \in [j_i : z_{i,1}-1]\}} n_i^{-k+1}(j - j_i + 1)^{k-1}, \\ u_{i,l} &:= 1_{\{j \in [z_{i,l} : z_{i,l+1}-1]\}} n_i^{-k+1}(j - j_i + 1)^{k-1}, \quad l \in [1:d-1], \\ u_{i,d} &:= 1_{\{j \in [z_{i,d} : n_i]\}} n_i^{-k+1}(j - j_i + 1)^{k-1}. \end{aligned}$$

Note that $\{u_{i,l}\}_{l=0}^d$ are mutually orthogonal.

Thanks to the decomposition of the factors, the following decomposition of the dictionary atom $\phi^k_{j_1,\dots,j_d}$ holds:

$$\phi^k_{j_1,\dots,j_d} = \sum_{(l_1,\dots,l_d) \in [0:d]^d} \prod_{i=1}^{d} u_{i,l_i},$$

where $\{\prod_{i=1}^d u_{i,l_i}\}_{(l_1,\ldots,l_d)\in[0:d]^d}$ are mutually orthogonal. We therefore obtain a decomposition of a product of sums into a sum of products.

Partitioning the decomposition. We now partition $\{(l_1,\ldots,l_d) \in [0:d]^d\}$ into two subsets: Σ and Σ^c. Define

$$\Sigma := \big\{(l_1,\ldots,l_d) \in [0:d]^d : |\{i \in [d] : l_i \leq z\}| \leq z, \forall z \in [0:d]\big\}.$$

This means that Σ contains tuples $(l_1,\ldots,l_d)$ having at most d entries with value at most d *and* at most $d-1$ entries with value at most $d-1$ *and* ... *and* at most 1 entry with value at most 1 *and* no entry with value 0.

Connecting the decomposition with the enlarged mesh grid. We now want to show that, for any $(l_1,\ldots,l_d) \in \Sigma$, $\prod_{i=1}^d u_{i,l_i}$ can be obtained as a linear combination of $\{\phi^k_{j_1,\ldots,j_d}\}_{(j_1,\ldots,j_d)\in\widetilde{S}}$. These components will approximate the projection of any $\phi^k_{j_1,\ldots,j_d}$ onto the linear span of $\{\phi^k_{j_1,\ldots,j_d}\}_{(j_1,\ldots,j_d)\in\widetilde{S}}$.

For $l_i \in [1:d-1]$ it holds that

$$u_{i,l_i}(j) = 1_{\{z_{i,l_i}\leq j\}} n_i^{-k+1}(j-j_i+1)^{k-1} - 1_{\{z_{i,l_i+1}\leq j\}} n_i^{-k+1}(j-j_i+1)^{k-1}.$$

In analogy to the proof of Lemma 5.12 (use the binomial theorem and equation (C.1)) it holds that

$$u_{i,l_i} \in \operatorname{span}\big(\{\phi^k_{z_{i,l_i}+h}\}_{h=0}^{k-1} \cup \{\phi^k_{z_{i,l_i+1}+h}\}_{h=0}^{k-1}\big).$$

For $l_i \in [d]$ it holds that $u_{i,d} \in \operatorname{span}\big(\{\phi^k_{z_{i,l_i}+h}\}_{h=0}^{k-1}\big)$.

We need a claim. We now claim that

$$(l_1,\ldots,l_d) \in \Sigma \Longrightarrow (l_1',\ldots,l_d') \in \Sigma,$$

where $l_i' \geq l_i, \forall i \in [d]$ by proving that

$$(l_1,\ldots,l_d) \in \Sigma \Longrightarrow (l_1,\ldots,l_{d-1},l_d+1) \in \Sigma,$$

where without loss of generality we choose the index l_d and assume that $l_d \leq d-1$.

As a consequence it will follow that, for any $(l_1,\ldots,l_d) \in \Sigma$, $\prod_{i=1}^d u_{i,l_i}$ can be obtained as a linear combination of $\{\phi^k_{j_1,\ldots,j_d}\}_{(j_1,\ldots,j_d)\in\widetilde{S}}$.

We now prove the claim: assume that $(l_1,\ldots,l_d) \in \Sigma$, i.e.,

$$|\{i \in [d] : l_i \leq z\}| \leq z, \quad \forall z \in [0:d].$$

Take $(l_1',\ldots,l_d')$ as $l_i' = l_i, i \in [d-1]$ and $l_d' = l_d + 1$. Then

$$\begin{aligned}|\{i \in [d] : l_i' \leq z\}| &= |\{i \in [d-1] : l_i \leq z\}| + 1_{\{z\geq l_d+1\}} \\ &\leq z - 1_{\{z\geq l_d\}} + 1_{\{z\geq l_d+1\}} \leq z.\end{aligned}$$

Therefore, $(l_1',\ldots,l_d') \in \Sigma$ and the claim is proved.

Approximating the antiprojections. Thanks to the above claim and to the mutual orthogonality of the elements of $\{\prod_{i=1}^d u_{i,l_i}\}_{(l_1,\dots,l_d)\in[0:d]^d}$, we can approximate as follows:

$$\|\mathrm{A}_{\widetilde{S}}\phi^k_{j_1,\dots,j_d}\|_2^2/n \leq \sum_{(l_1,\dots,l_d)\notin\Sigma} \left\| \prod_{i=1}^d u_{i,l_i} \right\|_2^2/n = \sum_{(l_1,\dots,l_d)\notin\Sigma} \prod_{i=1}^d \|u_{i,l_i}\|_2^2/n_i.$$

Now we use the following property:

$$\|u_{i,l_i}\|_2^2/n_i = \mathcal{O}\big(s^{-\frac{2k-1}{(l_i+1)H(d)}}\big).$$

The larger l_i, the larger the contribution of $\|u_{i,l_i}\|_2^2/n_i$.

It therefore only remains to find the order of the largest contribution(s) indexed by Σ^c. A tuple of indices in Σ^c giving the contribution highest in order is

$$(d-1,\dots,d-1).$$

It holds that

$$\|\mathrm{A}_S\phi^k_{j_1,\dots,j_d}\|_2^2/n = \mathcal{O}\left(\prod_{i=1}^d \|u_{i,l_i}\|_2^2/n_i\right) = \mathcal{O}\big(s^{-\frac{2k-1}{H(d)}}\big).$$

Since the upper bound does not depend on $(j_1,\dots,j_d)$ we read directly that

$$\widetilde{\gamma} = \mathcal{O}\big(s^{-\frac{2k-1}{2H(d)}}\big).$$

■

E. The bound on the effective sparsity by Lemma 5.15 is tight (in the noiseless case, up to constants)

We show that the bound on the effective sparsity by Lemma 5.15 is tight in the noiseless case, up to constants, by providing lower bounds on the noiseless effective sparsity. The noiseless effective sparsity is the effective sparsity as defined in Definition 5.9, but with $v_{-S} = 0$.

It holds that

$$\Gamma^2_{D^k}(S,0) \geq \frac{\big(\|(Df)_S\|_1 - \|(Df)_{-S}\|_1\big)^2}{\|f\|_2^2/n}, \quad \forall f \in \mathbb{R}^{n_1\times\cdots\times n_d}.$$

To prove a lower bound on $\Gamma^2_{D^k}(S,0)$ we therefore need to choose an appropriate $f \in \mathbb{R}^{n_1\times\cdots\times n_d}$.

E.1. Lower bounds in one dimension

Let $S = \{t_1, \dots, t_s\}$. For $k \in \{1, 2, 3, 4\}$ assume a tessellation of $[k+1 : n]$ with $t_m^+ = t_{m+1}^-$, $m \in [s-1]$ and $d_m^+ = d_{m+1}^-$, $m \in [s-1]$, where $d_m^- = t_m - t_m^-$ and $d_m^+ = t_m^+ - t_m$ for $m \in [s]$. Note that $t_1^- = k+1$. Assume that

$$d_1^- + k = d_1^+ = \cdots = d_s^- = d_s^+ =: d \quad \text{and} \quad d > k.$$

E.1.1. A piecewise constant function, $k = 1$.

Define

$$K_1 := \frac{1}{2} \sum_{m=1}^{s} \left(\frac{n}{d_m^-} + \frac{n}{d_m^+} \right).$$

Lemma E.1. *For $k = 1$ and $d = 1$, the bound on the effective sparsity by Lemma 5.15 is tight in the noiseless case up to a constant, i.e., it holds that*

$$\Gamma_{D^1}^2(S, 0) \geq K_1.$$

Proof of Lemma E.1. Consider the vector $f^* \in \mathbb{R}^n$ given by

$$f_j^* = \begin{cases} -(1^m)\dfrac{n}{d_m^-}, & j \in [t_m - d : t_m],\ m \in [s] \\ (-1)^{m+1}\dfrac{n}{d_m^+}, & j \in [t_m + 1 : t_m + d],\ m \in [s]. \end{cases}$$

It holds that $\|(Df^*)_{-S}\|_1 = 0$ and $\|(Df^*)_S\|_1 = 2K_1$. Moreover,

$$\|f^*\|_2^2/n = (d_1^- + 1)\frac{n}{(d_1^-)^2} + \sum_{m=1}^{s-1} (d_m^+ + d_{m+1}^-)\frac{n}{(d_m^-)^2} + \frac{n}{(d_s^+)} \leq 4K_1.$$

Therefore, $\Gamma_{D^1}^2(S, 0) \geq K_1$. ∎

E.1.2. A piecewise linear function, $k = 2$.

For $k \geq 1$ we need to choose f^* with continuous $0, 1, \dots, k-2$ derivatives. In the case $k = 2$, f^* needs to be a discretization of a continuous function.

Define

$$K_2 := \frac{3}{2^9} \sum_{m=1}^{s} \left(\left(\frac{n}{d_m^-} \right)^3 + \left(\frac{n}{d_m^+} \right)^3 \right).$$

Lemma E.2. *For $k = 2$ and $d = 1$, the bound on the effective sparsity by Lemma 5.15 is tight in the noiseless case up to a constant, i.e., it holds that*

$$\Gamma_{D^2}^2(S, 0) \geq K_2.$$

Proof of Lemma E.2. Consider the vector $f^* \in \mathbb{R}^n$ given by

$$f_j^* = (-1)^m \frac{n^2}{d^3} |t_m - j|, j \in [t_m - d : t_m + d], \quad m \in [s].$$

It holds that $\|(Df^*)_{-S}\|_1 = 0$ and $\|(Df^*)_S\|_1 \geq 2^6 K_2/3$. Indeed, for $d > k$ it holds that $(d+k)^{-x} \geq 2^{-x} d^{-x}$, $x \in \mathbb{N}$. Moreover,

$$\|f^*\|_2^2/n \leq \frac{n^3}{d^6} 2s \frac{(d+1)^3}{3} \leq 2^{12} K_2/3^2.$$

Therefore, $\Gamma^2_{D^2}(S, 0) \geq K_2$. ∎

E.1.3. A piecewise constant function, $k = 3$. In the case $k = 3$, f^* needs to be a discretization of a continuous function with continuous first derivative.

Define

$$K_3 := \frac{5}{2^{13}} \sum_{m=1}^{s} \left(\left(\frac{n}{d_m^-} \right)^5 + \left(\frac{n}{d_m^+} \right)^5 \right).$$

Lemma E.3. *For $k = 3$ and $d = 1$, the bound on the effective sparsity by Lemma* 5.15 *is tight in the noiseless case up to a constant, i.e., it holds that*

$$\Gamma^2_{D^3}(S, 0) \geq K_3.$$

Proof of Lemma E.3. Consider the vector $f^* \in \mathbb{R}^n$ given by

$$f_j^* = \begin{cases} -\dfrac{n^3}{d^5}(j^2 - d^2), & j \in R_1^- \cup [1:3], \\ (-1)^{m+1} \dfrac{n^3}{d^5} \big((t_m^+ - j)^2 - d^2 \big), & j \in R_m^+, m \in [1:s], \\ (-1)^m \dfrac{n^3}{d^5} \big((t_m^- - j)^2 - d^2 \big), & j \in R_m^-, m \in [2:s]. \end{cases}$$

It holds that $\|(Df^*)_{-S}\|_1 = 0$ and $\|(Df^*)_S\|_1 \geq \frac{2^9}{5} K_3$. Moreover,

$$\|f^*\|_2^2/n \leq \frac{n^5}{d^{10}} 2s \frac{(d+1)^5}{5} \leq \frac{2^{18}}{5^2} K_3.$$

Therefore, $\Gamma^2_{D^3}(S, 0) \geq K_3$. ∎

E.1.4. A piecewise cubic function, $k = 4$. In the case $k = 4$, f^* needs to be a discretization of a continuous function with continuous first and second derivative.

Define

$$K_4 := \frac{3^2}{2^{12} 7^2} \sum_{m=1}^{s} \left(\left(\frac{n}{d_m^-} \right)^7 + \left(\frac{n}{d_m^+} \right)^7 \right).$$

Lemma E.4. *For $k = 4$ and $d = 1$, the bound on the effective sparsity by Lemma 5.15 is tight in the noiseless case up to a constant, i.e., it holds that*

$$\Gamma^2_{D^4}(S, 0) \geq K_4.$$

Proof of Lemma E.4. Consider the vector $f^* \in \mathbb{R}^n$ given by

$$f^*_j = \begin{cases} (-1)^m \frac{n^4}{d^7}(j - t_m - d/2)(j - t_m + d)^2, & j \in [t_m - d : t_m], m \in [s], \\ (-1)^m \frac{n^4}{d^7}(j - t_m + d/2)(j - t_m - d)^2, & j \in [t_m : t_m + d], m \in [s]. \end{cases}$$

It holds that $\|(Df^*)_{-S}\|_1 = 0$ and $\|(Df^*)_S\|_1 \geq \frac{2^6 7^2}{3} K_4$. Moreover,

$$\|f^*\|_2^2/n \leq \frac{2^{12}7^4}{3^2} K_4.$$

Therefore, $\Gamma^2_{D^4}(S, 0) \geq K_4$. ∎

E.2. Lower bounds in higher dimensions

We consider the case $d \geq 1$. For $k \in \{1, 2, 3, 4\}$ define for each coordinate $i \in [d]$ the active set $S_i := \{t_{1,i}, \dots, t_{s_1,i}\}$ and the bounds $K_{k,i}$ as in Lemmas E.1–E.4 and their proofs.

Let $S := S_1 \times \cdots \times S_d$. Assume a tessellation of $[k + 1 : n_1] \times \cdots \times [k + 1 : n_d]$ being a product of one-dimensional tessellations with $t^+_{m,i} = t^-_{m+1,i}$, $m \in [s_i - 1]$, $i \in [d]$ and $d^+_{m,i} = d^-_{m+1,i}$, $m \in [s_i - 1]$, $i \in [d]$, where $d^-_{m,i} = t_{m,i} - t^-_{m,i}$ and $d^+_{m,i} = t^+_{m,i} - t_{m,i}$ for $m \in [s_i]$, $i \in [d]$. Note that $t^-_{1,i} = k + 1$, $i \in [d]$.

Lemma E.5. *For $k \in \{1, 2, 3, 4\}$ and $d \geq 1$, the bound on the effective sparsity by Lemma 5.15 is tight in the noiseless case up to a constant, i.e., it holds that*

$$\Gamma^2_{D^k}(S, 0) \geq \prod_{i=1}^d K_{k,i}.$$

Proof of Lemma E.5. For $i \in [d]$, define the vectors $f^*_i \in \mathbb{R}^{n_i}$ as in the proofs of Lemmas E.1–E.4. Define $f^* := f^*_1 \times \cdots \times f^*_d$.

Because of the product structure of f^*, by the proofs of Lemmas E.1–E.4 it holds that

$$\|D^k f^*\|_1 = \prod_{i=1}^d \|D^k_i f^*_i\|_1 = \prod_{i=1}^d \|(D^k_i f^*_i)_{S_i}\|_1 = \|(D^k f^*)_S\|_1 \geq \prod_{i=1}^d K_{k,i}$$

and

$$\|f^*\|_2^2/n = \prod_{i=1}^{d} \|f_i^*\|_2^2/n_i \le \prod_{i=1}^{d} K_{k,i}. \qquad \blacksquare$$

Acknowledgments. We are very grateful to two anonymous reviewers for their valuable comments.

Funding. We would like to acknowledge support for this project from the Swiss National Science Foundation (SNF grant 200020_169011).

References

[1] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data. Methods, theory and applications*. Springer Ser. Statist., Springer, Heidelberg, 2011 Zbl 1273.62015 MR 2807761

[2] S. Chatterjee and S. Goswami, Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees. *Ann. Statist.* **49** (2021), no. 5, 2531–2551 Zbl 07438259 MR 4338374

[3] S. Chatterjee and S. Goswami, New risk bounds for 2D total variation denoising. *IEEE Trans. Inform. Theory* **67** (2021), no. 6, part 2, 4060–4091 Zbl 07374382 MR 4289366

[4] A. S. Dalalyan, M. Hebiri, and J. Lederer, On the prediction performance of the Lasso. *Bernoulli* **23** (2017), no. 1, 552–581 Zbl 1359.62295 MR 3556784

[5] M. Elad, P. Milanfar, and R. Rubinstein, Analysis versus synthesis in signal priors. *Inverse Problems* **23** (2007), no. 3, 947–968 Zbl 1138.93055 MR 2329926

[6] B. Fang, A. Guntuboyina, and B. Sen, Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. *Ann. Statist.* **49** (2021), no. 2, 769–792 Zbl 1471.62372 MR 4255107

[7] J. H. Friedman, Multivariate adaptive regression splines. *Ann. Statist.* **19** (1991), no. 1, 1–141 Zbl 0765.62064 MR 1091842

[8] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** (2007), no. 2, 302–332 Zbl 1378.90064 MR 2415737

[9] A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen, Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** (2020), no. 1, 205–229 Zbl 1439.62100 MR 4065159

[10] A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen, Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** (2020), no. 1, 205–229 Zbl 1439.62100 MR 4065159

[11] J.-C. Hütter and P. Rigollet, Optimal rates for total variation denoising. In *29th Annual Conference on Learning Theory*, pp. 1115–1146, Proceedings of Machine Learning Research 49, PMLR, 2016

[12] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, l_1 trend filtering. *SIAM Rev.* **51** (2009), no. 2, 339–360 Zbl 1171.37033 MR 2505584

[13] K. Lin, J. Sharpnack, A. Rinaldo, and R. J. Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Proceedings of the 31st International Conference on Neural Information Processing System*, pp. 6887–6896, Curran Associates Inc., 2017

[14] E. Mammen and S. van de Geer, Locally adaptive regression splines. *Ann. Statist.* **25** (1997), no. 1, 387–413 Zbl 0871.62040 MR 1429931

[15] F. Ortelli and S. van de Geer, Synthesis and analysis in total variation regularization. 2019, arXiv:1901.06418v1

[16] F. Ortelli and S. van de Geer, On the total variation regularized estimator over a class of tree graphs. *Electron. J. Stat.* **12** (2018), no. 2, 4517–4570 Zbl 1411.62208 MR 3892703

[17] F. Ortelli and S. van de Geer, Adaptive rates for total variation image denoising. *J. Mach. Learn. Res.* **21** (2020), Paper No. 247 Zbl 7306926 MR 4213428

[18] F. Ortelli and S. van de Geer, Oracle inequalities for square root analysis estimators with application to total variation penalties. *Inf. Inference* **10** (2021), no. 2, 483–514 Zbl 07409387 MR 4270756

[19] F. Ortelli and S. van de Geer, Prediction bounds for higher order total variation regularized least squares. *Ann. Statist.* **49** (2021), no. 5, 2755–2773 Zbl 07438267 MR 4338382

[20] V. Sadhanala, Nonparametric methods with total variation type regularization, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2019

[21] V. Sadhanala and R. J. Tibshirani, Additive models with trend filtering. *Ann. Statist.* **47** (2019), no. 6, 3032–3068 Zbl 1436.62450 MR 4025734

[22] V. Sadhanala, Y. X. Wang, J. Sharpnack, and R. Tibshirani, Higher-order total variation classes on grids: minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems 30*, pp. 5801–5811, 2017.

[23] V. Sadhanala, Y.-X. Wang, and R. Tibshirani, Total variation classes beyond 1d: minimax rates, and the limitations of linear smoothers. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3521–3529, Curran Associates Inc., 2016

[24] J. Sharpnack, A. Rinaldo, and A. Singh, Sparsistency of the edge lasso over graphs. In *International Conference on Artificial Intelligence and Statistics*, pp. 1028–1036, Proceedings of Machine Learning Research 22, PMLR, 2012.

[25] R. Tibshirani, Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. 2020, arXiv:2003.03886

[26] R. Tibshirani, Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996), no. 1, 267–288 Zbl 0850.62538 MR 1379242

[27] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** (2005), no. 1, 91–108 Zbl 1060.62049 MR 2136641

[28] R. J. Tibshirani, Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** (2014), no. 1, 285–323 Zbl 1307.62118 MR 3189487

[29] S. van de Geer, *Estimation and testing under sparsity*. Lecture Notes in Math. 2159, Springer, Cham, 2016 MR 3526202

[30] S. van de Geer, On tight bounds for the Lasso. *J. Mach. Learn. Res.* **19** (2018), Paper No. 46 Zbl 1467.62127 MR 3874154
[31] S. van de Geer and P. Hinz, The Lasso with structured design and entropy of (absolute) convex hulls. To appear in *Foundations of Modern Statistics*, Springer
[32] Y.-X. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani, Trend filtering on graphs. *J. Mach. Learn. Res.* **17** (2016), Paper No. 105 Zbl 1369.62082 MR 3543511

Received 26 January 2021; revised 9 July 2021.

Francesco Ortelli
Seminar für Statistik, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland;
fortelli@ethz.ch

Sara van de Geer
Seminar für Statistik, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland;
vsara@ethz.ch