Prediction bounds for (higher order) total variation regularized least squares
Sara van de Geer and Francesco Ortelli
Seminar for Statistics, ETH Zürich

Abstract We establish oracle inequalities for the least squares estimator \( \hat{f} \) with penalty on the total variation of \( f \) or on its higher order differences. Our main tool is an interpolating vector that leads to lower bounds for compatibility constants. This allows one to show that for any \( N \in \mathbb{N} \) the \( N \)th order differences penalty leads to an estimator \( \hat{f} \) that can adapt to the number of jumps in the \((N - 1)\)th order differences.

1 Introduction

Total variation (TV) penalties have been introduced by Rudin and Osher [1992] and Steidl et al. [2006]. The present paper builds further on the theory as developed in Tibshirani [2014], Sadhanala and Tibshirani [2017] and Guntuboyina et al. [2017]. We show that for any \( N \in \mathbb{N} \) the \( N \)th order TV regularized least squares estimator can adapt to the number of jumps in the \((N - 1)\)th order differences. Inspired by Candès and Fernandez-Granda [2014], our main tool is the use of an interpolating vector which interpolates between the signs of the jumps. We will moreover base our theory on an oracle inequality for the general “analysis” problem given in (1) below. This allows one to generalize the findings to graphs. We only briefly elaborate on this in the concluding section.

In Elad et al. [2007] it is shown that every analysis problem has an equivalent “synthesis” formulation. The synthesis problem is called the Lasso (Tibshirani [1996]). The paper Dalalyan et al. [2017] introduces a new “compatibility constant” for the synthesis problem and derives oracle inequalities. We establish oracle inequalities for the analysis problem without taking the detour via a synthesis problem. Moreover, we provide bounds on the compatibility constant using interpolating vectors. We furthermore generalize the projection arguments from Dalalyan et al. [2017] by allowing for “mock” variables. In this way we arrive at better weights in the compatibility constant which in turn lead to the desired oracle results.

Having observed a vector \( Y \in \mathbb{R}^n \) the analysis problem is

\[
\min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_2^2/n + 2\lambda\|Df\|_1 \right\}
\]  

(1)

where \( D \in \mathbb{R}^{m \times n} \) is a given “analysis operator” and \( \lambda > 0 \) is a tuning parameter. We denote the solution of (1) by \( \hat{f} \). The aim is to show that \( \hat{f} \) is close to the mean \( f^0 := \mathbb{E}Y \) of \( Y \), or to some approximation \( f \in \mathbb{R}^n \) thereof that has \( \|Df\|_0 \) “small”. Throughout we assume that the noise \( \epsilon := Y - f^0 \) is a vector of i.i.d. (unobservable) Gaussian random variables with known variance \( \sigma^2 \). Without
loss of generality we take $\sigma^2$ equal to 1. For the case of unknown variance one may apply for example the analysis version of the square-root Lasso introduced by Belloni et al. [2011]. The paper Ortelli and van de Geer [2019] derives oracle results for square-root analysis.

1.1 Organization of the paper

In the next section we explain the idea of interpolating vectors to arrive at bounds for the compatibility constant. In Section 3 we present the main result for the least squares estimator with (higher order) total variation penalty. We explain there “in words” how this result can be derived using interpolating vectors. Sections 6 and 7 study in detail the case $N = 1$ and $N = 2$ respectively. The theory for general $N$ is laid out in Section 8. The main point is the “matching” of higher order differences. This means one needs to solve a system of linear equations. The details for $N = 3$ are given in Subsection 8.4. For the case $N \geq 4$ we do not give explicit constants but only describe the system of equations. (When $N = 3$ the number of equations is also 3, but when $N = 4$ the number of equations is 8.) The results of Sections 3, 6 and 7 are based on the oracle inequality for the general analysis problem presented in Section 5. This inequality is (potentially) based on mock variables, which we describe in Section 4. Section 9 presents an inequality without compatibility assumptions nor entropy calculations, showing the minimax rate up to log-terms. Section 10 concludes. Since the way the results are obtained may be of interest in itself, the proofs are given in the main text. However, the general oracle inequality of Theorem 5.1 and the almost minimax result of Theorem 9.1 form an exception: their proof, given in Section 11, follows the arguments used in Ortelli and van de Geer [2019] but with as new element the introduction of mock variables.

1.2 Some notation

The row vectors of $D$ are indexed by a set $\mathcal{D}$ with of size $|\mathcal{D}| = m$. One may take $\mathcal{D} = \{1, \ldots, m\}$ but in our examples a different indexing is more convenient. For example, if $D$ is the incidence matrix of a graph one may index its columns by the edges of the graph, i.e. by pair of nodes sharing an edge. We write the row vectors of $D$ as $\{d_j\}_{j \in \mathcal{D}}$. Moreover, $D_{-S}$ is the sub-matrix of the remaining rows. We write $\mathcal{N} := \mathcal{N}(D), \mathcal{N}_S := \mathcal{N}(D_S)$ and $\mathcal{N}_{-S} := \mathcal{N}(D_{-S})$. The dimension of $\mathcal{N}_{-S}$ is denoted by $r_S$. We let $\bar{S} \supset S$
be such that $D_{-\bar{S}}$ has full row rank and let $\bar{s} := |\bar{S}|$ be its size. For two vectors $a$ and $b$ with the same dimension the vector $ab := \{a_j b_j\}$ denotes the entry-wise product.

## 2 Compatibility via interpolating vectors

**Definition 2.1** Let $w_{-S} := \{w_j\}_{j \in D \setminus \bar{S}} \in \{0, 1\}^{m-s}$ be a vector of weights. The weighted effective sparsity is

$$\Gamma^2(S, w_{-S}) = \left( \min \left\{ \|f\|_2^2/n : \|D_S f\|_1 - \|(1 - w_{-S}) D_{-S} f\|_1 = 1 \right\} \right)^{-1}.$$

Thus the weighted effective sparsity is up to scaling the inverse of the weighted “compatibility constant”

$$\kappa^2(S, w_{-S}) := r_S \min \left\{ \|f\|_2^2/n : \|D_S f\|_1 - \|(1 - w_{-S}) D_{-S} f\|_1 = 1 \right\}$$

which is the analysis version of the compatibility constant given in [Dalalyan et al. 2017]. The scaling by $r_S := \text{dim}(N_\bar{S})$ is in a sense natural: it has to do with the different scaling of the $\ell_1$-norm as compared to the $\ell_2$-norm. However, in the present context this scaling is not too helpful, as $r_S$, the dimension of the “oracle problem” where $S$ is known to be a good active set (which could be the active set of $f_0$ for instance) is potentially going to be replaced by a larger dimension due to the adding of “mock” variables as discussed in Section 4. We will express our results in terms of the effective sparsity $\Gamma^2(S, w_{-S})$ rather than in terms of the compatibility constant $\kappa^2(S, w_{-S})$.

Given weights $w_{-S} \in \{0, 1\}^{m-s}$ we can define for all $z_S \in \pm 1^s$ (sign-)interpolating vectors $q(z_S, w_{-S}) \in \mathbb{R}^m$ such that

$$(q(z_S, w_{-S}))_S = z_S, \quad |(q(Z_S, w_{-S}))_j| \leq 1 - w_j, \quad j \in D \setminus S.$$

We let $Q(z_S, w_{-S})$ the set of all such interpolating vectors $q(z_S, w_{-S})$. The following lemma says that given an interpolating vector for the worst case sign configuration, one immediately has an upper bound for the effective sparsity, i.e. a lower bound for the compatibility constant. We took the idea from [Candès and Fernandez-Granda 2014] which has a qualitative result concerning the so-called null space property using interpolating polynomials. Lemma 2.1 can be seen as a quantitative version of this idea and moreover concerns the (in this context easier) analysis problem instead of a synthesis problem.

**Lemma 2.1** It holds that

$$\Gamma^2(S, w_{-S}) \leq n \max_{z_S \in \{\pm 1\}^s} \min_{q \in Q(z_S, w_{-S})} \|D' q\|_2^2.$$
Proof of Lemma 2.1. Let \( f \in \mathbb{R}^n \) be an arbitrary vector with \( d'_jf_j \neq 0 \) for all \( j \in S \), and let \( z_S \) be the sign vector of \( D_S f \). Then we have for all \( q \in Q(z_S, w_{-S}) \)

\[
\|D_S f\|_1 - \|(1 - w_{-S})D_S f\|_1 \leq q'Df \leq \|D'q\|_2\|f\|_2.
\]

Lemma 2.1 is simple yet powerful. The fact that it can be invoked for synthesis/analysis problems is to be credited to Dalalyan et al. [2017] because they show (for the synthesis problem) that the weights for the active part can be taken equal to zero, i.e. they introduced a version of the weighted compatibility constant that is susceptible for study using interpolating vectors. However, as far as we know the present paper is first in pointing out that the new compatibility constant opens the door for interpolating vectors.

3 TV regularization of general higher order differences

In this section, we present the main result for the TV regularized least squares estimator as a special case of the general oracle inequality of Theorem 5.1. We will explain after its statement the line of reasoning we use.

Fix some \( N \in \mathbb{N} \). Define \( \Delta^0f = f \) and for \( j \geq 2 \)

\[
(\Delta f)_j := f_j - f_{j-1}.
\]

Consider \( \Delta f \) as vector in \( \mathbb{R}^{n-1} \) with index set \( \{2, \ldots, n\} \). For \( j \geq l + 1, 1 \leq l \leq N \), define

\[
(\Delta^l f)_j := (\Delta(\Delta^{l-1} f))_j
\]

where \( \Delta^{l-1} f \in \mathbb{R}^{n-(l-1)} \) has index set \( \{l, \ldots, n\} \). The \( N \)th order TV regularized least squares estimator is

\[
\hat{f} := \arg \min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_2^2/n + 2\lambda\|\Delta^N f\|_1 \right\}.
\]

This corresponds to the analysis problem with \( \mathcal{D} := \{N+1, \ldots, n\} \) and with analysis operator \( D \) the \( N \)th order difference operator

\[
(Df)_j := (\Delta^N f)_j, \quad j \in \mathcal{D}.
\]

Take some arbitrary \( f \in \mathbb{R}^n \) and some arbitrary \( S := \{t_1, \ldots, t_s\} \subset \mathcal{D} \), where \( t_1 < \cdots < t_s \). Write \( t_0 := N \) and \( t_{s+1} := n \). Let \( d_\infty := \max_{1 \leq k \leq s+1}(t_k - t_{k-1}) \).

Theorem 3.1 For all \( u > 0, v > 0 \), and for

\[
\lambda \geq c_N \lambda_S(u)n^{-\frac{1}{2}}d_\infty^{2N-1},
\]

Theorem 3.1
where
\[ \lambda_S(u) = \sqrt{\frac{2 \log(2(n - (s + N))) + 2u}{n}}, \]
we have with probability at least \(1 - \exp[-u] - \exp[-v]\)
\[ \|\hat{f} - f^0\|_2/n \leq \|f - f^0\|_2/n + \left( \sqrt{\frac{N(s + 1)}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma(S, w_S) \right)^2 + 4\lambda \|D_Sf\|_1 \]
with
\[ \Gamma^2(S, w_S) \leq C_N^2 \sum_{k=1}^{s+1} \frac{n \log(t_k - t_{k-1})}{(t_k - t_{k-1})^{2N+1}}. \]
Here \(c_N\) and \(C_N\) are constants depending only on \(N\) and \(w_S\) is a suitable vector of weights.

The main new point in this theorem is the bound given for the effective sparsity \(\Gamma^2(S, w_S)\). The vector \(w_S\) consists of weights that dominate up to a scaling the length of the residuals after projecting the non-active variables with indices in \(D\setminus S\) on the active variables with indices in \(S\) plus possibly some extra mock variables. If there were no noise (a situation only of theoretical interest) one could take all the weights equal to zero. To calculate \(\Gamma(S, 0)\) one argues as follows. One needs to interpolate the sign vector \(z_S\) with alternating signs because alternating signs forms the worst case (the most difficult interpolation problem). Note now that the linear map \(D'\), appearing in the bound of Lemma 2.1 for the effective sparsity, is the discrete variant of taking the \(N\)th derivative (up to a minus sign when \(N\) is odd). Therefore, roughly speaking (up to discretization) the interpolation problem boils down to finding an interpolation of the sign vector \(z_S\) with \(N\)th derivative piecewise constant. Because this interpolation has to build a bridge from the, say, the +1 sign at \(t_{k-1}\) to the −1 sign at \(t_k\), it is clear that an \(N\)th derivative of order \((t_k - t_{k-1})^{-N} \) on an interval \((t_{k-1}, t_k]\) is needed. The squared \(\ell_2\)-norm over this interval is then up to scaling of order \((t_k - t_{k-1})^{-2N+1}\).

However, in the noisy case the interpolating vector is to drop quicker at \(t_{k-1}\) and slower at \(t_k\), so it will no longer consist of piecewise polynomials of degree \(N\). This is why there appear extra log-factors in the effective sparsity \(\Gamma^2(S, w_S)\).

**Corollary 3.1** If one takes \(S := \{j : d^j f = 0\}\) (the active set of \(f \in \mathbb{R}^n\)) the term \(\|D_Sf\|_1\) vanishes in the oracle result of Theorem 3.1. The result then says that \(f\) is up to a “variance term” at least as close to \(f^0\) as the “sparse” vector \(f\). If one then chooses \(f = f^0\) the “bias” term vanishes as well and there is only the “variance” term left involving the active set \(S_0 := \{j : d^j f^0 \neq 0\}\) of \(f^0\) (with size \(s_0 := |S_0|\)).

**Corollary 3.2** We note that Theorem 3.1 gives good results when the distances between jumps are more or less all equal. Suppose now that they are indeed all equal:
\[ d_\infty = t_1 - t_0 = \cdots = t_{s+1} - t_s = \frac{n - N}{s+1} \in \mathbb{N} \]
Then we may take
\[ \lambda = c_N \lambda_S(u)n^{-\frac{1}{2}} \left( \frac{n - N}{s + 1} \right)^{\frac{2N-1}{2}}, \]
and then
\[ \Lambda(S,w) \leq c_NC_N \lambda_S(u) \sqrt{s + 1} \sqrt{\log((n - N)/(s + 1))}. \]
For \( N \) fixed and \( u \) of order \( \log n \) this is of order
\[ \sqrt{(s + 1) \log^2 n}. \]
With \( s = n^{-\frac{1}{2N+1}} \), one obtains up to log-terms the minimax rate \( n^{-\frac{N}{2N+1}} \) for estimating a vector \( f^0 \) with, after scaling with \( n^N \), \( \ell_\infty \)-bounded \( N^{th} \) differences (as one can approximate such a function \( f^0 \) on a bounded interval by a function \( f \) with \( D_{-S}f = 0 \), \( s = |S| = n^{-\frac{1}{2N+1}} \) and \( \|f - f^0\|_2/n \leq \text{const}.n^{-\frac{2N}{2N+1}} \)). One can show that this rate is still achieved when \( f^0 \) has, after scaling with \( n^{N-1} \), only \( \ell_1 \)-bounded \( N^{th} \) differences. We derive this result with an extra log-factor in Section 4. The reason for the log-factor is that we use similar projection arguments as in the proof of Theorem 5.1 instead of more refined entropy bounds.

4 Adding some mock variables

We will see that the anti-projection of the “non-active” variables with indices in \( D \setminus S \) on an appropriate space, which we shall call \( V^S \) will play an important role. If \( V^S \) is a rich space, these anti-projections will have small length, which is good. On the other hand we do not want to have \( V^S \) too rich because its dimension \( r(V^S) \) will occur in the upper bound for the prediction error.

The space \( V^S \) will be spanned by a basis for the null-space \( N_{-S} \) and possibly some additional “mock” variables \( U^S \in \mathbb{R}^{n \times r(U^S)} \). Without loss of generality we take the matrix \( U^S \) of full rank \( r(U^S) \) (not adding any mock variables is a special case, where we take \( r(U_S) = 0 \)). Let \( U^S \) be a linear subspace of \( \mathbb{R}^n \) spanned by the columns of the matrix \( U^S \) and define
\[ V^S := N_{-S} \oplus U^S. \]

The main point is now that one can write
\[ (I - \Pi_{V^S})f = A^SD_{-S}f \]
for a properly chosen matrix \( A^S \) (see Lemma 4.1 below). This is easy to see but important, because as is usual for \( \ell_1 \) penalized problems, we will need the dual norm inequality
\[ |a'_{-S}D_{-S}f| \leq \|a_{-S}\|_\infty \|D_{-S}f\|_1, \quad a_{-S} \in \mathbb{R}^{n-s}. \]
We assume
\[ r(VS) := \dim(VS) = r_S + r(U^S) \]
i.e. that there are no redundant mock variables. In fact we assume non-redundancy in the sense that
\[ \operatorname{rank}(\Pi_{\mathcal{N}^\perp} U^S) = r(U^S) \]
i.e. the anti-projections of the mock variables remain linearly independent.

When \( r(U^S) \neq 0 \) we define
\[ B^S := (D_{-\bar{S}} D_{-\bar{S}}')^{-1} D_{-\bar{S}} U^S, \]
and we let
\[
A^S := \begin{cases} 
  D_{-\bar{S}}' \left( (D_{-\bar{S}} D_{-\bar{S}}')^{-1} - B^S (B^{St} D_{-\bar{S}} D_{-\bar{S}}' B^S)^{-1} B^{St} \right) & r(U^S) \neq 0 \\
  D_{-\bar{S}}'(D_{-\bar{S}} D_{-\bar{S}}')^{-1} & r(U^S) = 0
\end{cases}.
\]

**Lemma 4.1** It holds that
\[ (I - \Pi_{VS}) f = A^S D_{-\bar{S}} f. \]

**Proof of Lemma 4.1.** By standard projection arguments
\[ \Pi_{VS} f = \Pi_{\mathcal{N}^\perp} f + \Pi_{U^S_{\mathcal{N}^\perp}} f \]
where \( U^S_{\mathcal{N}^\perp} \) is the space spanned by \( \Pi_{\mathcal{N}^\perp} U^S \). Hence
\[ (I - \Pi_{VS}) f = (I - \Pi_{\mathcal{N}^\perp}) f - \Pi_{U^S_{\mathcal{N}^\perp}} f. \]
But
\[ (I - \Pi_{\mathcal{N}^\perp}) f = D_{-\bar{S}}' (D_{-\bar{S}} D_{-\bar{S}}')^{-1} D_{-\bar{S}} f. \]
Moreover, when \( r(U^S) \neq 0 \),
\[
\Pi_{\mathcal{N}^\perp} U^S = (I - \Pi_{\mathcal{N}^\perp}) U^S = D_{-\bar{S}}' (D_{-\bar{S}} D_{-\bar{S}}')^{-1} D_{-\bar{S}} U^S = D_{-\bar{S}}' B^S.
\]
So then
\[ \Pi_{U^S_{\mathcal{N}^\perp}} f = D_{-\bar{S}}' B^S (B^{St} D_{-\bar{S}} D_{-\bar{S}}' B^S)^{-1} B^{St} D_{-\bar{S}} f. \]

\[ \square \]
4.1 Definition of the noise weights

Let $\Omega := A^S A^S \in \mathbb{R}^{(m-\bar{s}) \times (m-\bar{s})}$. Note that $\Omega$ depends on $S$ although we do not express this in our notation. We call the diagonal elements of the matrix $\Omega$ the squared noise weights $\{\omega^2_j\}_{j \in D \setminus \bar{S}}$. They will play a role in the compatibility constant. Small noise weights are good. Note further that

$$\Omega = \begin{cases} (D_S D_S' - \bar{B}^S)^{-1} - B^S (B^S D_S D_S' B^S)^{-1} B^S & r(U^S) \neq 0 \\ (D_S D_S')^{-1} & r(U^S) = 0 \end{cases}.$$ 

Thus, one sees that adding mock variables reduces the noise weights. Define for $\omega_j := 0$ for $j \in \bar{S} \setminus S$: for indices in $\bar{S} \setminus S$ the noise is already taken care of by the indices in $D \setminus \bar{S}$. One has much freedom is choosing $\bar{S}$ and $U^S$. It is good to choose $\bar{S}$ and $U^S$ in such a way that $\|\omega_{-S}\|_\infty$ is small (or even minimized), with the restriction that $r(U^S)$ should (typically) be of the same order as $r_S$.

5 Analysis of the analysis problem

Let $S \subset D$ and take $\lambda > 0$ such that

$$\frac{\lambda_S(u)}{\sqrt{n\lambda}} \|\omega_{-S}\|_\infty \leq 1$$

where

$$\lambda_S(u) = \sqrt{\frac{2\log(2(n - r(V^S))) + 2u}{n}}$$

with $u > 0$ playing a role in the confidence level of the oracle result of Theorem 5.1 below. Then we let in the effective sparsity $\Gamma^2(S, w_{-S})$ the collection of weights $w_{-S}$ be such that for all $j \in D \setminus S$

$$w_j \geq \frac{\lambda_S(u)}{\sqrt{n\lambda}} \omega_j.$$

(2)

**Theorem 5.1** Fix an arbitrary $f \in \mathbb{R}^n$. For all $u > 0$, $v > 0$, and for

$$\lambda \geq \lambda_S(u) n^{-\frac{1}{2}} \|\omega_{-S}\|_\infty,$$

we have with probability at least $1 - \exp[-u] - \exp[-v]$}

$$\|\hat{f} - f^0\|_2^2/n \leq \|f - f^0\|_2^2/n + \left(\frac{r(V^S)}{n} + \frac{2v}{n} + \lambda \Gamma(S, w_{-S})\right)^2 + 4\lambda \|D_{-S} f\|_1$$

where $w_{-S}$ is assumed to satisfy (2).

A proof can be found in Subsection 11.1.
6 TV regularization of $f$

Consider the total variation penalty

$$\|Df\|_1 := \sum_{j=2}^{n} |f_j - f_{j-1}|.$$ 

Let $\{\phi_j\}_{j=1}^{n}$ be the step functions

$$\phi_j(i) = 1\{i \geq j\}, \ i,j \in \{1, \ldots, n\}.$$ 

Then we can write

$$f = \sum_{j=1}^{n} \beta_j \phi_j,$$

where $\beta_1 = f_1$ and for $j = 2, \ldots, n$

$$\beta_j = f_j - f_{j-1}.$$ 

Moreover

$$\|Df\|_1 = \sum_{j=2}^{n} |\beta_j|.$$ 

Let $D := \{2, \ldots, n\}$ and $S := \{t_1, \ldots, t_s\}$, $t_1 < \cdots < t_s$, and $t_{s+1} := n - t_s$. One may think of $S$ as the locations of the jumps of (an approximation $f$ of) $f^0$. Assume for simplicity that each distance $t_k - t_{k-1}$ is even and define $d_k := (t_k - t_{k-1})/2$, $k = 2, \ldots, s$. Let $d_1 := t_1 - 1$ and $d_{s+1} = n - t_s$. We take $V^S := N_{-S}$ so that

$$r(V^S) = s + 1.$$ 

In other words, in this case we do not add any mock variables. One can calculate $\omega_{-S}$ exactly as is done in [Ortelli and van de Geer 2018]. We alternatively present here an upper bound. This facilitates the comparison with the results for the total variation penalty on higher order differences as given in Sections 7 and 8.

**Theorem 6.1** We have $\omega_j^2 \leq \tilde{\omega}_j^2$ for all $j \in D \setminus S$, where

$$\tilde{\omega}_j^2 := \begin{cases}
|t_1 - j| & 2 \leq j \leq t_1 + d_2, \ j \neq t_1 \\
|t_2 - j| & t_1 + d_2 \leq j \leq t_2 + d_3, \ j \neq t_2 \\
& \vdots \\
|t_s - j| & t_{s-1} + d_s \leq j \leq n, \ j \neq t_s
\end{cases}$$

**Proof of Theorem 6.1** Clearly, for $j \geq t$

$$\phi_i(i) - \phi_j(i) = \begin{cases}
0 & i < t \\
1 & t \leq i < j \\
0 & i \geq j
\end{cases}$$
so that

$$\|\phi_t - \phi_j\|^2 = j - t.$$  

One sees that

$$\|\Pi_{\mathcal{N}^+} \phi_j\|^2 \leq \min_{k \in \{t_1, \ldots, t_s\}} |j - k|$$

Hence (for \( j \notin \{1, t_0, \ldots, t_s\} \))

$$\|\Pi_{\mathcal{N}^+} \phi_j\|^2 \leq \begin{cases} |t_1 - j| & 2 \leq j < t_1 \\ |t_1 - j| & t_1 < j \leq t_1 + d_2 \\ |t_2 - j| & t_1 + d_2 \leq j < t_2 \\ \vdots & \vdots \\ |t_s - j| & t_s < j \leq n \end{cases}$$

\( \Box \)

**Theorem 6.2** Let \( \tilde{w}^2_{-S} \) be as in Theorem 6.1 and define for \( j \in \mathcal{D}\setminus S \)

$$w_j^2 := \begin{cases} \tilde{w}_j^2/d_1 & 2 \leq j < t_1 \\ \tilde{w}_j^2/d_k & t_{k-1} < j < t_k, \ k \in \{2, \ldots, s+1\} \\ \tilde{w}_j^2/d_{s+1} & t_s < j \leq n \end{cases}$$

Then

$$\Gamma^2(S, w_{-S}) \leq \sum_{k=1}^{s+1} n \log(d_k + 1)/d_k.$$  

**Proof of Theorem 6.2.** To be able to write explicit expressions, let \( s \) be even (say). Take \( z_S := (+1, -1, \ldots, -1)' \). This is one of the two hardest cases for an interpolating vector \( q = q(z_S) \) in Lemma 2.1 (the other case being \( -z_S \)). The following vector \( q := (q_2, \ldots, q_n)' \) will be the interpolating vector for \( z_S \):

$$q_j := \begin{cases} +1 - \sqrt{t_1 - j}/\sqrt{d_1} & 2 \leq j \leq t_1 \\ +1 - \sqrt{j - t_1}/\sqrt{d_2} & t_1 + 1 \leq j \leq t_1 + d_2 \\ -1 + \sqrt{t_2 - j}/\sqrt{d_2} & t_1 + d_2 + 1 \leq j \leq t_2 \\ \vdots & \vdots \\ -1 + \sqrt{j - t_s}/\sqrt{d_{s+1}} & t_s + 1 \leq j \leq n \end{cases}$$

Observe that \( q_j \) can be seen as the weight attached to the edge between node \( j \) and node \( j - 1, j = 2, \ldots, n \). Moreover, \( q_S = z_S \).

Then

$$D'q = \begin{pmatrix} q_2 \\ q_3 - q_2 \\ \vdots \\ q_n - q_{n-1} \\ -q_n \end{pmatrix} = \begin{pmatrix} 1 - \sqrt{d_1 - 1}/\sqrt{d_1} \\ \sqrt{d_1 - 2}/\sqrt{d_1} - \sqrt{d_1 - 1}/\sqrt{d_1} \\ \vdots \\ -1 + \sqrt{d_{s+1} - 1}/\sqrt{d_{s+1}} \\ 0 \end{pmatrix}.$$
It follows that

\[
\|D'q\|_2^2 = \sum_{j=2}^{t_1} \left( \sqrt{d_1 + 1 - j} / \sqrt{d_1} - \sqrt{d_1 + 2 - j} / \sqrt{d_1} \right)^2 + \sum_{j=t_1+1}^{t_1+d_2} \left( \sqrt{j - t_1} / \sqrt{d_2} - \sqrt{j - 1 - t_1} / \sqrt{d_2} \right)^2 + \sum_{j=t_1+d_2+1}^{t_2} \left( \sqrt{t_2 - j} / \sqrt{d_2} - \sqrt{t_2 + 1 - j} / \sqrt{d_2} \right)^2 + \cdots + \sum_{j=t_{s+1}}^{n} \left( \sqrt{j - t_s} / \sqrt{d_{s+1}} - \sqrt{j - 1 - t_s} / \sqrt{d_{s+1}} \right)^2.
\]

We have for any \( d \in \mathbb{N} \)

\[
\sum_{j=1}^{d} \left( \sqrt{j} - \sqrt{j - 1} \right)^2 \\
\sum_{j=1}^{d} \left( \frac{1}{\sqrt{j + \sqrt{j - 1}}} \right)^2 \leq \sum_{j=1}^{d} \frac{1}{j} \leq \log(d + 1).
\]

It follows that

\[
\|D'q\|_2^2 \leq \sum_{k=1}^{s+1} \frac{\log(d_k + 1)}{d_k}.
\]

\(\square\)

**Theorem 6.3** Let \( f \in \mathbb{R}^n \) be arbitrary. Let

\[
d_{\text{max}} := \max_{1 \leq k \leq s+1} d_k.
\]

For all \( u > 0 \), \( v > 0 \), and for

\[
\lambda \geq \lambda_S(u) \sqrt{d_{\text{max}} / n},
\]

we have with probability at least \( 1 - \exp[-u] - \exp[-v] \)

\[
\|f - f^0\|_2^2 / n \leq \|f - f^0\|_2^2 / n + \left( \sqrt{\frac{s+1}{n}} + \sqrt{\frac{2v}{n}} + \lambda \Gamma(S, w_{-S}) \right)^2 + 4\lambda \|D_{-S}f\|_1.
\]

with

\[
\Gamma^2(S, w_{-S}) \leq \sum_{k=1}^{s+1} \frac{n \log(d_k + 1)}{d_k}.
\]

11
7 TV regularization of the first differences of $f$

Let

$$\|Df\|_1 := \sum_{j=3}^{n} |f_j - 2f_{j-1} + f_{j-2}|.$$ 

We can write

$$f = \sum_{j=1}^{n} \beta_j \psi_j(\cdot)$$

where

$$\beta_1 := f_1, \beta_2 := f_2 - f_1, \beta_j := f_j - 2f_{j-1} + f_{j-2}, j \geq 3.$$ 

and where

$$\psi_1(i) \equiv 1, \psi_2(i) = (i - 1), \psi_j(i) = (i - j + 1) \mathbb{I}\{i \geq j\}, j \geq 3.$$ 

These are sometimes called ReLU (Rectifier Linear Unit) functions. Define moreover the step functions

$$\phi_j(i) := \mathbb{I}\{j \geq i\}, i, j \in \{1, \ldots, n\}.$$ 

Let $\mathcal{D} := \{3, \ldots, n\}$, $S := \{t_1, \ldots, t_s\}$, $2 < t_1 < \cdots < t_s$, and $t_{s+1} := n - t_s$. On may think of $S$ as the location of the kinks of (an approximation $f$ of) $f_0$. Assume for simplicity that each distance $t_k - t_{k-1}$ is even and define $d_k := (t_k - t_{k-1})/2, k = 1, \ldots, s+1$, where $t_0 := 2$ and $t_{s+1} = n$. Our mock variables will be $\{\phi_{tk}\}_{k=1}^{s}$: we take

$$\mathcal{V}^S := \mathcal{N}_S \oplus \text{span}(\{\phi_{tk}\}_{k=1}^{s}).$$

**Theorem 7.1** We have $\omega_j \leq \tilde{\omega}_j$ for all $j \notin S$ where

$$\tilde{\omega}_j^2 := 2 \begin{cases} 
|j - 2|^3 & 3 \leq j \leq d_1 + 2 \\
|t_1 - j|^3 & d_1 + 2 \leq j < t_1 + d_2, j \neq t_1 \\
|t_2 - j|^3 & t_1 + d_2 \leq j \leq t_2 + d_3, j \neq t_2 \\
\vdots & \vdots \\
|t_s - j|^3 & t_{s-1} + d_s \leq j \leq t_s + d_{s+1}, j \neq t_s 
\end{cases}.$$
Proof of Theorem 7.1. Fix a \( t \in [2, n-2] \) and let \( j > t \). Then

\[
\psi_t(i) - (j - t)\phi_t(i) - \psi_j(i) = \begin{cases} 
0 & i < t \\
(t - i) & i \leq j - 2 \\
(i - j + 1) & i \geq j - 1 
\end{cases}.
\]

It follows that \( \psi_t - (j - t)\phi_t - \psi_j = 0 \) when \( j = t + 1 \) and for \( j \geq t + 2 \)

\[
\|\psi_t - (j - t)\phi_t - \psi_j\|_2^2 \leq \frac{(j - t)^3}{3}.
\]

If \( j < t \) we get

\[
\psi_j(i) - (t - j)\phi_t(i) - \psi_t(i) = \begin{cases} 
0 & i < j \\
(t - i) & j \leq i \leq t - 1 \\
(i - j + 1) & i \geq t 
\end{cases}
\]

and

\[
\|\psi_j - (t - j)\phi_t - \psi_t\|_2^2 \leq 2(t - j)^3.
\]

Thus with

\[
V^S := \mathcal{N}_S \oplus \text{span}(\{\phi_k\}_{k=1}^s)
\]

we find for \( t_{k-1} < j < t_k, k \in \{1, \ldots, s+1\} \) (where \( t_0 := 2 \) and \( t_{s+1} := n \))

\[
\|\Pi_{V^S} \psi_j\|_2^2 \leq 2 \min\left\{ (j - t_{k-1})^3, (t_k - j)^3 \right\}.
\]

In other words, for \( j \notin S, j \geq 3 \)

\[
\|\Pi_{V^S} \psi_j\|_2^2 \leq 2 \begin{cases} 
(j - 2)^3 & 3 \leq j \leq d_1 + 2 \\
(t_1 - j)^3 & d_1 + 2 \leq j < t_1 \\
(j - t_1)^3 & t_1 < j \leq t_1 + d_2 \\
(t_2 - j)^3 & t_1 + d_2 \leq j < t_2 \\
\vdots & \vdots \\
(j - t_s)^3 & t_s < j \leq t_s + d_{s+1} \\
(n - j)^3 & t_s + d_{s+1} \leq j \leq n
\end{cases}.
\]

\( \square \)

Theorem 7.2 Let \( \tilde{\omega}_S^2 \) be as in Theorem 7.1 and define for \( j \notin S \)

\[
w_j^2 := \begin{cases} 
\tilde{\omega}_j^2/(8d_1^2) & 3 \leq j < t_1 \\
\tilde{\omega}_j^2/(2d_k^3) & t_{k-1} < j < t_k, \ k \in \{2, \ldots, s\} \\
\tilde{\omega}_j^2/(8d_{s+1}) & t_s < j \leq n
\end{cases}.
\]

Then for a universal constant \( C_2 \)

\[
\Gamma^2(S, w_S) \leq C_2^2 \sum_{k=1}^{s+1} \frac{n \log d_k}{d_k}.
\]
Proof of Theorem 7.2

Assume for simplicity that \( s \) is even and take as in the proof of Theorem 7.2 \( z_S := (+1, -1, \ldots, -1)' \), which is (modulo a sign flip) the hardest case for the interpolating vector \( q \) in Lemma 2.1. As interpolating \( q = (q_3, \ldots, q_n)' \) we have

\[
q_j = \begin{cases}
  +(j - 2)^{3/2}/(2d_1^{3/2}) & 3 \leq j \leq d_1 + 2 \\
  +1 - (t_1 - j)^{3/2}/(2d_1^{3/2}) & d_1 + 2 \leq j \leq t_1 \\
  +1 - (j - t_1)^{3/2}/d_2^{3/2} & t_1 \leq j \leq t_1 + d_2 \\
  -1 + (t_2 - j)^{3/2}/d_2^{3/2} & t_1 + d_2 \leq j \leq t_2 \\
  \vdots & \\
  -1 + (j - t_s)^{3/2}/(2d_{s+1}^{3/2}) & t_s \leq j \leq t_s + d_{s+1} \\
  - (n - j)^{3/2}/(2d_{s+1}^{3/2}) & t_s + d_{s+1} \leq j \leq n
\end{cases}
\]  

(3)

Since

\[
D'q = \begin{pmatrix}
  q_3 \\
  -2q_3 + q_4 \\
  q_3 - 2q_4 + q_5 \\
  \vdots \\
  q_{n-2} - 2q_{n-1} + q_n \\
  q_{n-1} - 2q_n \\
  q_n
\end{pmatrix}
\]
we get, tacitly assuming that \( d_1 \geq 5 \),

\[
\frac{\|D'q\|^2}{2} = (q_3)^2 + (-2q_3 + q_4)^2 + \sum_{j=5}^{d_1} (q_j - 2q_{j-1} + q_{j-2})^2 \\
+ (q_{d_1+1} - 2q_{d_1} + q_{d_1-1})^2 + \sum_{j=d_1+2}^{t_1} (q_j - 2q_{j-1} + q_{j-2})^2 \\
+ (q_{t_1+1} - 2q_{t_1} + q_{t_1-1})^2 + \sum_{j=t_1+2}^{t_1+d_2} (q_j - 2q_{j-1} + q_{j-2})^2 \\
+ (q_{t_1+d_2+1} - 2q_{t_1+d_2} - q_{t_1+d_2-1})^2 \\
+ \sum_{j=t_1+d_2+2}^{t_2} (q_j - 2q_{j-1} + q_{j-2})^2 \\
+ (q_{t_2+1} - 2q_{t_2} + q_{t_2-1})^2 + \sum_{j=t_2+2}^{t_2+d_3} (q_j - 2q_{j-1} + q_{j-2})^2 \\
+ \ldots \\
+ \sum_{j=t_s+d_{s+1}+2}^{t_s+d_{s+1}} (q_j - 2q_{j-1} + q_{j-2})^2 \\
+ (q_{t_s+d_{s+1}+1} - 2q_{t_s+d_{s+1}} - q_{t_s+d_{s+1}-1})^2 \\
+ \sum_{j=t_s+d_{s+1}+2}^{n} (q_j - 2q_{j-1} + q_{j-2})^2 + (-2q_n + q_{n-1})^2 + q_n^2
\]

Insert now the value given in (3) for \( q \) and note that for \( k = 1, \ldots, s + 1 \), at the point of change of regime \( j = t_{k-1} + d_k \):

\[
q_{t_{k-1}+d_k+1} - q_{k} + d_k = q_{t_{k-1}+d_k} - q_{t_{k-1}+d_k-1}.
\]

In other words, at these points there is no contribution to the second order
One may use that
\[ \|D'q\|_2^2 = 1/(4d_3^4) \]
\[ + \frac{1}{4d_1^4} \sum_{j=4}^{d_1+2} \left( (j - 2)^{3/2} - 2(j - 3)^{3/2} + (j - 4)^{3/2} \right)^2 \]
\[ + \frac{1}{4d_1^4} \sum_{j=d_1+4}^{t_1} \left( (t_1 - j)^{3/2} - 2(t_1 - j + 1)^{3/2} + (t_1 - j + 2)^{3/2} \right)^2 \]
\[ + \left( 1/(2d_1^{3/2}) + 1/d_1^{3/2} \right)^2 \]
\[ + \frac{1}{d_2^3} \sum_{j=t_1+2}^{t_1+d_2} \left( (j - t_1)^{3/2} - 2(t - t_1 - 1)^{3/2} + (j - t_1 - 2)^{3/2} \right)^2 \]
\[ + \frac{1}{d_2^3} \sum_{j=t_1+d_2+2}^{t_2} \left( (t_2 - j)^{3/2} - 2(t_2 - j + 1)^{3/2} + (t_2 - j - 1 + 2)^{3/2} \right)^2 \]
\[ + \left( 1/d_2^{3/2} + 1/d_2^{3/2} \right)^2 \]
\[ + \frac{1}{d_3^3} \sum_{j=t_2+2}^{t_2+d_3} \left( (j - t_2)^{3/2} - 2(j - t_2 - 1)^{3/2} + (j - t_2 - 2)^{3/2} \right)^2 \]
\[ + \ldots \]
\[ + \frac{1}{4d_{s+1}^3} \sum_{j=t_s+2}^{t_s+d_{s+1}} \left( (j - t_s)^{3/2} - 2(j - t_s - 1)^{3/2} + (j - t_s - 2)^{3/2} \right)^2 \]
\[ + \frac{1}{4d_{s+1}^3} \sum_{j=t_s+d_{s+1}+2}^{t_{s+1}} \left( (t_{s+1} - j)^{3/2} - 2(t_{s+1} - j + 1)^{3/2} + (t_{s+1} - j + 2)^{3/2} \right)^2 \]
\[ + 1/(4d_{s+1}^3). \]

One may use that
\[ j^{3/2} - 2(j - 1)^{3/2} + (j - 2)^{3/2} = \frac{3}{2} \sqrt{u} - \frac{3}{2} \sqrt{v} \]
where \( u \in [j - 1, j] \) and \( v \in [j - 2, j - 1] \). So
\[ |j^{3/2} - 2(j - 1)^{3/2} + (j - 2)^{3/2}| = \frac{3}{2} \frac{u - v}{\sqrt{u} + \sqrt{v}} \]
\[ \leq 3 \frac{1}{\sqrt{u} + \sqrt{v}} \leq 3 \frac{1}{\sqrt{j - 1}}. \]

To conclude, for a universal constant \( C_2 \) we have
\[ \|q\|_2^2 \leq C_2^2 \sum_{k=1}^{s+1} \frac{\log d_k}{d_k^3}. \]
\[ \square \]
**Theorem 7.3**  Fix an arbitrary $f \in \mathbb{R}^n$. Let
\[
 d_{\text{max}} := \max_{1 \leq k \leq s+1} d_k.
\]

For all $u > 0$, $v > 0$, and for
\[
 \lambda \geq \lambda_S(u) \sqrt{2} n^{-\frac{1}{2}} d_{\text{max}}^{\frac{3}{2}}
\]
we have with probability at least $1 - \exp[-u] - \exp[-v]$
\[
 \|\hat{f} - f^0\|_2^2/n \leq \|f - f^0\|_2^2/n + \left( \sqrt{\frac{2(s+1)}{n}} + \sqrt{\frac{2v}{n} + \lambda \Gamma(S, w_-)} \right)^2 + 4\lambda \|D_S f\|_1.
\]

with for a universal constant $C_2$
\[
 \Gamma^2(S, w_-) \leq C_2^2 \sum_{k=1}^{s+1} \frac{n \log d_k}{d_k^2}.
\]

**Proof of Theorem 7.3**  We have with $\tilde{w}_-$ given in Theorem 7.1
\[
 \|\tilde{w}_-\|_\infty \leq 2d_{\text{max}}^3.
\]

So for $j \in \mathcal{D}\setminus S$
\[
 \frac{\lambda_S(u) \tilde{w}_j}{\sqrt{n}\lambda} \leq w_j
\]
with $w_-$ given in Theorem 7.2. Moreover $\dim(V^S) = 2(s+1)$. The result thus follows from combining Theorems 5.1, 7.1 and 7.2.

---

**8 Proof of Theorem 3.1**

Fix some $N \in \mathbb{N}$ and recall the notation of Section 3
\[
 \Delta^0 f = f, \ \Delta^1 := \Delta(\Delta^0 f)
\]
The analysis operator penalty is taken as
\[
 \|Df\|_1 := \sum_{j=N+1}^{n} |(\Delta^N f)_j|.
\]

Let $\mathcal{D} = \{N+1, \ldots, n\}$, $S = \{t_1, \ldots, t_s\}$, $N+1 < t_1 < \cdots < t_s$. Let $t_0 := N$ and $t_{s+1} := n$.  

---

17
8.1 The dictionary

One may write

$$ f = \sum_{j=1}^{n} \beta_j \psi_j^{(N)}(\cdot), $$

where for $i = 1, \ldots, n$

$$ \psi_j^{(N)}(i) = \phi_j^{(N)}(i) \mathbb{1}\{i \geq j\} $$

with, for $j \leq N$, $\phi_j^{(N)}$ a polynomial of degree $j - 1$, and with, for $j > N$, $\phi_j^{(N)}$ a polynomial of degree $N - 1$. We choose the constants in these polynomials properly so that

$$ \beta_j = \begin{cases} (\Delta j^{-1} f)_j & j = 1, \ldots, N \\ (\Delta N f)_j & j > N \end{cases}. $$

8.2 Mock variables and projections

By taking $V^S$ as the direct product of $\mathcal{N}_S$ and the span of the mock variables $\{\psi_l^{(l)}, k = 1, \ldots, s, l = 1, \ldots, N\}$ one sees that the length of the projection of an inactive variable $\psi_j^{(N)}$ on $V^S \perp$ is at most $\tilde{c}_N \min_{k \in \{0, 1, \ldots, s+1\}} |t_k - j| 2^{N-1}$ for a constant $\tilde{c}_N$ depending only on $N$. Moreover, $r(V^S) = N(s + 1)$.

8.3 Building a system of linear equations

Fix some $k \in \{1, \ldots, s + 1\}$, say some $k$ in the interior $\{2, \ldots, s\}$. Say we aim at interpolation the value $+1$ at $t_{k-1}$ to the value $-1$ at $t_k$. We choose $q_j$ odd around $(t_{k-1} + t_k)/2$, i.e.

$$ q_{tk-1+t_k+j} = -q_{tk-1+t_k-j}, \; j \in [0, (t_k - t_{k-1})/2]. $$

For $j \geq t_{k-1}$ near $t_{k-1}$ we let

$$ q_{t_{k-1}-j} = 1 - a_k(j - t_{k-1}) \frac{2^{N-1}}{2}. $$

where $a_k > 0$ is to be determined. Near $t_k$ we will then have

$$ q_{tk-j} = -1 + a_k(t_k - j) \frac{2^{N-1}}{2}. $$

Note that we need not match $N$th differences at the $N$ points $j$ to the left of $t_{k-1}$ because there $q_{j-1}$ is small enough in itself (that is, of order $(t_{k-1} - t_{k-2})^{-\frac{2N-1}{2}}$).

Thus, we can define the polynomials separately on each interval $[t_{k-1}, t_k]$, $k = 1, \ldots, s + 1$ (in that sense the problem is localized).

We continue with some fixed $k \in \{2, \ldots, s\}$. Because we decided for $q_j$ odd near the midpoint $(t_{k-1} + t_k)/2$ we take $q_j$ here as a polynomial in $(t_{k-1} - j)^l$ where $l \leq N$ is odd.
We now split an interval \((t_{k-1}, t_k)\) in enough pieces of equal length to have \(\Delta^N q\) almost piecewise constant. For \(N = 3\) we split each interval \((t_{k-1}, t_k)\) into 4 pieces. Let us assume therefore that \(d_k := (t_k - t_{k-1})/4\) is an integer, \(k = 1, \ldots, s + 1\). For \(N = 4\) we split each interval \((t_{k-1}, t_k)\) into 6 pieces. In general for \(N\) is even we split the first half of the interval in \(N/2 + 1\) pieces. We then have 1 unknown (the coefficient of \((t_{k-1} - j)^N\) near \(j = t_{k-1}\)) plus \(N/2\) unknowns (the coefficients of \((t_{k-1} - j)^l\) near \(j = (t_{k-1} + t_k)/2\) for the \(N/2\) odd values of \(l \in \{1, 3, \ldots, N - 1\}\)). To join the endpoints of the splits we have \(N\) equations. So \(N/2 - 1\) extra splits give \(N(N/2)\) equations and \(N/2 + 1 + (N/2 - 1)(N + 1) = (N/2)/N\) unknowns. Thus for \(N\) even we split the interval into \(N + 2\) subintervals.

When \(N\) is odd we have for the first half interval 1 unknown (the coefficient \(a_k > 0\) of \((t_{k-1} - j)^{2N-1} / 2\) near \(j = t_{k-1}\)) plus \((N + 1)/2\) unknowns (the coefficients of \((t_{k-1} - j)^l\) near \(j = (t_{k-1} + t_k)/2\) for the \((N + 1)/2\) odd values of \(l \in \{1, 3, \ldots, N\}\)). So \((N - 1)/2 - 1\) extra splits give \(N(N - 1)/2\) equations and \((N + 1)/2 + 1 + ((N - 1)/2 - 1)(N + 1) = (N - 1)/2\) unknowns. Thus for \(N\) odd we split the interval into \(N + 1\) subintervals.

Note that \(N = 1\) is odd, leading to split in two intervals, as we did in Section 6. The value \(N = 2\) is even so then we would split in 4 intervals. However, in Section 7 we saw that a split in two suffices. This is due to the fact that \(q_j\) is odd around \((t_{k-1} + t_k)/2\). The operation \(\Delta^2\) concerns only three indices. The only triple \((j, j - 1, j - 2)\) near \((t_{k-1} - t_k)/2\) that involves a change is regime is the one with \(j = (t_{k-1}s + t_k)/2\) and at that value of \(j\), \(\Delta^2 q_j = 0\). In other words

\[
q_{t_{k-1} + d_k} = 0, \quad \Delta q_{t_{k-1} + d_k + 1} = \Delta q_{t_{k-1} + d_k}
\]

We could however also have chosen for a split into 4 intervals for the case \(N = 2\), with in the middle two intervals a linear interpolation such that the differences match at the endpoint of the first quarter and then also (building up the interpolation anti-symmetrically around the midpoint) at the endpoint of the third quarter.

To simplify the notation it helps to assume that each distance \(t_k - t_{k-1}\) is a multiple of \(N + 2\) when \(N\) is even and a multiple of \(N + 1\) when \(N\) is odd, and then define \(d_k := (t_k - t_{k-1})/(N + 2)\), when \(N\) is even and \(d_k := (t_k - t_{k-1})/(N + 1)\) when \(N\) is odd, \(k = 1, \ldots, s + 1\).

### 8.4 Details of the system for the case \(N = 3\)

When \(N = 3\) we take for \(i = 1, \ldots, n\)

\[
\psi_1^{(3)}(i) = 1, \quad \psi_2^{(3)}(i) = (i - 1)\mathbb{1}\{i \geq 2\}, \quad \psi_3^{(3)}(i) = \frac{(i - 1)(i - 2)}{2}\mathbb{1}\{i \geq 3\}
\]

and

\[
\psi_j^{(3)}(i) = \frac{(i - j + 1)(i - j + 2)}{2}\mathbb{1}\{i \geq j\}, \quad j \geq 4.
\]
Then
\[ f = \sum_{j=1}^{n} \beta_j \psi_j^{(3)} \]
with
\[ \beta_j = \begin{cases} 
  f_1 & j = 1 \\
  f_2 - f_1 & j = 2 \\
  f_3 - 2f_2 + f_1 & j = 3 \\
  (\Delta^3 f)_j & j > 3 
\end{cases} \]

We consider in this subsection the interpolating vector for \( N = 3 \). A prototype has the following form.

**Lemma 8.1** Let \( d \in \mathbb{N} \) and define
\[
\begin{align*}
\alpha_1 & := \frac{[\Delta(d+1)^{5/2}]}{d^{5/2}} = \frac{(d+1)^{5/2} - d^{5/2}}{d^{5/2}} \\
\gamma_1 & := \frac{\Delta^3}{d^2} := \frac{d^3 - (d-1)^3}{d^2} \\
\alpha_2 & := \frac{[\Delta^2(d+2)^{5/2}]}{d^{1/2}} := \frac{[\Delta(d+2)^{5/2}] - [\Delta(d+1)^{5/2}]}{d^{1/2}} \\
\gamma_2 & := \frac{\Delta^2 d^3}{d} := \frac{[\Delta^3] - [\Delta(d-1)^3]}{d}.
\end{align*}
\]

Let
\[
\begin{align*}
a & := \frac{\gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1 \alpha_2 + \alpha_1 \gamma_2)} \\
b & := \frac{\alpha_2}{\gamma_2 - \alpha_2 + (\gamma_1 \alpha_2 + \alpha_1 \gamma_2)} \\
c & := \frac{\gamma_1 \alpha_2 + \alpha_1 \gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1 \alpha_2 + \alpha_1 \gamma_2)}
\end{align*}
\]

and for \( j \in \{d, d+1, d+2\} \)
\[
\begin{align*}
q_j & := 1 - a_j^{5/2}/d^{5/2} \\
p_j & := -b(2d-j)^3/d^3 + c(2d-j)/d.
\end{align*}
\]

Then
\[ \Delta^l q_{d+l} = \Delta^l p_{d+l}, \ l \in \{0, 1, 2\}. \]

**Proof of Lemma 8.1.** First
\[
\begin{align*}
\Delta^2 q_{d+2} & = -a \frac{[\Delta^2(d+2)^{5/2}]}{d^{5/2}} = -\frac{a \alpha_2}{d^2} \\
& = -\frac{1}{d^2} \times \frac{\alpha_2 \gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1 \alpha_2 + \alpha_1 \gamma_2)} \\
\Delta^2 p_{d+2} & = -\frac{b \Delta^2 d^3}{d^3} = -\frac{b \gamma_2}{d^2} = -\frac{1}{d^2} \times \frac{\alpha_2 \gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1 \alpha_2 - \alpha_1 \gamma_2)}.
\end{align*}
\]
Second

\[ \Delta q_{d+1} = -\frac{a[\Delta(d + 1)^{5/2}]}{d^{5/2}} = -\frac{a\alpha_1}{d} = -\frac{1}{d} \frac{\gamma_2\alpha_1}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)} \]

\[ \Delta p_{d+1} = \frac{b[\Delta d^3]}{d^3} - \frac{c}{d} = \frac{b\gamma_1}{d} - \frac{c}{d} = \frac{1}{d} \frac{\alpha_2\gamma_1 - (\gamma_1\alpha_2 + \alpha_1\gamma_2)}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)} = -\frac{1}{d} \frac{\alpha_1\gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)}. \]

Finally

\[ q_d = 1 - a = 1 - \frac{\gamma_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)} = \frac{\gamma_1\alpha_2 + \alpha_1\gamma_2 - \alpha_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)} \]

\[ p_d = -b + c = \frac{\gamma_1\alpha_2 + \alpha_1\gamma_2 - \alpha_2}{\gamma_2 - \alpha_2 + (\gamma_1\alpha_2 + \alpha_1\gamma_2)}. \]

\[ \Box \]

The values of the parameters \( a, b \) and \( c \) in the above prototype lemma depend on \( d \), but one easily checks that one can bound them from above and below by universal constants. For \( k \in \{1, \ldots, s + 1\} \) we replace \( d \) by \( d_k \) and call \((a_k, b_k, c_k)\) the corresponding value for \((a, b, c)\). For each \( k \in \{2, \ldots, s\} \) and for \( j \) in the interval the interval \([t_{k-1}, t_k]\) we take

\[
q_j := \begin{cases} 
1 - \frac{a_k(j-t_{k-1})^{5/2}}{d_k^{5/2}} & t_{k-1} \leq j \leq t_{k-1} + d_k + 2 \\
-b_k \frac{(t_{k-1}+2d_k-j)^3}{d_k^3} + c_k \frac{(t_{k-1}+2d_k-j)^2}{d_k^2} & t_{k-1} + d_k \leq j \leq t_k - d_k \\
-1 + \frac{a_k(t_k-j)^{5/2}}{d_k^{5/2}} & t_k - d_k - 2 \leq j \leq t_k 
\end{cases}
\]

For the two intervals \([t_0, t_1]\) and \([t_s, t]\) one uses the same formulas, but rescaled because one has to interpolate from \(+1\) (or \(-1\)) to \(0\) instead of from \(+1\) to \(-1\).

### 8.5 \( N \)th order differences of the interpolation

One easily checks that for a polynomial of degree \( N \)

\[ p_j = a_0 + a_1 j + \cdots + j^N, \quad j = 1, \ldots, d \]

it holds that

\[ (\Delta^{(N)} p)_j = N!, \quad N + 1 \leq j \leq d. \]

Fix some \( k \in \{1, \ldots, s + 1\} \). We take \( q_j \) as a polynomial of degree \( N \) in the interior subintervals of \([t_{k-1}, t_k]\), so here, for \( j \in [t_{k-1} + d_k, t_k - d_k] \) the \( N \)th order differences are piecewise constant.

At \( j \) in the two boundary intervals \([t_{k-1}, t_{k-1} + d_k]\) and \([t_k - d_k, t_k]\) we make sure that \(|q_j|\) is small enough to take care of the noise. For \( j \in [t_{k-1} + d_k, t_{k-1} + \]

21
For some $d \in \mathbb{N}$, $d \geq 2N$, we let $q_j$ match with the polynomial of degree $N$. Moreover at these boundary intervals, say the left boundary where $j \in [t_k - 1, t_k - d + 1]$, we use that for $q_j := 1 - a_k (j - t_k - d + 1)^{2N} / d_k^{2N}$, where the constant $a_k$ is bounded from above and below by a constant depending only on $N$. The following lemma is inserted to control the $N^{th}$ order differences in the two boundary intervals $[t_k - 1, t_k - d + 1]$ and $[t_k - d, t_k]$, it uses that $N^{th}$ order differences behave like $N^{th}$ order derivatives.

**Lemma 8.2** Let for some $d \in \mathbb{N}$, $d \geq 2N$,

$$q_j := \frac{j^{2N-1}}{2N-1}, \ j = N, \ldots, d.$$

Then for some constant $\tilde{C}_N$

$$\|\Delta^N q\|_2^2 \leq \tilde{C}_N^2 \log(1 + d).$$

**Proof of Lemma 8.2.** We have for $j \geq N$

$$\Delta^N j^{2N-1} = \sum_{l=0}^{N} \binom{N}{l} (-1)^l (j - l)^{2N-1}$$

$$= j^{2N-1} \left[ \sum_{l=0}^{N} \binom{N}{l} (-1)^l \left(1 - \frac{l}{j}\right)^{2N-1} \right].$$

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

$$(1 - x)^{2N-1} = \sum_{k=0}^{N-1} a_k x^k + \text{rem}(x)$$

where $a_0 = 1$, $a_1 = -\frac{2N-1}{2}$, $\ldots$ are the coefficients of the Taylor expansion and where the remainder $\text{rem}(x)$ satisfies for some constant $C_N$

$$\sup_{0 \leq x \leq 1/2} |\text{rem}(x)| \leq C_N |x|^N.$$

Thus

$$\sum_{l=0}^{N} \binom{N}{l} (-1)^l \left(1 - \frac{l}{j}\right)^{2N-1}$$

$$= \sum_{l=0}^{N} \binom{N}{l} (-1)^l \sum_{k=0}^{N-1} a_k \frac{l^k}{j^k} + \sum_{l=0}^{N} \binom{N}{l} (-1)^l \text{rem} \left( \frac{l}{j} \right)$$

$$= \Delta^N p_l + \sum_{l=0}^{N} \text{rem} \left( \binom{N}{l} (-1)^l \left( \frac{l}{j} \right) \right)$$

where

$$p_l = (-1)^N \sum_{k=0}^{N-1} a_k \frac{l^k}{j^k}, \ l = 0, \ldots, N - 1$$
is a polynomial of degree \(N - 1\) and hence \(\Delta^N p = 0\). It follows that for \(j \geq 2N\),
\[
\left| \sum_{l=0}^{N} \binom{N}{l} (-1)^l \left( 1 - \frac{l}{j} \right)^{2N-1} \right| \leq \sum_{l=0}^{N} \binom{N}{l} \left| \text{rem} \left( \frac{l}{j} \right) \right| \leq \tilde{C}_N \frac{1}{j^N}.
\]
But then for \(j \geq 2N\)
\[
\Delta^N j^{\frac{2N-1}{2}} \leq \tilde{C}_N \frac{1}{j^{1/2}}.
\]
So
\[
\sum_{j=2N}^{d} \left| \Delta^N j^{\frac{2N-1}{2}} \right|^2 \leq \tilde{C}_N^2 \log(1 + d).
\]
Finally, for \(N \leq j < 2N\),
\[
\Delta^N j^{\frac{2N-1}{2}} \leq j^{\frac{2N-1}{2}} \sum_{l=0}^{N} \binom{N}{l} \leq 2N \cdot j^{\frac{2N-1}{2}}.
\]
Thus
\[
\sum_{j=N}^{d} \left| \Delta^N j^{\frac{2N-1}{2}} \right|^2 \leq (2N)^{2N} + \tilde{C}_N^2 \log(1 + d) \leq \tilde{C}_N^2 \log(1 + d)
\]
for some constant \(\tilde{C}_N\).

### 8.6 Interpolation and weights

Recall we want the interpolating vector \(q\) to satisfy
\[
q_S = z_S, \quad q_j \leq 1 - w_j, \quad j \in \mathcal{D}\setminus S
\]
where, as required in inequality (2),
\[
w_j \geq \frac{\lambda S(u)}{\sqrt{n\lambda}} \omega_j, \quad j \in \mathcal{D}\setminus S.
\]
In the present context, \(\omega_j \leq \tilde{c}_N \min_{t_k \in \{0, \ldots, s+1\}} \left\| t_k - j \right\|^{\frac{2N-1}{2}}\) (see Subsection 8.2). Say \(j \in (t_k-1, t_k)\) is closest to \(t_k\) so that \(\omega_j \leq \tilde{c}_N (t_k - j)^{\frac{2N-1}{2}}\). By the construction in Subsection 8.3 with \(w_j = a_k(t_k - j)^{\frac{2N-1}{2}}\) near \(t_k\), inequality (2) is for met for this value of \(j\). Moving \(j\) more to the middle, we see that \(w_j\) gets larger, and once we reach the first split \(w_j\) stays, for \(N\) fixed, away from zero. In other words, one can make sure (2) holds for all \(j\) by choosing
\[
\lambda \geq c_N \lambda S(u) n^{-\frac{1}{4}} d_{\infty}^{\frac{2N-1}{2}}
\]
with the constant \(c_N\), depending only on \(N\), sufficiently large.
8.7 Finalizing the proof of Theorem 3.1.

We have thus built an interpolating vector $q$:

$$q_S = z_S, \quad q_j = 1 - w_j, \quad j \in \mathcal{D} \setminus S,$$

with

$$\|D' q\|_2^2 \leq C N^s \sum_{k=1}^{s+1} \frac{\log(t_k - t_{k-1})}{(t_k - t_{k-1})^{2N-1}},$$

and

$$w_j \geq \frac{\lambda_S(u)}{\sqrt{n} \lambda} \omega_j, \quad j \in \mathcal{D} \setminus S$$

when $\lambda \geq c_N \lambda_S(u) n^{-\frac{1}{2}} d_\infty^{2N-1}$. Theorem 3.1 now follows from an application of Theorem 5.1.

9 An almost minimax rate

In this section we do not rely on effective sparsity (or compatibility constants) and derive up to log-terms the minimax rate over the class

$$\mathcal{F}_N := \{f^0 : \rho_{f^0} \leq 1\}$$

(4)

where

$$\rho_f := \frac{2(2N - 1)/N}{n} \|\Delta^N f\|_1, \quad f \in \mathbb{R}^n$$

(the scaling with $2(2N - 1)/N$ is merely to simplify the expressions). The point is that the result does not use Dudley’s entropy integral, but instead the same projection arguments as for the adaptive oracle results. This shows that such projection arguments are capable of catching the right rates up to log-terms. Moreover, the result is non-asymptotic with “good” constants. Theorem 9.1 is proved in Subsection 11.2.

**Theorem 9.1** Let $f \in \mathbb{R}^n$ and $s \in \mathbb{N}$ be arbitrary. Take

$$\lambda \geq \frac{(n/s)^{2N-1}}{2} \lambda_S(u)/\sqrt{n}.$$

Then with probability at least $1 - \exp[-u] - \exp[-v]$

$$\|\hat{f} - f^0\|^2_2/n \leq \|f - f^0\|^2_2/n + 4\|Df\|_1 + \left( \sqrt{\frac{N(s+1)}{n}} + \sqrt{\frac{2v}{n}} \right)^2.$$

One can trade off the linear term $s$ against the term $(1/s)^{2N-1}$ appearing in the lower bound for $\lambda$. This gives as optimal $s$:

$$s_f := \left( \lambda_S(u) n \rho_f \right)^{\frac{2}{2N+1}}.$$
which we for simplicity assume to be an integer. So when \( N \) is fixed, \( \rho_f \) remains bounded and \( u \) is of order \( \log n \), one sees that
\[
\| \hat{f} - f^0 \|_2^2 / n \leq \frac{1}{2} N / (2N - 1) n^{-\frac{2N}{2N+1}} (n \lambda_S^2(u))^{\frac{1}{2N+1}}.
\]
with probability at least \( 1 - \exp[-u] - \exp[-v] \). The term \( (n \lambda_S^2(u))^{\frac{1}{2N+1}} \) is an additional log-term of (order \( (\log n) \frac{1}{2N+1} \) when \( u \) is of order \( \log n \)) as compared to the minimax rate in \( \| \cdot \Le_2 \| / n \) over \( \mathcal{D} \), which is \( n^{-\frac{2N}{2N+1}} \).

10 Conclusion

We showed that the approach using interpolating vectors can upper-bound the effective sparsity, i.e. lower-bound the compatibility constant. This can be used in the analysis formulation, as well as in the synthesis formulation. In the latter case one needs that the interpolating vector is in the range of \( X' \), where \( X \) is the design matrix.

We furthermore showed that the use of mock variables can be profitable for the analysis problem. The same it true for the synthesis problem.

In this paper we considered higher order differences of a vector \( f \in \mathbb{R}^n \). One may regard \( f \) as a path graph with nodes \( \{1, \ldots, n\} \) and edges between the nodes \( j \) and \( j - 1, j = 2, \ldots, n \). It is not difficult to extend results to more general graphs. One may think for instance of a cycle which adds to the edge set of the path graph an edge between node 1 and \( n \). In that situation, the
interpolation problem is on the cycle and we always interpolate between $+1$ and $-1$ (and not between $+1$ (or $-1$) and $0$ as was needed for the boundary of the path graph). For general graphs, TV results depend on the configuration $S$ of the potential jumps. We refer to Ortelli and van de Geer [2018] for results using first order differences on tree graphs.

Finally, as we have seen the choice of the tuning parameter $\lambda$ depends on $S$. (One may also consider $\lambda$ as given and then one has to choose $S$ properly, depending on $\lambda$.) We want $S$ to be an approximate active set of $f^0$. In that sense the choice of $\lambda$ depends on what one believes about $f^0$. If these beliefs are violated we do not adapt properly. Thus, in fact the results do not do what “oracle inequalities” promise. This also puts the oracle results for the Lasso in a different daylight: if the non-active variables are highly correlated with the active ones, one should take the tuning parameter smaller. So also for the Lasso, the choice of the tuning parameter generally depends on what one believes to be approximately the active set.

11 Proof of Theorems 5.1 and 9.1

The following lemma is standard and its proof is omitted.

Lemma 11.1 For any $a \in \mathbb{R}^n$ with $\|a\|_2 = 1$ it holds that

$$\mathbb{P}(\epsilon' a > \sqrt{2u}) \leq \exp[-u], \quad \forall \ u > 0.$$  

Moreover, for a linear space $\mathcal{V} \subset \mathbb{R}^n$ with $\dim(\mathcal{V}) = r$ we have

$$\mathbb{P}\left( \sup_{a \in \mathcal{V}, \|a\|_2 = 1} \epsilon' a > \sqrt{r} + \sqrt{2v} \right) \leq \exp[-v], \quad \forall \ v > 0.$$  

We let for some $u > 0$

$$\lambda_S(u) := \sqrt{2 \log(2(n - r_S))/n + 2u/n}.$$  

The next ingredient is a lemma shown in Ortelli and van de Geer [2019], Lemma 2.1, which is based on the KKT conditions.

Lemma 11.2 For all $f \in \mathbb{R}^n$ we have

$$\|\hat{f} - f^0\|_2^2/n + \|\hat{f} - f^0\|_2^2/n - \|f - f^0\|_2^2/n \leq 2\epsilon' (\hat{f} - f)/n + 2\lambda \|Df\|_1 - 2\lambda \|D\hat{f}\|_1.$$  

The following result may be of interest in itself. It is deriving a bound on the empirical process using projection arguments. The lemma is an extension using mock variables of Ortelli and van de Geer [2019], Lemma 2.2.

Lemma 11.3 For all $u > 0$ and $v > 0$ with probability at least $1 - \exp[-u] - \exp[-v]$

$$\epsilon^T f/n \leq \lambda_S(u)^2 \|\omega_{S^c} D_{S^c} f\|_1/\sqrt{n} + \left( \sqrt{r(\mathcal{V}^S)} + \sqrt{2v} \right) \|f\|_2/n, \forall f.$$  

26
Proof of Lemma 11.3. It holds that
\[ \epsilon^T f/n = \epsilon' \Pi_{V S} f/n + \epsilon'(I - \Pi_{V S}) f/n. \]

But
\[ \epsilon'(I - \Pi_{V S}) f/n = \epsilon' A^S D_{-S} f/n \leq \|w^{-1}_S A^S\epsilon\|_{\infty} \|w_{-S} D_{-S} f\|_1/n. \]

where \( \omega^{-1}_S = \{\omega_j^{-1}\}_{j \in D \setminus S}. \)
Let \( T \) be the event
\[ \|w^{-1}_S A^S\epsilon\|_{\infty} / \sqrt{n} \leq \lambda_S(u). \]
Then \( \mathbb{P}(T) \geq 1 - \exp[-u] \) by Lemma 11.1 and the union bound. On \( T \)
\[ \epsilon'(I - \Pi_{V S}) f/n \leq \lambda_S(u) \|w_{-S} D_{-S} f\|_1 / \sqrt{n}. \]

Let \( \mathcal{X} \) be the event
\[ \epsilon' \Pi_{V S} f/n \leq \left( \sqrt{r(V^S)} + \sqrt{2v} \right) \|\Pi_{V S} f\|_2/n. \]
Then \( \mathbb{P}(\mathcal{X}) \geq 1 - \exp[-v] \) by Lemma 11.1 and clearly on \( \mathcal{X} \) also
\[ \epsilon' \Pi_{V S} f/n \leq \left( \sqrt{r(V^S)} + \sqrt{2v} \right) \|\Pi_{V S} f\|_2/n. \quad (5) \]
We have \( \mathbb{P}(T \cap \mathcal{X}) \geq 1 - \exp[-u] - \exp[-v] \). On \( T \cap \mathcal{X} \)
\[ \epsilon^T f/n \leq \|w_{-S}\|_{\infty} \lambda_S(u) \|D_{-S} f\|_1 + \left( \sqrt{r(V^S)} + \sqrt{2v} \right) \|f\|_2/n. \]

\( \square \)

11.1 Proof of Theorem 5.1

For \( j \in D \setminus S \)
\[ \lambda_S(u) \omega_j / \sqrt{n} \leq \lambda w_j. \]
Thus by Lemma 11.3 with probability at least \( 1 - \exp[-u] - \exp[-v] \)
\[ \epsilon'(\hat{f} - f)/n \leq \lambda \|w_{-S} D_{-S}(\hat{f} - f)\|_1 + \left( \sqrt{r(V^S)} + \sqrt{2v} \right) \|\hat{f} - f\|_2/n \]
\[ = \lambda \|w_{-S} D_{-S}(\hat{f} - f)\|_1 + \left( \sqrt{r(V^S)} + \sqrt{2v} \right) \|\hat{f} - f\|_2/n. \]

Moreover
\[ \|Df\|_1 - \|D\hat{f}\|_1 = \|D_Sf\|_1 + \|D_{-S}f\|_1 - \|D_S\hat{f}\|_1 - \|D_{-S}\hat{f}\|_1 \leq \|D_S(\hat{f} - f)\|_1 + \|D_{-S}f\|_1 - \|D_{-S}\hat{f}\|_1. \]
and
\[ \|D-sf\|_1 - \|D-s\hat{f}\|_1 + \|w-sD-s(\hat{f} - f)\|_1 \leq 2\|D-sf\|_1 - \|D-s(\hat{f} - f)\|_1 + 2\|D-sf\|_1. \]

Thus
\[ \|Df\|_1 - \|D\hat{f}\|_1 + \|w-sD-s(\hat{f} - f)\|_1 \leq \|D_s(\hat{f} - f)\|_1 - \|(1-w-s)D_s(\hat{f} - f)\|_1 + 2\|D-sf\|_1 \]
\[ \leq \Gamma(S, w-s)\|\hat{f} - f\|_2/\sqrt{n} + 2\|D-sf\|_1. \]

In view of Lemma 11.2 and using the definition of effective sparsity, we thus see that with probability at least \(1 - \exp[-u] - \exp[-v]\)
\[ \parallel f - f^{0}\parallel_{2}^{2}/n + \parallel f - f^{0}\parallel_{2}^{2}/n - \parallel f^{0}\parallel_{2}^{2}/n \]
\[ \leq 2\lambda\|D_s(\hat{f} - f)\|_1 - 2\lambda\|(1-w-s)D_s(\hat{f} - f)\|_1 + 4\lambda\|D-sf\|_1 \]
\[ + 2\left(\sqrt{r(V_S)} + \sqrt{2v}\right)\|\hat{f} - f\|_2/n \]
\[ \leq 2\left(\sqrt{r(V_S)/n + \sqrt{2v}/n + \lambda\Gamma(S, w-S)}\right)\|\hat{f} - f\|_2/\sqrt{n} + 4\lambda\|D-sf\|_1 \]

The proof is completed by observing that
\[ 2\left(\sqrt{r(V_S)/n + \sqrt{2v}/n + \lambda\Gamma(S, w-S)}\right)\|\hat{f} - f\|_2/\sqrt{n} \]
\[ \leq \left(\sqrt{r(V_S)/n + \sqrt{2v}/n + \lambda\Gamma(S, w-S)}\right)^2 + \|\hat{f} - f\|_2^2/n. \]

\[ \square \]

11.2 Proof of Theorem 9.1.

We apply Lemma 11.3. Let \(S := \{t_1, \ldots, t_s\}\) where
\[ t_0 := N, \ t_1 - t_0 = \cdots = t_{s-1} - t_{s-1} = t_{s+1} - t_s. \]

Define
\[ d_{\infty} := \max_{1 \leq k \leq s+1} (t_k - t_{k-1}). \]

Then
\[ d_{\infty} \leq \frac{n}{s}. \]

Lemma 11.3 tells us that with probability at least \(1 - \exp[-u] - \exp[-v]\)
\[ \epsilon'(\hat{f} - f)/n \leq \frac{\|w-s\|_\infty \lambda_S(u)}{\sqrt{n}} \|D-s(\hat{f} - f)\|_1 + \left(\sqrt{N(s + 1) + \sqrt{2v}}\right)\|\hat{f} - f\|_2/\sqrt{n}. \]

(6)
Arguing as in Subsection 8.2 we see that
\[ \| \omega - S \|_\infty \leq d_n^2 \]
where in this case \( d_n \leq n/s \). Moreover,
\[ \| D - \bar{S} (\hat{f} - f) \|_1 \leq \| D (\hat{f} - f) \|_1 \leq \| D \hat{f} \|_1 + \| D f \|_1. \]
Finally,
\[ 2(\sqrt{N(s + 1)} + \sqrt{2v}) \| \hat{f} - f \|_2 / n \leq (\sqrt{N(s + 1)/n} + \sqrt{2v/n})^2 + \| \hat{f} - f \|_2^2 / n. \]
Inserting these three bounds into (6) one arrives at
\[ 2\epsilon (\hat{f} - f) / n \leq \| \hat{f} - f \|_2^2 / n + 2(n/s)^{2N-1} \lambda_S(u) \| Df \|_1 / \sqrt{n} + 2(n/s)^{2N-1} \lambda_S(u) \| D \hat{f} \|_1 / \sqrt{n} + (\sqrt{N(s + 1)/n} + \sqrt{2v/n})^2. \]
with probability at least \( 1 - \exp[-u] - \exp[-v] \). By assumption
\[ (n/s)^{2N-1} \lambda_S(u) / \sqrt{n} \leq \lambda. \]
Lemma 11.2 completes the proof. \( \square \)

References

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007.

A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering, 2017. arXiv:1702.05113.

F. Ortelli and S. van de Geer. On the total variation regularized estimator over a class of tree graphs. *Electronic Journal of Statistics*, 12(2):4517–4570, 2018.

F. Ortelli and S. van de Geer. Oracle inequalities for square root analysis estimators with application to total variation penalties, 2019. arXiv:1902.11192.
L. I. Rudin and E. Osher, S. and Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

V. Sadhanala and R. J. Tibshirani. Additive models with trend filtering, 2017. arXiv:1702.05037.

G. Steidl, S. Didas, and J. Neumann. Splines in higher order TV regularization. *International Journal of Computer Vision*, 70(3):241–255, 2006.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.