Adaptive Risk Bounds in Univariate Total Variation Denoising and Trend Filtering

Adityanand Guntuboyina*, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen†

423 Evans Hall
Berkeley, CA 94720
e-mail: aditya@stat.berkeley.edu

367 Evans Hall
Berkeley, CA 94720
e-mail: dlieu333@berkeley.edu

117 Illini Hall
725 S. Wright St. M/C 374
Champaign, IL 61820
e-mail: sci1706@illinois.edu

1255 Amsterdam Avenue
New York, NY 10027
e-mail: bodhi@stat.columbia.edu

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

Keywords and phrases: Adaptive splines, discrete splines, fat shattering, higher order total variation regularization, metric entropy bounds, nonparametric function estimation, risk bounds, subdifferential, tangent cone.

*Supported by NSF CAREER Grant DMS-1654589
†Supported by NSF Grant DMS-1712822
1. Introduction

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

\[ Y_i = f^*(i/n) + \xi_i, \quad i = 1, \ldots, n, \]

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

\[ Y = \theta^* + \xi, \]

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

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

\[ \hat{\theta}_V^{(r)} := \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Y - \theta\|^2 : \|D^{(r)} \theta\|_1 \leq V n^{1-r} \right\} \]  

(3)

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

Alternatively, the trend filtering estimator in the penalized form is

\[ \hat{\theta}_\lambda^{(r)} := \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Y - \theta\|^2 + \sigma n^{r-1} \lambda \|D^{(r)} \theta\|_1 \right\} \]  

(4)

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

For \( r = 1 \), (4) reduces to the one-dimensional discrete version of total variation regularization or total variation denoising which was first proposed by Rudin, Osher and Fatemi.
and has since been heavily used in the image processing community. The penalized estimator (4), for general \( r \geq 1 \), was first proposed by Steidl, Didas and Neumann [44] in the image processing literature who termed it higher order total variation regularization. The same estimator was later rediscovered by Kim et al. [25] who coined the name trend filtering for it. Many properties of the estimator have been studied in Tibshirani [46] and Wang, Smola and Tibshirani [51]. It should also be mentioned here that a continuous version of (4), where the discrete differences are replaced by continuous derivatives, was proposed much earlier in the statistics literature by Mammen and van de Geer [31] under the name locally adaptive regression splines.

The presence of the \( L^1 \) norm in the constraint in (3) (resp. penalty in (4)) promotes sparsity of the vector \( D^{(r)} \hat{\theta}^{(r)}_V \) (resp. \( D^{(r)} \hat{\theta}^{(r)}_\lambda \)). Now for every vector \( \theta \in \mathbb{R}^n \), \( \|D^{(r)} \theta\|_0 = k \) if and only if \( \theta \) equals \((f(1/n), \ldots, f(n/n))\) for a discrete spline function \( f \) that is made of \( k + 1 \) polynomials each of degree \((r - 1)\) (here \( \|x\|_0 \) denotes the number of entries of the vector \( x \) that are non-zero). Discrete splines are piecewise polynomials with regularity at the knots. They differ from the usual (continuous) splines in the form of the regularity condition at the knots: for splines, the regularity condition translates to (higher order) derivatives of adjacent polynomials agreeing at the knots, while for discrete splines it translates to discrete differences of adjacent polynomials agreeing at the knots; see Mangasarian and Schumaker [32] for details. This fact about the connection between \( \|D^{(r)} \theta\|_0 \) and discrete splines is standard (see e.g., Steidl, Didas and Neumann [44]) but we included a proof in Subsection D.3 for the convenience of the reader.

Thus the presence of the \( L^1 \) norm in (3) (resp. penalty in (4)) implies that \( \hat{\theta}^{(r)}_V \) (resp. \( \hat{\theta}^{(r)}_\lambda \)) can be written as \((\hat{f}(1/n), \ldots, \hat{f}(n/n))\) for a discrete spline \( \hat{f} \) of degree \((r - 1)\) made up of not too many polynomial pieces. Trend filtering thus presents a way of fitting (discrete) splines to the data. Note that the knots of the discrete splines are automatically chosen by the optimization algorithms underlying (3) and (4) without any input from the user (except for the value of the tuning parameter \( V \) or \( \lambda \)). Because of this automatic selection of the knots, trend filtering can be regarded as a spatially adaptive method (in the terminology of Donoho and Johnstone [9]). Note that such spatial adaptation is not exhibited by classical nonparametric regression methods such as local polynomials, kernels and splines, with a fixed tuning parameter. On the other hand, methods such as CART (Breiman et al. [4]), MARS (Friedman [14]), variable-bandwidth kernel/spline methods (see e.g., Brockmann, Gasser and Herrmann [5], Müller and Stadmüller [33], Pintore, Speckman and Holmes [36] and Zhou and Shen [54]) and wavelets (Donoho and Johnstone [9]) are also spatially adaptive.

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

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

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

It is natural to study the risk properties of (3) and (4) under the following two kinds of assumptions on $\theta^*$: (a) $n^{-1} \|D^{(r)} \theta^*\|_1 \leq V$ for some $V > 0$ (possibly dependent on $n$), and (b) $\|D^{(r)} \theta^*\|_0 \leq k$ for some $k$ that is much smaller than $n$. We shall refer to these two regimes as weak sparsity and strong sparsity respectively. This breakdown into weak and strong sparsity settings is inspired by corresponding terminology in the study of risk properties of thresholding based estimators in Gaussian sequence models [24] and the prediction risk properties of the LASSO estimators in regression [6]. Indeed, as demonstrated in Tibshirani [46], there is a close connection between the trend filtering estimators and LASSO (more details are provided in Subsection 5.4).

A thorough study on the performance of the penalized trend filtering estimator (4) under weak sparsity has been done by Tibshirani [46] and Wang, Smola and Tibshirani [51] building on earlier results of Mammen and van de Geer [31]. It is proved there that, when the tuning parameter $\lambda$ is appropriately chosen, the penalized estimator (4) is minimax optimal in the weak sparsity setting. Actually, the weak sparsity results of [46, 51] are broader and hold under more general settings (see Remark 2.1 for more details).

The present paper focuses on the strong sparsity setting. Compared to available results in the weak sparsity setting, relatively little is known about the performance of the trend filtering estimators in the strong sparsity setting. In fact, all existing results [8, 21, 29, 30, 34, 48] for strong sparsity deal with the case $r = 1$ (where trend filtering is the same as total variation denoising). To the best of our knowledge, the present paper is the first to prove risk bounds for trend filtering under strong sparsity for arbitrary $r \geq 1$. We also improve, in certain aspects, existing results for $r = 1$.

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

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

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

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

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

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

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

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

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

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

The implication of our results is the following. As mentioned earlier, the trend filtering estimators are given by discrete spline functions of degree $r-1$. The knots of these splines are chosen automatically by the algorithm (the user only needs to specify the tuning parameter $V$ or $\lambda$). Our results indicate that under the assumption $\|D^{(r)}\theta^*\|_0 = k$ (i.e., $\theta^*$ is a discrete spline of degree $r-1$ with $k+1$ polynomial pieces) with a minimum length condition on the polynomial pieces of $\theta^*$, the automatic selection of knots by the trend filtering estimators (when appropriate choices of $V$ or $\lambda$) happens in a way that the overall risk is comparable to the oracle risk of $r\sigma^2(k+1)/n$. In fact, when $k = O(1)$, the risks of the ideally tuned trend filtering estimators is only off compared to the oracle risk by a factor that is logarithmic in $n$ (we also prove in Lemma 2.4 that this logarithmic factor cannot be completely removed in general). The automatic knot selection of trend filtering can therefore be interpreted as being done adaptively depending on the structure of the unknown $\theta^*$ in order to approximate the oracle risk. This is the reason why we refer to our results as adaptive risk bounds. It should be mentioned here that a similar adaptation story can also be used to describe the weak sparsity results [46, 51] where the knots are adaptively chosen to attain the minimax rate under the $L^1$ constraint on $D^{(r)}\theta^*$. Therefore, our results (together with those of [46, 51]) provide support for the use of the trend filtering estimators in both weak and strong sparsity settings.

We would like to mention here that theoretical analysis of spatially adaptive nonparametric regression methods under strong sparsity is non-trivial. Indeed, among various
such methods including CART, MARS, variable-bandwidth kernel/spline methods and wavelets, rigorous theoretical risk results under strong sparsity only exist for wavelets [9] and variable-bandwidth kernel methods [17, 28]. The analysis of trend filtering estimators is more involved compared to estimators based on wavelets and variable-bandwidth kernels because the trend filtering estimators are given by the output of an optimization algorithm and have no closed form expressions.

The rest of this paper is organized as follows. Our main results are described in Section 2: Subsection 2.1 deals with the constrained estimator where we provide risk bounds under both weak sparsity (which was not known previously) and strong sparsity. Subsection 2.2 deals with the penalized estimator and here we separate our presentation into two parts: results for \( r = 1 \) and results for \( r \geq 2 \); our results for \( r \geq 2 \) are weaker (there is an additional \((k + 1)^{2r}/n\) term in the risk) than the results for \( r = 1 \). Throughout, we focus on nonasymptotic upper bounds for the risk (expected loss) although all our results can be converted into high probability upper bounds on the loss (see Remark 2.3). All proofs are given in the supplementary material at the end of the paper and a high level overview of the proofs is provided in Section 3. Section 4 contains some simulation studies supporting some of our theoretical results. Finally several interesting issues related to our results are described in Section 5.

2. Main Results

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

2.1. Results for the Constrained Estimator

We start with the bound of \( n^{-2r/(2r+1)} \) for risk of \( \hat{\theta}_V^{(r)} \) under the condition that the tuning parameter \( V \) satisfies \( \|D^{(r)}\theta^*\|_1 \leq Vn^{1-r} \). This result is similar to results in Mammen and van de Geer [31], Tibshirani [46] and Wang, Smola and Tibshirani [51] who focussed on the penalized estimator (4) (see Remark 2.1 for details). We also explicitly state the dependence of the bound on \( V \) and \( \sigma \).

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

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

Also for every \( x > 0 \), we have

\[
\frac{1}{n} \|\hat{\theta}_V^{(r)} - \theta^*\|^2 \leq C_r \max \left( \left( \frac{\sigma^2 V^{1/r}}{n} \right)^{2r/(2r+1)}, \frac{\sigma^2}{n} \log(en) \right) + \frac{4\sigma^2 x}{n}
\]
with probability at least $1 - e^{-x}$.

**Remark 2.1.** As mentioned earlier, bounds similar to (8) and (9) have been proved in Mammen and van de Geer [31], Tibshirani [46] and Wang, Smola and Tibshirani [51] for the penalized trend filtering estimator. Actually, the bounds in these earlier papers hold under more general assumptions than the assumptions of the current paper. For example, their analyses also holds under the assumption that the (continuous) variation norm of the function $(f^*)^{(r-1)}$ (this is the $(r-1)^{th}$ derivative of $f^*$) is at most $V$, where $f^*$ is the true function with $\theta^* = (f^*(1/n), \ldots, f^*(1))$. Note that there is subtle difference between this and our assumption of an upper bound on $\|D^{(r)}\theta^*\|_1$ in the sequence model (2). An assumption on the variation norm of $(f^*)^{(r-1)}$ does not directly lead to a bound on $\|D^{(r)}\theta^*\|_1$ which makes the analysis difficult (see Wang, Smola and Tibshirani [51] for more details on the relation between the two variation norms). Also, the results in these earlier papers studied the general setting with $\theta^* := (f^*(x_1), \ldots, f^*(x_n))$ where $x_1, \ldots, x_n$ are design points that are not necessarily equally spaced. We restrict ourselves to the equally spaced design setting in this paper (see Subsection 5.1).

**Remark 2.2.** $n^{-2r/(2r+1)}$ is the minimax rate of estimation over the class of $\theta \in \mathbb{R}^n$ with $\|D^{(r)}\theta\|_1 \leq Vn^{1-r}$ (see e.g., Donoho and Johnstone [10]). This means that the constrained trend filtering estimator with tuning parameter $V$ is minimax optimal over \{ $\theta \in \mathbb{R}^n$ : $\|D^{(r)}\theta\|_1 \leq Vn^{1-r}$ \}. This result was known previously for the penalized estimator; see Tibshirani [46]. Note also that $V$ here can change with $n$ as well and inequality (8) implies that $\hat{\theta}^{(r)}_V$ is minimax optimal even in terms of the dependence of the rate on $V$. Before we state results for strong sparsity, we need some notation. Fix an integer $r \geq 1$ and let $n \geq r + 1$. For a vector $\theta \in \mathbb{R}^n$ and an index $2 \leq j \leq n - r + 1$, we say that $j$ is an $r^{th}$ order knot (or knot of order $r$) of $\theta$ provided $(D^{(r-1)}\theta)_{j-1} \neq (D^{(r-1)}\theta)_j$. Note that first order knots are just jumps and second order knots are points of change of slope. We also say that an $r^{th}$ order knot $j$ has sign $+1$ if $(D^{(r-1)}\theta)_{j-1} < (D^{(r-1)}\theta)_j$ and sign $-1$ if $(D^{(r-1)}\theta)_{j-1} > (D^{(r-1)}\theta)_j$. For $\theta \in \mathbb{R}^n$, we let

$$k_r(\theta) := \|D^{(r)}\theta\|_0 \quad \text{and} \quad V^{(r)}(\theta) := n^{r-1}\|D^{(r)}\theta\|_1.$$  

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

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

Suppose $k_r(\theta) = k$ and let $2 \leq j_1 < \cdots < j_k \leq n - r + 1$ denote all the $r^{th}$ order knots of $\theta$ with associated signs $\tau_1, \ldots, \tau_k \in \{-1, 1\}$. Also let $\tau_0 = \tau_{k+1} = 0$. Further, let $n_0 := j_1 + r - 2$, $n_i := j_{i+1} - j_i$, for $1 \leq i \leq k - 1$, and $n_k := n - r + 2 - j_k$, and observe
that \( \sum_{i=0}^{k} n_i = n \). Finally, let

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

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

\[
\delta_r(\theta) := \left( r_0^{1-2r} + n_k^{1-2r} + \sum_{i=1}^{k-1} n_i^{1-2r} I\{r_i \neq r_{i+1}\} \right)^{1/2}
\]

and

\[
\Delta_r(\theta) := \frac{k + 1}{n} \log \frac{cn}{k+1} + \frac{\delta_r^2(\theta)}{n} \left( \frac{n}{k+1} \right)^{2r-1} \log \frac{en}{k+1} + \left( \frac{\delta_r(\theta)}{\sqrt{n}} \right)^{1/r}
\]

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

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

\[
\min_{0 \leq i \leq k; r_i \neq r_{i+1}} n_i \geq \frac{cn}{k+1}
\]

for a positive constant \( c \leq 1 \) (here \( r_1, \ldots, r_k \in \{-1, 1\} \) are the signs of the \( r \)th order knots of \( \theta \) while \( r_0 \) and \( r_{k+1} \) are taken to be zero), then

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

and consequently

\[
\Delta_r(\theta) \leq \left\{ 1 + c^{1-2r} \right\} \frac{k + 1}{n} \log \frac{en}{k+1} + \frac{c^{(1-2r)/(2r)} k + 1}{n}
\]

\[
\leq \left\{ 1 + c^{1-2r} + c^{(1-2r)/(2r)} \right\} \frac{k + 1}{n} \log \frac{en}{k+1}.
\]

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

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

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

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

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

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

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

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

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

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

\[
R(\hat{\theta}_V^{(r)}, \theta^*) \leq \left( V - V^{(r)}(\theta^*) \right)^2 + C_r(c) \frac{\sigma^2 \left(k_r(\theta^*) + 1\right)}{n} \log \frac{en}{k_r(\theta^*) + 1} \tag{17}
\]

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

\[
\left( V - V^{(r)}(\theta^*) \right)^2 \leq C \frac{\sigma^2 \left(k_r(\theta^*) + 1\right)}{n} \log \frac{en}{k_r(\theta^*) + 1}
\]

for a positive constant \( C \), then we have

\[
R(\hat{\theta}_V^{(r)}, \theta^*) \leq C_r(c,C) \frac{\sigma^2 \left(k_r(\theta^*) + 1\right)}{n} \log \frac{en}{k_r(\theta^*) + 1} \tag{18}
\]

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

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

Remark 2.4 (On the minimum length condition). The minimum length condition (13) required for Corollary 2.3 is weaker than existing minimum length conditions in the literature (this comparison is only for \( r = 1 \) because no results exist for \( r \geq 2 \)) which are all
of the form
\[
\min_{0 \leq i \leq k} n_i \geq \frac{cn}{k + 1} \quad \text{where } k = k_1(\theta^*).
\] (19)
Indeed our condition (13) requires that \(n_i \geq cn/(k + 1)\) be true only for those \(i\) for which \(r_i \neq r_{i+1}\) while (19) requires this for all \(i\). To see why our condition can be substantially weaker, consider, for example, the situation when \(D(\theta^*)\) is a monotonic vector (for \(r = 1\), this means that \(\theta^*\) is itself monotone while for \(r = 2\), this means that \(\theta^*\) is convex/concave). In this case, condition (13) is equivalent to requiring that \(n_i \geq cn/(k + 1)\) only for \(i = 0\) and \(i = k\) which is much weaker than requiring it for all \(0 \leq i \leq k\).

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

![A continuous piecewise Linear Function](image)

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

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

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

(20)

and consider estimating $\theta^*$ from an observation $Y \sim N_n(\theta^*, I_n)$ (i.e., $\sigma = 1$) by $\hat{\theta}^{(1)}_V$ (i.e., $r = 1$) with tuning parameter $V = V^{(1)}(\theta^*) = 5$. It is clear here that $k_1(\theta^*) = 1$. The minimum length condition (13) is not satisfied because $n_0 = n - 1$ and $n_1 = 1$. The risk $R(\hat{\theta}^{(1)}_V, \theta^*)$ can be computed via simulation. In Figure 2 (left panel), we have plotted $\log R(\hat{\theta}^{(1)}_V, \theta^*)$ against $\log n$ for values of $n$ between 1000 and 5000 (chosen to be equally spaced on the log-scale). For each value of $n$, we calculated the risk using 100 Monte Carlo replications. The slope of the least squares line through these points turned out to be close to $-2/3$ which indicates that the risk $R(\hat{\theta}^{(1)}_V, \theta^*)$ decays at the rate $n^{-2/3}$. This rate is slower than the rate given by Corollary 2.3 indicating that inequality (17) is not true for this $\theta^*$. On the other hand, the $n^{-2/3}$ rate here makes sense in light of Theorem 2.1. Therefore, even though the vector $D\theta^*$ is sparse (with $\|D\theta^*\|_0 = 1$), the rate of convergence of $\hat{\theta}^{(1)}$ is equal to the $n^{-2/3}$ and not the faster rate given by Corollary 2.3. This points to the necessity of the minimum length condition (13).

![Fig 2: Left](image1)

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

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

$$\theta_1^* = \cdots = \theta_{[n/2]}^* = 0 \quad \text{and} \quad \theta_{[n/2]+1}^* = \theta_{[n/2]+2}^* = \cdots = \theta_n^* = 5.$$  

(21)

Here consider the problem of estimating $\theta^*$ by the estimator $\hat{\theta}^{(2)}_V$ (i.e., $r = 2$) with tuning parameter $V = V^{(2)}(\theta^*) = 10n$. It is clear that $k_2(\theta^*) = 2$, $n_0 = \lfloor n/2 \rfloor$, $n_1 = 1$ and $n_2 = n - \lfloor n/2 \rfloor - 1$. The minimum length condition (13) is not satisfied as $n_1$ is too small. The risk $\log R(\hat{\theta}^{(2)}_V, \theta^*)$ is plotted against $\log n$ in the right panel of Figure 2 (the values...
of $n$ are chosen as before). The slope of the least squares line here is close to $-2/5$ which suggests that the risk decays slowly than what is given by Corollary 2.3. Note that $n^{-2/5}$ is exactly the rate given by Theorem 2.1 (take $r = 2$ and $V = 10n$ in (8)).

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

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

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

### 2.2. Results for the Penalized Estimator

In this section, we present risk results for the penalized estimator defined in (4). An important role in these results will be played by the subdifferential of the convex function $f(\theta) := \|D^r\theta\|_1$ at the true parameter value $\theta^*$. Recall that the subdifferential of a convex function $g : \mathbb{R}^n \to \mathbb{R}$ at a point $\theta \in \mathbb{R}^n$ is the set consisting of all subgradients of $g$ at $\theta$ and will be denoted by $\partial g(\theta)$. For every finite convex function $g$ on $\mathbb{R}^n$ and $\theta \in \mathbb{R}^n$, the subdifferential $\partial g(\theta)$ is non-empty, closed, convex and bounded (see, for example, Rockafellar [39, Page 218]).

The following is the reason why $\partial f(\theta^*)$ (for $f(\theta) := \|D^r\theta\|_1$) plays a key role in understanding the risk of (4). It has been proved by Oymak and Hassibi [35, Theorem 2.2] that for a general penalized estimator:

$$\hat{\theta}_\lambda^\circ := \arg\min_{\theta \in \mathbb{R}^n} \left( \frac{1}{2} \|Y - \theta\|^2 + \sigma \lambda g(\theta) \right)$$

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

$$R(\hat{\theta}_\lambda^\circ, \theta^*) \leq \frac{\sigma^2}{n} \mathbb{E} \left( \inf_{v \in \lambda \partial g(\theta^*)} \|Z - v\|^2 \right)$$

(22)

where $\lambda \partial g(\theta^*) := \{\lambda v : v \in \partial g(\theta^*)\}$ and the expectation on the right hand side is with respect to the standard Gaussian vector $Z \sim N_n(0, I_n)$. Moreover, inequality (22) cannot in general be improved, because, as proved in [35, Proposition 4.2], it is tight in the low $\sigma$ limit, i.e., the limit (as $\sigma \to 0$) of the left hand side of (22) scaled by $\sigma^2/n$ equals the expectation on the right hand side of (22). Inequality (22) will be our main technical tool for studying the risk of (4) and thus it will be important to understand the subdifferentials of the function $\theta \mapsto \|D^r\theta\|_1$.

The next result (proved in Subsection C.4) characterizes the subdifferential of $f(\theta) := \|D^r\theta\|_1$.  

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

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


Proposition 2.5. Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\alpha) := \|D^{(r)}\alpha\|_1$. Fix $\theta \in \mathbb{R}^n$. Then $\partial f(\theta)$ consists of vectors $v \in \mathbb{R}^n$ such that

$$\sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r} \right) v_i = 0 \quad \text{for } 1 \leq j \leq r,$$

and

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

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

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

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

$$v^* := \arg\min_{v \in \partial f(\theta^*)} \|v\| \quad \text{and} \quad v_0 := \arg\min_{v \in \text{aff}(\partial f(\theta^*))} \|v\|$$

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

$$\sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r} \right) v_i = \text{sgn}((D^{(r)}\theta^*)_{j-r})$$

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

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

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

where $\lambda_{\theta^*}(z)$ is the value of $\lambda$ which minimizes the distance of the vector $z$ from the set $\lambda \partial f(\theta^*)$. Lemma B.5 proves that $\lambda_{\theta^*}(z)$ is uniquely defined for each $z \in \mathbb{R}^n$ (under the assumption that $D^{(r)}\theta^* \neq 0$) and also that $\mathbb{E}\lambda_{\theta^*}(Z) < \infty$ where the expectation is taken with respect to $Z \sim N_n(0, I_n)$. We are now ready to state our first result on the risk of the penalized trend filtering estimators (recall $\Delta_r(\theta)$ from (12)).
Theorem 2.6. Fix \( r \geq 1 \) and suppose \( \theta^* \in \mathbb{R}^n \) with \( D^{(r)}\theta^* \neq 0 \). Let
\[
\lambda^* := n^{1-r} \left( \mathbb{E}_{\theta^*} \|Z\| + \frac{2}{\|v_0\|} \right)
\] (27)
where the expectation is taken with respect to the standard Gaussian vector \( Z \sim N_n(0,I_n) \). Then for every regularization parameter \( \lambda \geq \lambda^* \), we have
\[
R(\hat{\theta}_\lambda^{(r)}, \theta^*) \leq C_r \sigma^2 \Delta_r(\theta^*) + \frac{64 \sigma^2}{n} \|v^*\|^2 + \frac{4 \sigma^2}{n^{3-2r}} (\lambda - \lambda^*)^2 \|v^*\|^2
\] (28)
for a constant \( C_r \) that only depends on \( r \).

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

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

1. Suppose \( r = 1 \). Then \( v_0 = v^* \). Further suppose that \( \theta^* \) has \( k \geq 1 \) jumps (first order knots) with signs \( r_1, \ldots, r_k \) and let \( n_0, n_1, \ldots, n_k \) denote the lengths of the constant pieces of \( \theta^* \). Then
\[
\|v_0\|^2 = \|v^*\|^2 = \frac{1}{n_0} + \frac{1}{n_k} + 4 \sum_{i=1}^{k-1} \frac{I\{r_i \neq r_{i+1}\}}{n_i}.
\] (29)

2. For \( r \geq 2 \), we have
\[
\|v_0\| \geq \frac{(r-1)!}{(r+1)2^{r-1}} n^{-r+1/2}.
\] (30)

3. Suppose \( r \geq 2 \) and \( \theta^* \) satisfies the minimum length condition (13) with constant \( c \), then
\[
\|v^*\| \leq C_r c^{-r+1/2}(k+1)^r n^{-r+1/2}
\] (31)
where \( C_r \) is a constant depending only on \( r \).

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

Corollary 2.8. Suppose \( \theta^* \in \mathbb{R}^n \) has \( k \geq 1 \) jumps with signs \( r_1, \ldots, r_k \) and suppose that \( n_0, n_1, \ldots, n_k \) denote the lengths of the constant pieces of \( \theta^* \). Then, with \( \lambda^* \) as in (27), we have
\[
R(\hat{\theta}_\lambda^{(1)}, \theta^*) \leq C \sigma^2 \left( \Delta_1(\theta^*) + \frac{(\lambda - \lambda^*)^2}{n} \sum_{i=0}^{k} \frac{I\{r_i \neq r_{i+1}\}}{n_i} \right)
\] (32)
for every $\lambda \geq \lambda^*$. Here $C$ is a universal constant. Also, we use our usual convention $r_0 = r_{k+1} = 0$.

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

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

(33)

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

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

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

(34)

where $k$ is the number of jumps of $\theta^*$, i.e., $k = k_1(\theta^*)$. Moreover, the logarithmic term above cannot be removed in general. This is due to the following reason. First, note that, for every non-random $\lambda$ possibly depending on $\lambda^*$, the penalized estimator $\hat{\theta}^{(1)}_{\lambda}$ has worse risk compared to the ideally tuned constrained estimator i.e., $\hat{\theta}^{(1)}_{V}$ with $V = V^{(r)}(\theta^*)$. This fact (which is noted and explained in Subsection 5.2), together with Lemma 2.4, implies clearly that the logarithmic factor in (34) cannot be removed in general.

**Remark 2.6 (Comparison to existing results).** Among the class of existing results for the risk of $\hat{\theta}^{(1)}_{\lambda}$, the strongest (in terms of giving the smallest bound on the risk) is due to Lin et al. [30] who proved that, when $\lambda$ is appropriately selected (depending on $\theta^*$), $\hat{\theta}^{(1)}_{\lambda}$ satisfies:

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

(35)

provided

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

(36)

for a positive constant $c$. Here $n_0, \ldots, n_k$ are the lengths of the constant pieces of $\theta^*$. This bound from Lin et al. [30] is smaller compared to an earlier result of Dalalyan, Hebiri and Lederer [8] and to a very recent result of Ortelli and van de Geer [34] (although the results of [8, 34] apply to a universal choice of the tuning parameter $\lambda$; see Remark 2.7). The bound (35) is weaker than (34) in two respects: (a) there are additional terms in (35) involving $\log n$ and $k$ compared to (34), and (b) our minimum length condition (13) is weaker than (36): (13) requires that $n_i \geq cn/(k+1)$ only for those $i$ for which $r_i \neq r_{i+1}$ while (36) requires this for all $i$.

Note that the regularization parameter $\lambda^*$ (for which the near parametric risk bound (34) holds) depends on $\theta^*$. Further, the exact nature of its dependence on $\theta^*$ is not apparent from its definition (27). In the next result, we provide a more explicit upper bound for $\lambda^*$. For this, we require a stronger length condition than (13). Note that we are still in the $r = 1$ case.
Lemma 2.9. Consider the same setting as in Corollary 2.8. Assume that the length condition:
\[
\min_{0 \leq i \leq k; t_i \neq t_{i+1}} n_i \geq \frac{c_1 n}{k + 1} \quad \text{and} \quad \max_{0 \leq i \leq k; t_i \neq t_{i+1}} n_i \leq \frac{c_2 n}{k + 1}
\]  
(37)
holds for two positive constants \( c_1 \leq 1 \) and \( c_2 \geq 1 \). Let \( \lambda^* \) be as defined in (27). Then there exists a positive constant \( C^*(c_1, c_2) \) (which depends only on \( c_1 \) and \( c_2 \)) such that
\[
\lambda^* \leq C^*(c_1, c_2) \left( \frac{n}{\sum_{i=0}^{k} I\{r_i \neq r_{i+1}\}} \log \left( \frac{en}{k+1} \right) \right).
\]  
(38)
Lemma 2.9 can be used, in conjunction with the risk bound (33) (which holds for every \( \lambda \geq \lambda^* \)) to yield the following result which provides bounds similar to (34) for explicit choices of \( \lambda \).

Corollary 2.10. Consider the same setting as in Lemma 2.9 and assume the length condition (37). Then if the regularization parameter \( \lambda \) satisfies
\[
\lambda = \Gamma \sqrt[4]{\frac{n}{\sum_{i=0}^{k} I\{r_i \neq r_{i+1}\}} \left( \log \frac{en}{k+1} \right)},
\]  
(39)
we have
\[
R(\hat{\theta}^{(1)}_{\lambda}, \theta^*) \leq C(c_1)\sigma^2 (1 + \Gamma^2) \left( \frac{k+1}{n} \log \frac{en}{k+1} \right)
\]  
(40)
for every \( \Gamma \geq C^*(c_1, c_2) \) (where \( C^*(c_1, c_2) \) is the constant given by Lemma 2.9). Also \( C(c_1) \) depends only on \( c_1 \). Also, if the regularization parameter \( \lambda \) satisfies
\[
\lambda = \Gamma \sqrt{n \log(en)},
\]  
(41)
we have
\[
R(\hat{\theta}^{(1)}_{\lambda}, \theta^*) \leq C(c_1)\sigma^2 (1 + \Gamma^2) \left( \frac{k+1}{n} \log \frac{en}{k+1} \right) \left( 1 + \Gamma^2 \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\} \right)
\]  
(42)
for every \( \Gamma \geq C^*(c_1, c_2) \).

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

Remark 2.7 (Comparison to existing results). We now compare Corollary 2.10 to existing results for the penalized estimator in Lin et al. [30], Dalalyan, Hebiri and Lederer [8] and Ortelli and van de Geer [34]. Note first that the choice (39) of \( \lambda \) depends on certain aspects of \( \theta^* \); in particular, it depends on \( k, \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\} \) and the values \( c_1 \) and \( c_2 \) in the length condition (37). The bound (35) of Lin et al. [30] holds for \( \lambda_1 = (n \min_{0 \leq i \leq k} n_i)^{1/4} \)

which also depends on the true vector $\theta^*$ through the lengths $n_1, \ldots, n_k$. If we assume that each $n_i$ is of order $n/(k+1)$, then

$$\lambda_1 \sim \sqrt{\frac{n}{k+1}}.$$  \hspace{1cm} (43)

Note that the leading term in our choice (39) of $\lambda$ as well as in $\lambda_1$ is $\sqrt{n}$. Corollary 2.10 also applies to the choice (41) for which the bound (42) holds. Note that (41) has considerably less dependence on $\theta^*$ as it only depends on the constants $c_1$ and $c_2$ appearing in the length condition (37). On the other hand, the bound (42) is weaker compared to (40). However, (42) needs to be compared to the results of Dalalyan, Hebiri and Lederer [8, Proposition 3] and Ortelli and van de Geer [34, Corollary 4.4]. Indeed, Dalalyan, Hebiri and Lederer [8] considered the choice

$$\lambda_2 := 2 \sqrt{2n \log(n/\delta)}$$  \hspace{1cm} (44)

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

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

This result has been improved slightly in the very recent paper Ortelli and van de Geer [34] (see also van de Geer [48]) where the $\log(en) \log(en/\delta)$ term in the right hand side above is replaced by $\log(en/(k+1)) \log(en/\delta)$ (i.e., one of the $\log(en)$ terms is replaced by $\log(en/(k+1))$). An expectation (risk) bound has not been proved in these two papers.

Note the the choice of $\lambda$ in (41) is similar to that of $\lambda_2$ in (44) although our choice needs $\Gamma$ to be sufficiently large while the choice $\lambda_2$ is universal (although it depends on $\delta$). On the other hand, the high probability bound implied by (42) is (see Remark 2.3) the statement that

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

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

We reiterate here that our length condition (37) involves an upper bound on $n_i$ for $r_i \neq r_{i+1}$. From an examination of the proof of Lemma 2.9, it will be clear that we will obtain a weaker upper bound for $\lambda^*$ in the sense of having additional multiplicative factors involving $k$ if this upper bound assumption on $n_i$ is removed. No such upper bound is needed for the results in Dalalyan, Hebiri and Lederer [8], Lin et al. [30], Ortelli and van de Geer [34]. On the other hand, our lower bound (and our upper bound in (37)) involves only those $i$ satisfying $r_i \neq r_{i+1}$ while the assumptions in these earlier papers required a lower bound on every $n_i$. 


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

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

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

for every \( \lambda \geq \lambda^* \). Here \( k := k_r(\theta^*) \) and \( C_r(c) \) depends only on \( c \).

Corollary 2.11 implies that when \( \theta^* \) satisfies the minimum length condition (13), then

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

for \( \lambda = \lambda^* \). (47)

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

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

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

\[
\min_{0 \leq i \leq k_r, \neq r_{i+1}} n_i \geq \frac{c_1 n}{k+1} \quad \text{and} \quad \max_{0 \leq i \leq k_r, \neq r_{i+1}} n_i \leq \frac{c_2 n}{k+1}
\]

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

\[
\lambda^* \leq C^*_r(c_1, c_2) \sqrt{n \log \left( \frac{en}{k+1} \right)}
\]

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

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

Compared to (38), the bound (49) is weaker because there is no \( \sum_{i=0}^k I\{r_i \neq r_{i+1}\} \) in the denominator in (49).
Combining Lemma 2.12 with the risk bound (46), we obtain the following result which provides bounds similar to (47) for explicit choices of $\lambda$.

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

$$\lambda = \Gamma \sqrt{n \log \left( \frac{en}{k+1} \right)},$$

we have

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

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

Further, if the regularization parameter $\lambda$ satisfies

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

we have

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

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

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

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

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

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

### 3. Proof Ideas

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

$$\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \| Y - \theta \|^2 : \theta \in K \right\}.$$
for a closed convex set $K$. Clearly $\hat{\theta}^{(r)}_V$ is a special case of this estimator when $K$ is taken to be the set $K^{(r)}(V)$ defined as

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

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

$$\mathbb{E} \sup_{\theta \in K^{(r)}(V)} \langle \xi, \theta - \theta^* \rangle$$

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

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

We then bound the metric entropy of $S_r(V, t)$ via its fat-shattering dimension (it is well known that fat-shattering dimension can be used to control metric entropy; see e.g., Rudelson and Vershynin [41]). Metric entropy and fat-shattering dimension are formally defined in Subsection C.1 and Subsection D.6 respectively. Our idea of using fat shattering to establish the metric entropy of $S_r(V, t)$ and thereby bounding (54) seems novel. Previous bounds on quantities similar to (54) in the context of trend filtering used eigenvector incoherence (see, for example, Wang et al. [52]) and the ideas here are quite different from our methods.

To prove the strong sparsity risk bound, Theorem 2.2, we use another strand of results from the general theory of convex-constrained least squares estimators. Specifically, a result from Oymak and Hassibi [35] implies that the risk of $\hat{\theta}^{(r)}_V$ at $V = V^* := V^{(r)}(\theta^*)$ can be obtained by controlling the Gaussian width of the tangent cone of the convex set $K^{(r)}(V^*)$ at $\theta^*$. These general results, along with the definitions of tangent cones and Gaussian width, are again recalled in Subsection A. Understanding the tangent cone to $K^{(r)}(V^*)$ at $\theta^*$ then becomes key to proving Theorem 2.2.

We provide a precise characterization of the tangent cones of $K^{(r)}(V^*)$ in Lemma C.3. These tangent cones have a complicated structure (especially for $r \geq 2$) and calculating their Gaussian width is non-trivial. Our idea behind these calculations is the fact (proved in Lemma B.2) that, under a unit norm constraint, every vector $\alpha$ in the tangent cone of $K^{(r)}(V^*)$ at $\theta^*$ is nearly made up of two $(r - 1)^{th}$ order convex/concave sequences in each polynomial part of $\theta^*$ (note that a sequence $\theta \in \mathbb{R}^n$ is said to be $(r - 1)^{th}$ order convex/concave if the vector $D^{(r-1)}\theta$ is monotone; see e.g., Kuczma [26]). The special case of this observation for $r = 1$ implies that every vector $\alpha$ with $\|\alpha\| \leq 1$ in the tangent cone to $K^{(1)}(V^*)$ at $\theta^*$ is nearly made up of two monotonic sequences in each constant piece of $\theta^*$. For $r = 2$, it means that every vector $\alpha$ with $\|\alpha\| \leq 1$ in the tangent cone to $K^{(2)}(V^*)$ at $\theta^*$ is nearly made up of two convex/concave sequences in each linear piece of $\theta^*$. 


The above observation allows us to compute the Gaussian width of these tangent cones using metric entropy results (established again via connections between metric entropy and fat shattering) and also available results (from Bellec [3]) on the Gaussian widths of shape constrained cones. The set of all \((r - 1)^{th}\) order convex sequences in \(\mathbb{R}^n\) forms a convex cone in \(\mathbb{R}^n\) and these cones have been studied in the literature on shape constrained estimation.

For \(r = 1\), the above idea bears strong similarities with the method employed in Lin et al. [30] for studying the penalized estimator (4) for \(r = 1\). In this paper, they use the key observation that for appropriate \(\lambda\), the vector \((I - P_0)(\hat{\theta}_\lambda^{(1)} - \theta^*\)) is well-approximated by a vector which is made of two monotonic sequences in each constant piece of \(\theta^*\). Here \(P_0\) is the projection matrix onto the piecewise constant structure determined by \(\theta^*\) and \(I\) is the identity matrix. This idea is similar in spirit to our observation on the tangent cone of \(K^{(1)}(V^*)\) at \(\theta^*\). The details differ though as we are working with the vectors in the tangent cone while Lin et al. [30] focus on a functional of \(\hat{\theta}_\lambda^{(1)} - \theta^*\) (note though that if \(\hat{\theta}\) has variation \(\leq V^*\), then \(\hat{\theta} - \theta^*\) does indeed belong to the tangent cone). Also our method for dealing with the Gaussian width of the set of these piecewise monotonic vectors is sharper than the analysis of Lin et al. [30] and our analysis also extends to every \(r \geq 2\).

The results in Subsection 2.2 for the penalized estimator are all based on (22). We use the precise characterization of the subdifferential of the penalty function \(\theta \mapsto \|D^{(r)}\theta\|_1\) given in Proposition 2.5 to control the right side of (22). Our idea here is to relate the right side of (22) to the risk of the constrained estimator (we use and extend ideas from Foygel and Mackey [13] for this). This allows us to derive risk results for the penalized trend filtering estimator as a corollary to our results for the constrained estimator.

4. Simulations

In this section, we present numerical evidence for our theoretical results. We generate data from a piecewise constant function \(f^*_1\) and a continuous piecewise affine function \(f^*_2\) on \([0, 1]\) and evaluate the performance of the trend filtering estimators for \(r = 1\) (total variation denoising) and \(r = 2\) respectively. The functions \(f^*_1\) and \(f^*_2\) (see Figure 3) are given by

\[
f^*_1(x) := 2I_{(0.2,0.4]}(x) + 4I_{(0.4,0.6]}(x) + I_{(0.6,0.8]}(x) + 4I_{(0.8,1]}(x)
\]

and

\[
f^*_2(x) := -44 \max(x - 0.25, 0) + 48 \max(x - 0.5, 0) - 56 \max(x - 0.75, 0) + 28x.
\]

The function \(f^*_1\) was used in the simulation study of Lin et al. [30]. In addition to these functions, we also performed a simulation study on another piecewise constant function \(f^*_3\) which is similar to the blocks function of Donoho and Johnstone [9]; results for \(f^*_3\) are in Section E.

From \(f^*_1\) and a value of \(n\) (chosen from a grid of size 30 between 100 and 10000; the grid being equally spaced on the logarithmic scale) we generated an \(n \times 1\) observation vector.
Y \sim N_n(\theta^*, I_n) where \theta^* is the vector obtained by sampling \( f_1^* \) at \( n \) equally spaced points with end-points 0 and 1. We then computed the following six estimators on the data vector \( Y \): (a) the ideal constrained estimator (3) with \( V = V^* = \| D\theta^* \|_1 \), (b) the ideal penalized estimator (4) with \( \lambda = \lambda^* \) (as defined in (27)), (c) two cross-validation (CV) based estimators, (d) the penalized estimator (4) with \( \lambda \) of the form (39) with \( \Gamma = 1 \), and (e) the penalized estimator (4) with \( \lambda \) of the form (41) with \( \Gamma = 0.5 \). Corollary 2.10 proves that the risk with these \( \lambda \) choices decays as \( (\log n)/n \) (ignoring terms involving \( k \)) provided \( \Gamma \) is taken to be a large enough constant. In our simulations for \( f_1^* \), we found that \( \Gamma = 1 \) in (39) and \( \Gamma = 0.5 \) in (41) were large enough to yield the desired performance. Higher values of \( \Gamma \) led to similar rates of decay of the risk with \( n \) (even though the risk itself seemed to become larger with \( \Gamma \)).

Here are some details behind the computation of these estimates. The constrained estimator was computed by the convex optimization software MOSEK (via the R package Rmosek). The penalized estimators were computed via the R package tvd for total variation denoising. The computation of the ideal penalized estimator requires computing the value of \( \lambda^* \) and, for this, we need to compute \( \mathbb{E} \lambda_{\theta^*} (Z) \) (where \( Z \sim N_n(0, I_n) \)) and \( 2/\|v_0\| \) (see (27)). \( 2/\|v_0\| \) was calculated by the formula (29). For \( \mathbb{E} \lambda_{\theta^*} (Z) \), we used the fact that \( \lambda_{\theta^*} (z) \) can be calculated by convex optimization for each \( z \in \mathbb{R}^n \) which implies that the expectation can be computed by Monte-Carlo averaging. More details behind this are provided in Section E. The CV estimators were calculated using the R package genlasso which provides two penalized estimates based on CV: one based on choosing \( \lambda \) so as to minimize the CV error (\( CV_1 \)) and the other based on choosing \( \lambda \) via the one standard error rule (\( CV_2 \)).

For each data set, we computed the value of the loss \( \|\hat{\theta} - \theta^*\|^2/n \) for each of these six estimates. We generated 600 replications of the data for each value of \( n \) to compute the average value of the loss which is an approximation of the risk of each estimator. Our results are provided in Figure 4. The top-left plot shows the different values of \( \lambda \) employed by the estimators based on (4). Here we plotted the \( \lambda^* \) values as well as those corresponding to (39) with \( \Gamma = 1 \) (penalty one) and (41) with \( \Gamma = 0.5 \) (penalty two).
Fig 4: Plots when the true function is \( f_1^* \). The top-left plot shows the \( \lambda^* \) values, the CV \( \lambda \) values (median and the first and third quartiles over 200 replications) and the values corresponding to the explicit penalties (39) with \( \Gamma = 1 \) and (41) with \( \Gamma = 0.5 \). The other three figures show the behavior of the risk as a function of \( n \). In the last two plots, the legend shows the value of \( R^2 \) and the slope respectively for the curves corresponding to each estimator.

In addition, we also plotted here the penalty levels chosen by the CV estimators. These are random so we plotted their median and quartile values over the 600 replications. The remaining three plots in Figure 4 show the risks of the six estimators. In the top-right plot, the risk is simply plotted as a function of \( n \) (from our theoretical results, the risk is supposed to decay like the curve \( n \mapsto (t_1/n) \log(t_2n) \) for two constants \( t_1 \) and \( t_2 \)). In the bottom-left plot, we plotted \( n \times \text{risk} \) against \( \log n \). These curves are supposed to be linear so we provided the squared correlation \( (R^2) \) values of each of the curves in this plot. One can see that the \( R^2 \) values are close to one for every estimator except CV. Finally, in the bottom-right plot, we plotted the logarithm of the risk against \( \log n \). We expect the curves here to have a near-linear relationship with negative slope of \(-1\). The least squares slope values for the different curves are given in the legend in this and it is clear that, for the non-CV estimators, the slope is indeed close to \(-1\).

The numerical results in Figure 4 for the non-CV estimates therefore clearly support our theoretical results. On the other hand, the behavior of the CV estimators seems more complicated and a theoretical study of their risk performance is beyond the scope of the present paper.
We also show results for $f_2^*$ where we evaluated the performance of trend filtering for $r = 2$. We did a simplified study here with the three estimators: (a) the ideal constrained estimator (3) with $V = V^* = n\|D^2\theta^*\|_1$, (b) the penalized estimator (4) with $\lambda$ taken to be (50) with $\Gamma = 1/16$, and (c) the penalized estimator (4) with $\lambda$ taken to be (52) with $\Gamma = 1/16$. Note that our theoretical results apply to (50) and (52) for a sufficiently large $\Gamma$. For $f_2^*$, we found in simulations that $\Gamma = 1/16$ was large enough to yield the desired rates. Higher values of $\Gamma$ inflated risk but gave similar risk decay rates. We could not compute the ideal penalized estimator with $\lambda = \lambda^*$ (defined in (27)) here as the convex optimization problem to compute $\lambda\theta^*(z)$ was highly ill-conditioned for $n \geq 1000$ so that MOSEK seemed unable to find the global minimum (see Section E for more details). We also did not compute CV estimates here as these are not the focus of this paper.

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

5. Discussion

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

5.1. Weakening our assumptions

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

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

Fig 5: Risk plots when the true function is $f_2^*$. 

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

5. Discussion

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

5.1. Weakening our assumptions

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

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

This basic setting (which is standard and used in many theoretical papers on univariate nonparametric regression) can be generalized in many ways and we mention two extensions involving the design points \( x_1, \ldots, x_n \) below. One is the situation where \( x_1, \ldots, x_n \) are not equally spaced. In this case, note that the penalty terms in (3) and (4) need to be changed for \( r \geq 2 \); see e.g., Tibshirani [46]. We believe that our results will still hold in this case provided \( x_1, \ldots, x_n \) satisfy \( \kappa_1/n \leq x_i - x_{i-1} \leq \kappa_2/n \) for two constants \( \kappa_1 \) and \( \kappa_2 \). However, this would make the notation in our proofs quite cumbersome.

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

### 5.2. Constrained and penalized estimators

As mentioned in the Introduction, we have studied both constrained and penalized versions of trend filtering while previous papers have focussed on the penalized estimator alone. When the noise level \( \sigma^2 \) tends to zero, it can be proved that the constrained estimator with \( V = V^* := V^{(r)}(\theta^*) \) is better than the penalized estimator for every choice of the tuning parameter \( \lambda \). More precisely,

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

Here \( \lambda \) is even allowed to depend on \( \theta^* \) as long as it is non-random. Inequality (55) follows from the results of Oymak and Hassibi [35] as described below: Oymak and Hassibi [35, Theorem 2.1] implies

\[
\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R(\hat{\theta}^{(r)}_{\nu}, \theta^*) = \frac{1}{n} \mathbb{E} \left( \inf_{v \in \text{cone}(\partial g(\theta^*))} \|Z - v\|^2 \right) \tag{56}
\]

and Oymak and Hassibi [35, Theorem 1.1] implies

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

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

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

5.3. Results for data-dependent tuning parameters

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

We would like to note that this feature is also present in earlier papers on the trend filtering estimators. The strong sparsity risk results of Lin et al. [30] hold for the tuning choice (43) which depends on $\theta^*$. The results of Dalalyan, Hebiri and Lederer [8] and Ortelli and van de Geer [34] hold for the tuning choice (44) which does not depend on $\theta^*$ but depends on the noise level $\sigma$ and the probability level $\delta$ (note that these results of [8, 34] give only high probability statements and not expectation (risk) bounds).

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

5.4. Connections to results for the LASSO

The trend filtering estimators are closely related to the LASSO estimator of Tibshirani [45]. Indeed, for $r = 1$, it is easy to see that the constrained estimator $\hat{\theta}^{(1)}_V$ is exactly equal to $X\hat{\beta}_V$ where $X$ is the $n \times n$ matrix whose $(i,j)$th entry equals $I\{i \geq j\}$ and $\hat{\beta}_V := \text{argmin}_{\beta \in \mathbb{R}^n} \{||Y - X\beta||^2 : \sum_{i=2}^n |\beta_i| \leq V\}$. Therefore our strong sparsity risk results for $\hat{\theta}^{(1)}_V$ can simply be seen as results for the LASSO estimator for this special design matrix $X$. This connection to LASSO also holds for $r \geq 2$ (see Tibshirani [46]).

Based on this link to the LASSO, it might seem possible to believe that our results might be derivable from general theorems about the LASSO. However, existing strong
sparsity risk bounds for the LASSO impose stringent conditions on the design matrix (such as the compatibility condition or the restricted eigenvalue condition) which do not hold for this particular design matrix $X$ (see Dalalyan, Hebiri and Lederer [8]). The relaxed compatibility condition of [8] does hold who use this condition to prove rates under strong sparsity but this argument is not strong enough to yield the $\frac{k+1}{n} \log \frac{en}{k+1}$ bound. More importantly, it is not clear if the relaxed compatibility condition of [8] or a modified version of it holds for $r \geq 2$.

5.5. Comparison to the $L^0$ estimators

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

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

(58)

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

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

(59)

A proof of this result for $r = 1$ can be found in the recent paper Gao, Han and Zhang [15, Theorem 2.1]. We could not find an exact reference for $r \geq 2$ but we believe that (59) should be true based on the regression connection described in the previous subsection and existing results for $L^0$-penalized estimators in linear regression (see e.g., [38, Theorem 4]).

From a comparison of (59) with (18), it might seem that the constrained trend filtering estimator (with $V = V^*$) has similar performance under strong sparsity as that of the $L^0$ estimator. However, it must be kept in mind here that (18) requires the minimum length condition (13) while the bound (59) for the $L^0$ estimator does not require any such minimum length condition. Without the minimum length condition, the $L^1$ estimator performs much worse compared to the $L^0$ estimator as proved in the recent paper Fan and Guan [12]. Note, however, that the minimum length condition is quite natural from the point of view of estimating piecewise polynomial functions.

From a computational viewpoint, (58) can be efficiently computed for $r = 1$ via dynamic programming (see e.g., Winkler and Liebscher [53]) but it is not clear how to compute it for $r \geq 2$. On the other hand, the trend filtering estimators are efficiently computable for every $r \geq 2$ via convex optimization (see e.g., Arnold and Tibshirani [2] and Kim et al. [25] for details).

5.6. Connection to shape constrained estimators

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

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

Here $D^{(r)} \theta \geq 0$ means that each component of the vector $D^{(r)} \theta$ is nonnegative. When $r = 1$, (60) coincides with the classical isotonic least squares estimator and when $r = 2$, (60) coincides with the convex least squares estimator (see Groeneboom and Jongbloed [19] for an introduction to shape constrained estimation). Like the trend filtering estimators, the shape constrained estimators enjoy the property that $D^{(r)} \hat{\theta}^{(r)}_{\text{shape}}$ is sparse. However, unlike the trend filtering estimators, there is no tuning parameter in (60) (of course, (60) is only applicable in situations where $\theta^*$ satisfies the constraint $D^{(r)} \theta^* \geq 0$ exactly or in some approximate sense).

The risk of (60) under the strong sparsity assumption (and the shape assumption $D^{(r)} \theta \geq 0$) has received much recent attention (see Guntuboyina and Sen [20] for a recent survey). In Bellec [3], it was proved that

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

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

**Acknowledgements**

We thank Ryan Tibshirani for informing us about the reference, Steidl, Didas and Neu- mann [44], and for many other helpful comments. We are also extremely thankful to the Associate Editor and the anonymous referees for very detailed comments on an earlier version of the paper. Their feedback greatly improved the quality of the paper.
Supplementary Material (including proofs of main results)

Here we provide proofs of the results in the paper and some additional simulation results. The material is organized as follows. Section A contains a summary of various existing results from the literature on convex-constrained least squares estimators as well as convex analysis and geometry that are needed for our main proofs. Section B contains proofs of our main results in Section 2 of the main paper. Section C contains proofs of various technical supporting results that were crucially used in the proofs of Section B. Section D contains additional technical results and proofs. Finally Section E contains some additional simulation results.

Appendix A: Preliminaries

In this section, we state some existing general results on the risk of constrained and penalized least squares estimators from the literature. These results will be used in the proofs of our main theorems from Section 2. We shall also state some standard results from convex analysis and convex geometry which will be used in our arguments.

Let us start with results for convex constrained least squares estimators. These are estimators of the form
\[
\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \| Y - \theta \|^2 : \theta \in K \right\}.
\] (62)
for a closed convex set \( K \). Note that the constrained trend filtering estimator \( \hat{\theta}^{(r)} \) is a special case of this estimator when \( K \) is taken to be the set \( K^{(r)}(V) \) defined as
\[
K^{(r)}(V) := \left\{ \theta \in \mathbb{R}^n : \| D^{(r)} \theta \|_1 \leq Vn^{1-r} \right\}.
\] (63)
The general theory of convex-constrained least squares estimators has a long history and is, by now, well established (see e.g., Chatterjee [7], Hjort and Pollard [23], Van de Geer [47], Van der Vaart and Wellner [50]). The following result, essentially from Chatterjee [7] (see Remark A.1) provides upper bounds for the risk of \( \hat{\theta} \). This result will be used in the proof of Theorem 2.1.

**Theorem A.1.** Suppose \( Y \sim N_n(\theta^*, \sigma^2 I_n) \) for some \( \theta^* \in K \) and consider the estimator (62). Then there exists a universal positive constant \( C \) such that
\[
R(\hat{\theta}, \theta^*) := \frac{1}{n} \mathbb{E}_{\theta^*} \| \hat{\theta} - \theta^* \|^2 \leq \frac{C}{n} \max(t_0^2, \sigma^2)
\] (64)
for every \( t_0 > 0 \) which satisfies
\[
\mathbb{E} \left[ \sup_{\theta \in K : \| \theta - \theta^* \| \leq t_0} \langle \xi, \theta - \theta^* \rangle \right] \leq \frac{t_0^2}{2} \quad \text{where} \ \xi \sim N_n(0, \sigma^2 I_n).
\] (65)
Remark A.1. The purpose of this remark is to describe how Theorem A.1 follows from the results of Chatterjee [7] which are all stated for $\sigma = 1$. Extending Chatterjee [7, Proof of Theorem 1.1] in a straightforward manner to the case of arbitrary $\sigma > 0$, one obtains that

$$
\mathbb{P}\left\{ \|\hat{\theta} - \theta^*\| - t_{\theta^*} \geq x\sqrt{t_{\theta^*}} \right\} \leq 3 \exp\left( \frac{-x^4}{32\sigma^2 (1 + x/\sqrt{t_{\theta^*}})^2} \right) 
$$

for every $x \geq 0$ where $t_{\theta^*}$ is defined as the maximizer of

$$
t \mapsto \mathbb{E} \left[ \sup_{\theta \in K: \|\theta - \theta^*\| \leq t} \langle \xi, \theta - \theta^* \rangle - \frac{t^2}{2} \right]
$$

over $t \geq 0$. Inequality (66) implies that whenever $x \geq \sqrt{t_{\theta^*}}$, we obtain

$$
\mathbb{P}\left\{ \|\hat{\theta} - \theta^*\| - t_{\theta^*} \geq x\sqrt{t_{\theta^*}} \right\} \leq 3 \exp\left( \frac{-t_{\theta^*}x^2}{128\sigma^2} \right).
$$

This is because $1 + x/\sqrt{t_{\theta^*}} \leq 2x/\sqrt{t_{\theta^*}}$ under the assumption that $x \geq \sqrt{t_{\theta^*}}$. Replacing $x$ by $u/\sqrt{t_{\theta^*}}$, we obtain

$$
\mathbb{P}\left\{ \|\hat{\theta} - \theta^*\| - t_{\theta^*} \geq u \right\} \leq 3 \exp\left( \frac{-u^2}{128\sigma^2} \right) \quad \text{for } u \geq t_{\theta^*}.
$$

Multiplying both sides by $u$ and integrating from $u = t_{\theta^*}$ to $u = \infty$, we get

$$
\mathbb{E} \left( \left( \|\hat{\theta} - \theta^*\| - t_{\theta^*} \right)^2 - t_{\theta^*}^2 \right) + 3 \int_0^\infty u \exp\left( \frac{-u^2}{128\sigma^2} \right) du \leq C\sigma^2.
$$

This implies that (via $a^2 \leq 2(a - b)^2 + 2b^2$)

$$
\mathbb{E} \left( \|\hat{\theta} - \theta^*\| - t_{\theta^*} \right)^2 \leq C\sigma^2 + 2t_{\theta^*}^2.
$$

which further implies that

$$
\mathbb{E}\|\hat{\theta} - \theta^*\|^2 \leq 6t_{\theta^*}^2 + C\sigma^2 \leq C \max \left( t_{\theta^*}^2, \sigma^2 \right).
$$

From here, we obtain (64) by noting that $t_{\theta^*} \leq t_0$ which follows from Chatterjee [7, Proposition 1.3].

The risk of $\hat{\theta}$ can also be related to the tangent cones of the closed convex set $K$ at $\theta^*$. To describe these results, we need some notation and terminology. The tangent cone of $K$ at $\theta \in K$ is defined as

$$
T_K(\theta) := \text{Closure}\{t(\eta - \theta) : t \geq 0, \eta \in K\}. \quad (67)
$$

Informally, $T_K(\theta)$ represents all directions in which one can move from $\theta$ and still remain in $K$. Note that $T_K(\theta)$ is a cone which means that $a\alpha \in T_K(\theta)$ for every $\alpha \in T_K(\theta)$ and $a \geq 0$. It is also easy to see that $T_K(\theta)$ closed and convex.
The statistical dimension of a closed convex cone $T \subseteq \mathbb{R}^n$ is defined as

$$
\delta(T) := \mathbb{E}\|\Pi_T(Z)\|^2 \quad \text{where } Z \sim N_n(0, I_n)
$$

and $\Pi_T(Z) := \arg\min_{u \in T} \|Z - u\|^2$ is the projection of $Z$ onto $T$. The terminology of statistical dimension is due to Amelunxen et al. [1] and we refer the reader to this paper for many properties of the statistical dimension. The statistical dimension $\delta(T)$ is closely related to the Gaussian width of $T$ which is defined as

$$
w(T) := \mathbb{E}\left[\sup_{\theta \in T, \|\theta\| \leq 1} \langle Z, \theta \rangle \right] \quad \text{where } Z \sim N_n(0, I_n). \quad (68)
$$

Indeed, it has been shown in Amelunxen et al. [1, Proposition 10.2] that

$$
w^2(T) \leq \delta(T) \leq w^2(T) + 1 \quad (69)
$$

for every closed convex cone $T$.

The relevance of these notions to the estimator $\hat{\theta}$ (defined in (62)) is that the risk of $\hat{\theta}$ can be related to the statistical dimension of tangent cones of $K$. This is the content of the following result due to Bellec [3, Corollary 2.2].

**Theorem A.2.** Suppose $Y \sim N_n(\theta^*, \sigma^2 I_n)$ for some $\theta^* \in \mathbb{R}^n$. Then

$$
R(\hat{\theta}, \theta^*) \leq \inf_{\theta \in K} \left[\frac{1}{n}\|\theta - \theta^*\|^2 + \frac{\sigma^2}{n}\delta(T_K(\theta))\right]. \quad (70)
$$

Moreover for every $x > 0$, we have

$$
\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \leq \inf_{\theta \in K} \left[\frac{1}{n}\|\theta - \theta^*\|^2 + \frac{2\sigma^2}{n}\delta(T_K(\theta))\right] + \frac{4\sigma^2 x}{n}
$$

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

**Remark A.2.** A useful lower bound corresponding to (70) has been proved by Oymak and Hassibi [35, Theorem 2.1]. This result states that when $\theta^* \in K$, we have

$$
\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R(\hat{\theta}, \theta^*) = \frac{1}{n}\delta(T_K(\theta^*)) \quad (71)
$$

which means that $\delta(T_K(\theta^*))/n$ provides a precise description of $R(\hat{\theta}, \theta^*)$ in the low $\sigma$ limit. The fact (71) will be used in the proof of Lemma 2.4.

An interesting aspect of Theorem A.2 is that $\theta^*$ is allowed to be any vector in $\mathbb{R}^n$; in particular, it is not necessary that $\theta^* \in K$. Note that combining Theorem A.2 with the bound $\delta(T) \leq w^2(T) + 1$ from (69), we obtain the following risk and loss bounds in terms of the Gaussian width of tangent cones:

$$
R(\hat{\theta}, \theta^*) \leq \inf_{\theta \in K} \left[\frac{1}{n}\|\theta - \theta^*\|^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{n}w^2(T_K(\theta))\right] \quad (72)
$$
and for every $x > 0$,
\[
\frac{1}{n} \| \hat{\theta} - \theta^* \|^2 \leq \inf_{\theta \in K} \left[ \frac{1}{n} \| \theta - \theta^* \|^2 + \frac{2\sigma^2}{n} + \frac{2\sigma^2}{n}w^2(T_K(\theta)) \right] + \frac{4\sigma^2x}{n} \tag{73}
\]
with probability at least $1 - e^{-x}$. The above pair of bounds will be our starting points in the proof of Theorem 2.2. With these bounds, the main task for proving Theorem 2.2 (as well as inequality (16) in Remark 2.3) will involve showing the existence of a constant $C_r$ depending only on $r$ such that
\[
w(T_K^{(V)}(\theta)) \leq C_r \sqrt{n\Delta_r(\theta)} \tag{74}
\]
for every $\theta \in \mathbb{R}^n$ with $V^{(r)}(\theta) = V$. Indeed, combining the inequalities (72) and (74), we obtain
\[
R(\hat{\theta}, \theta^*) \leq \inf_{\theta \in K} \left[ \frac{1}{n} \| \theta - \theta^* \|^2 + \frac{\sigma^2}{n} + C^2_r \sigma^2 \Delta_r(\theta) \right].
\]
Because $\Delta_r(\theta) \geq (k+1)/n \geq 1/n$, the above bound clearly implies (15). Similarly, (73), combined with (74), implies (16). The key therefore is to prove (74) which is accomplished in Subsection B.2.

Let us now describe results for penalized estimators of the form $\hat{\theta}_\lambda^0$ defined as
\[
\hat{\theta}_\lambda^0 := \arg\min_{\theta \in \mathbb{R}^n} \left( \frac{1}{2} \| Y - \theta \|^2 + \sigma \lambda g(\theta) \right) \tag{75}
\]
where $g: \mathbb{R}^n \to \mathbb{R}$ is a convex function. The risk of $\hat{\theta}_\lambda^0$ under $Y \sim N_n(\theta^*, \sigma^2 I_n)$ can be bounded by the Gaussian mean squared distance (defined next) of the set $\lambda \partial g(\theta^*) := \{ \lambda v : v \in \partial g(\theta^*) \}$ where $\partial g(\theta^*)$ is the subdifferential of $g$ at $\theta^*$. The Gaussian mean squared distance $D(C)$ of a nonempty set $C \subseteq \mathbb{R}^n$ is defined as
\[
D(C) := \mathbb{E} \left[ \text{dist}^2(Z, C) \right] \quad \text{where dist}(Z, C) := \inf_{x \in C} \| Z - x \| \tag{76}
\]
and $Z \sim N_n(0, I_n)$. The following result, due to Oymak and Hassibi [35, Theorem 2.2] bounds the risk of $\hat{\theta}_\lambda^0$ in terms of $D(\lambda \partial g(\theta^*))$.

**Theorem A.3.** Suppose $Y \sim N_n(\theta^*, \sigma^2 I_n)$. Then
\[
R(\hat{\theta}_\lambda^0, \theta^*) \leq \frac{\sigma^2}{n}D(\lambda \partial g(\theta^*)).
\]

Theorem A.3 will be our starting point for proving Theorem 2.6. Note that the penalized trend filtering estimator $\hat{\theta}_\lambda^{(r)}$ is a special case of (75) with $g(\theta) := n^{r-1}\|D^{(r)}\theta\|_1$ so that Theorem A.3 will imply that the risk of $\hat{\theta}_\lambda^{(r)}$ will be bounded by $(\sigma^2/n)$ times $D(\lambda n^{r-1}\partial f(\theta^*))$ where $f(\theta) := \|D^{(r)}\theta\|_1$. The goal then becomes that of bounding $D(\lambda n^{r-1}\partial f(\theta^*))$ from above in the case when $D^{(r)}\theta^* \neq 0$ (note that Theorem 2.6 does not deal with the case $D^{(r)}\theta^* = 0$; this case is dealt with in Lemma 2.14 whose proof is simpler and more direct).
Our idea for bounding $D(\lambda n^{-1}\partial f(\theta^*))$ is to relate it to the smaller quantity $D(\text{cone}(\partial f(\theta^*)))$ where $\text{cone}(\partial f(\theta^*))$ is the convex cone generated by $\partial f(\theta^*)$:

$$\text{cone}(\partial f(\theta^*)) := \bigcup_{\lambda \geq 0} [\lambda \partial f(\theta^*)].$$

It is clear that $\text{cone}(\partial f(\theta^*))$ contains the set $\lambda n^{-1}\partial f(\theta^*)$ for every $\lambda \geq 0$ and thus by definition of $D(\cdot)$, it follows that

$$D(\lambda n^{-1}\partial f(\theta^*)) \geq D(\text{cone}(\partial f(\theta^*)�).$$

However, we need an upper bound and not a lower bound for $D(\lambda n^{-1}\partial f(\theta^*))$. It turns out that an upper bound can indeed be given for $D(\lambda n^{-1}\partial f(\theta^*))$ in terms of $D(\text{cone}(\partial f(\theta^*)�)$ and additional terms (involving $\lambda$ and the vectors $v_0$ and $v^*$ defined in (25)). This result (formally stated in Proposition B.5) can be seen as a generalization of Foygel and Mackey [13, Proposition 1]. The advantage of Proposition B.5 is that it reduces the task to upper bounding $D(\text{cone}(\partial f(\theta^*)�)$. As we shall outline below, by some standard facts from convex analysis, it follows that

$$D(\text{cone}(\partial f(\theta^*)�) = \delta(T_{K^{(r)}(V^*)}'(\theta^*)) \leq 1 + w^2(T_{K^{(r)}(V^*)}'(\theta^*)) \quad (77)$$

where $V^* := V^{(r)}(\theta^*)$. This allows us to use the bound (74) (established in the course of the proof of Theorem 2.2) to bound $D(\text{cone}(\partial f(\theta^*)�).

We shall now explain why (77) is true. Note that we only need to prove the first equality (the second inequality is a consequence of (69)). For this, we need to introduce the notions of normal cone and polar cone from convex analysis (see, for example, Rockafellar [39] for background on these standard notions). The normal cone of a convex set $C \subseteq \mathbb{R}^n$ at a point $x \in C$ is defined by

$$N_C(x) := \{u \in \mathbb{R}^n : \langle y - x, u \rangle \leq 0 \text{ for every } y \in C \}.$$

Next let us define the notion of a polar cone. The polar $T^o$ of a nonempty closed convex cone $T \subseteq \mathbb{R}^n$ is defined as

$$T^o := \{u \in \mathbb{R}^n : \langle u, x \rangle \leq 0 \text{ for every } x \in T \} \quad (78)$$

The following result (see, for example, Rockafellar and Wets [40, Example 6.24]) states that for every convex set $C$ and $x \in C$, the normal cone $N_C(x)$ equals the polar of the tangent cone $T_C(x)$ (recall that $T_C(x)$ is defined in (67)).

**Lemma A.4.** For every convex set $C \subseteq \mathbb{R}^n$ and $x \in C$, we have

$$N_C(x) = (T_C(x))^o.$$

The next result states that $\text{cone}(\partial f(\theta^*))$ equals $N_C(\theta^*)$ where

$$\mathcal{C} := \{\theta \in \mathbb{R}^n : f(\theta) \leq f(\theta^*)\} \quad (79)$$

under some conditions on the convex function $f$ and $\theta^*$. This result follows from Rockafellar [39, Theorem 23.7 and Corollary 23.7.1].
Lemma A.5. Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\theta^* \in \mathbb{R}^n$ is such that $\partial f(\theta^*)$ is a compact convex set with $0 \notin \partial f(\theta^*)$. Then
\[
\text{cone}(\partial f(\theta^*)) = N_C(\theta^*)
\]
where $C$ is given by (79).

Observe now that when $f(\theta) := \|D(r)\theta\|_1$ and $\theta^*$ is such that $D(r)\theta^* \neq 0$, the conditions in Lemma A.5 hold as can be seen from the characterization of $\partial f(\theta^*)$ in Proposition 2.5.

The assumption that $0 \notin \partial f(\theta^*)$ holds because for every $v \in \partial f(\theta^*)$, we must have
\[
\sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r - 1} \right) v_i = \text{sgn}(\langle D(r)\theta^* \rangle_{j-r})
\]
for every $j$ such that $\langle D(r)\theta^* \rangle_{j-r} \neq 0$ (there must exist at least one such $j$ because of the assumption that $D(r)\theta^* \neq 0$).

Further, for $f(\theta) := \|D(r)\theta\|_1$, it is easy to see that the set $C$ in (79) satisfies
\[
C = \{ \theta \in \mathbb{R}^n : \|D(r)\theta\|_1 \leq \|D(r)\theta^*\|_1 \} = K^{(r)}(V^{(r)}(\theta^*)) = K^{(r)}(V^*)
\]
because $V^* := V^{(r)}(\theta^*)$ and $K^{(r)}(V)$ is defined as in (63). Putting together the conclusions of Lemma A.4 and Lemma A.5, we therefore deduce that
\[
\text{cone}(\partial f(\theta^*)) = \left( T_{K^{(r)}(V^*)}(\theta^*) \right)^o.
\]

From here, in order to prove (77), we need another standard fact from convex geometry (see, for example, Hiriart-Urruty and Lemaréchal [22, Theorem 3.2.5]). This result states that for every closed convex cone $T \subseteq \mathbb{R}^n$, we have
\[
\Pi_{T^o}(z) = z - \Pi_T(z) \quad \text{for every } z \in \mathbb{R}^n \tag{80}
\]
where $\Pi_K(z)$ denotes the projection of $z$ onto $K$.

Applying (80) to $T := \text{cone}(\partial f(\theta^*))$ (which is a closed convex cone when $D(r)\theta^* \neq 0$; closedness follows, for example, from Rockafellar [39, Corollary 9.6.1]), we obtain
\[
z - \Pi_{\text{cone}(\partial f(\theta^*))}(z) = \Pi_{T_{K^{(r)}(V^*)}}(z). \tag{81}
\]

From the above identity (and the definitions of $D(\cdot)$ and $\delta(\cdot)$), the fact (77) readily follows. The fact (77) will be crucially used in the proof of Theorem 2.6. Also, the identity (81) will play a key role in the proofs of Lemma 2.9 and Lemma 2.12.

Appendix B: Proofs of the Main Results

In the section, we provide the proofs of the following results in Section 2: Theorem 2.1, Theorem 2.2 (and inequality (16) in Remark 2.3), Corollary 2.3, Lemma 2.4, Theorem...
Corollary 2.8, Lemma 2.9 and Corollary 2.10, Corollary 2.11, Lemma 2.12 and Corollary 2.13 and finally, Lemma 2.14. In addition to these results, Section 2 also contains Proposition 2.5 and Lemma 2.7. These are proved in Subsection C.4.

Some of the proofs presented in this section will introduce and use additional technical results. These technical results will be proved in the Section C.

### B.1. Proof of Theorem 2.1

We prove Theorem 2.1 in this subsection. As mentioned at the start of Section A, our starting point for this proof is Theorem A.1; note that $\hat{\theta}_V^{(r)}$ is the least squares estimator subject to the constraint that $\theta \in K^{(r)}(V)$ (recall that the set $K^{(r)}(V)$ is defined in (63)). Theorem A.1 implies that we can bound the risk of $\hat{\theta}_V^{(r)}$ via upper bounds for

$$G(t) := \mathbb{E} \left[ \sup_{\theta \in K^{(r)}(V) : \|\theta - \theta^*\| \leq t} \langle \xi, \theta - \theta^* \rangle \right]$$

for $t > 0$. Our upper bound for $G(t)$ is proved from the following lemma. Let

$$S_r(V,t) := \{ \alpha \in \mathbb{R}^n : \|\alpha\| \leq t, \|D^{(r)}\alpha\|_1 \leq Vn^{1-r} \}.$$  \hspace{1cm} (83)

**Lemma B.1.** Fix an integer $r \geq 1$. Then there exists a positive constant $C_r$ such that for every $n \geq r$, $t \geq 0$ and $V \geq 0$, we have

$$\mathbb{E} \left[ \sup_{\theta \in S_r(V,t)} \langle \xi, \theta \rangle \right] \leq C_r \sigma t \left( \frac{\sqrt{n}V}{t} \right)^{1/(2r)} + C_r \sigma t \sqrt{\log(en)}. \hspace{1cm} (84)$$

Lemma B.1 is proved in Subsection C.3 and the ideas behind its proof are as follows. By Dudley’s entropy bound, the left hand side of (84) can be bounded from above by the metric entropy numbers (formally defined in Subsection C.3) of the set $S_r(V,t)$ (defined in (83)). The metric entropy of $S_r(V,t)$ will be bounded by controlling the fat shattering dimension (see Subsection C.1 for details).

Below, we provide the proof of Theorem 2.1 based on Lemma B.1.

**Proof of Theorem 2.1.** As $\|D^{(r)}\theta^*\|_1 \leq Vn^{1-r}$, it follows that $\theta^* \in K^{(r)}(V)$ (the set $K^{(r)}(V)$ is defined in (63)). Theorem A.1 implies that

$$R(\hat{\theta}_V^{(r)}, \theta^*) \leq \frac{C}{n} \max \left( t_0^2, \sigma^2 \right) \hspace{1cm} (85)$$

for a universal positive constant $C$, where $t_0 > 0$ is such that $G(t_0) \leq t_0^2/2$ with $G(t)$ defined as in (82). In order to apply this result, we need to bound the function $G(t)$ from above. By triangle inequality, $\|D^{(r)}(\theta - \theta^*)\|_1 \leq \|D^{(r)}\theta\|_1 + \|D^{(r)}\theta^*\|_1$ so that

$$G(t) = \mathbb{E} \sup_{\theta \in K^{(r)}(V) : \|\theta - \theta^*\| \leq t} \langle \xi, \theta - \theta^* \rangle \leq \mathbb{E} \sup_{\alpha \in \mathbb{R}^n : \|\alpha\| \leq t, \|D^{(r)}\alpha\|_1 \leq 2Vn^{1-r}} \langle \xi, \alpha \rangle.$$
The right hand side above is controlled in Lemma B.1 from which we deduce that

\[ G(t) \leq C_r \sigma t \left( \frac{\sqrt{n} V}{t} \right)^{\frac{1}{2r}} + C_r \sigma t \sqrt{\log(en)} \]

for a constant \( C_r \) depending on \( r \) alone. We now observe that

\[ C_r \sigma t \left( \frac{\sqrt{n} V}{t} \right)^{\frac{1}{2r}} \leq \frac{t^2}{4} \iff t \geq (4C_r)^{2r/(2r+1)} \sigma^{2r/(2r+1)} (V \sqrt{n})^{1/(2r+1)} \]

and

\[ C_r \sigma t \sqrt{\log(en)} \leq \frac{t^2}{4} \iff t \geq 4C_r \sigma \sqrt{\log(en)}. \]

It follows therefore that \( G(t_0) \leq \frac{t_0^2}{2} \) provided

\[ t_0 := \max \left( (4C_r)^{2r/(2r+1)} \sigma^{2r/(2r+1)} (V \sqrt{n})^{1/(2r+1)}, 4C_r \sigma \sqrt{\log(en)} \right). \]

The proof of inequality (8) is therefore complete by inequality (85).

Inequality (9) can be derived as a consequence of (8) and the fact that the map \( y \mapsto \|\hat{\theta}^{(r)}_V - \theta^*\| \) is 1-Lipschitz (see e.g., van de Geer and Wainwright [49, Section 2]). By the usual concentration inequality for Lipschitz functions of Gaussian variables, this gives

\[ \mathbb{P} \left\{ \|\hat{\theta}^{(r)}_V - \theta^*\| \geq \mathbb{E}_{\theta^*} \|\hat{\theta}^{(r)}_V - \theta^*\| + \sigma z \right\} \leq \exp \left( \frac{-z^2}{2} \right). \]

This gives that

\[ \frac{1}{n} \|\hat{\theta}^{(r)}_V - \theta^*\|^2 \leq 2R(\hat{\theta}^{(r)}_V, \theta^*) + \frac{4\sigma^2 x}{n} \]

with probability at least \( 1 - e^{-x} \) so that inequality (9) follows from (8).

\[ \Box \]

**B.2. Proof of Theorem 2.2**

Our starting points for proving Theorem 2.2 are the inequalities (72) and (73) applied to \( K = K^{(r)}(V) \). From here, it is clear that inequality (15) (as well as (16)) both follow from the inequality (74). Writing explicitly the Gaussian width \( w(T_{K^{(r)}(V)}(\theta)) \), we see that (74) is equivalent to proving that

\[ \mathbb{E} \left[ \sup_{\alpha \in T_{K^{(r)}(V)}(\theta) : \|\alpha\| \leq 1} \langle Z, \alpha \rangle \right] \leq C_r \sqrt{n \Delta_r(\theta)} \quad (86) \]

for every \( \theta \in \mathbb{R}^n \) with \( V^{(r)}(\theta) = V \). For this, we obviously need to understand the set \( T_{K^{(r)}(V)}(\theta) \) for \( \theta \in \mathbb{R}^n \) with \( V^{(r)}(\theta) = V \). The following result provides a necessary condition that is satisfied by every vector \( \alpha \in T_{K^{(r)}(V)}(\theta) \) with \( \|\alpha\| \leq 1 \). The proof of this
lemma is given in Subsection C.2. Recall, from Section 2, the notion of \(r\)th order knots (along with their signs) of vectors in \(\mathbb{R}^n\). We shall also use the following notation. For \(\alpha \in \mathbb{R}^m\) and \(1 \leq a \leq b \leq m\), we let

\[
V_{a,b}(\alpha) := V(\alpha_a, \ldots, \alpha_b) = |\alpha_{a+1} - \alpha_a| + \cdots + |\alpha_b - \alpha_{b-1}|.
\]

**Lemma B.2.** Fix \(V > 0\), \(r \geq 1\), \(n \geq r\) and \(\theta \in \mathbb{R}^n\) with \(V(\theta) = V\). Suppose \(2 \leq j_1 < \cdots < j_k \leq n - r + 1\) denote any set of indices which contains all the \(r\)th order knots of \(\theta\). Let \(\mathbf{r}_1, \ldots, \mathbf{r}_k\) be such that \(\mathbf{r}_i\) is the sign of the knot corresponding to \(j_i\); if \(j_i\) is a knot and \(\mathbf{r}_i\) is arbitrary in \([-1,0,1]\) if \(j_i\) is not a knot. Also let \(j_0 = 1\), \(j_{k+1} = n - r + 2\) and \(\mathbf{r}_0 = \mathbf{r}_{k+1} = 0\). The indices \(j_0, j_1, \ldots, j_k\) define a partition \(I_0, \ldots, I_k\) of \(\{1, \ldots, n\}\) in the following way: \(I_i := \{j_0, \ldots, j_i + r - 2\}\) and

\[
I_i = \{j_i + r - 1, \ldots, j_{i+1} + r - 2\} \quad \text{for} \ i = 1, \ldots, k.
\]

Let \(n_i\) denote the cardinality of \(I_i\) for \(i = 0, 1, \ldots, k\); i.e., \(n_0 := j_1 + r - 2\) and \(n_i = j_{i+1} - j_i\) for \(1 \leq i \leq k\). Then there exists a positive constant \(C_r\) (that depends on \(r\) alone) such that for every \(\alpha \in \mathcal{T}_{K(r)(V)}(\theta)\) with \(\|\alpha\| \leq 1\), there exist indices \(\ell_0 \in I_0, \ell_1 \in I_1, \ldots, \ell_k \in I_k\) such that

\[
\sum_{i=0}^{k} \Gamma_i(\alpha, \ell_i) \leq C_r \left[ \sum_{i=0}^{k} n_i^{1-2r} I\{\mathbf{r}_i \neq \mathbf{r}_{i+1}\} \right]
\]

where

\[
\Gamma_i(\alpha, \ell_i) := V_{j_i, j_{i+1}-1}(\Delta) - \mathbf{r}_{i+1}(\Delta_{j_{i+1}-1} - \Delta_{\ell_i}) - \mathbf{r}_i(\Delta_{\ell_i} - \Delta_{j_i})
\]

with \(\Delta = (\Delta_1, \ldots, \Delta_{n-r+1}) := D^{(r-1)}\alpha\).

**Remark B.1.** It may be noted that the indices \(j_1, \ldots, j_k\) in Lemma B.2 are not exactly the knots of \(\theta\). They are any set of indices that contain the knots of \(\theta\). We shall mostly work with the case when \(j_1, \ldots, j_k\) are exactly the set of knots of \(\theta\) but we shall need this additional generality to deal with one special situation when some of the distances between the knots of \(\theta\) are too large. In this case (see the last part of the proof of Theorem 2.2), we shall add additional indices to the knots in order to keep the inter-distances manageable.

The insight provided by Lemma B.2 into the structure of \(\{\alpha \in \mathcal{T}_{K(r)(V)}(\theta) : \|\alpha\| \leq 1\}\) (note that understanding this set is necessary for proving (86)) is as follows. Suppose that \(n_0, \ldots, n_k\) are such that the right hand side of (88) is small. In this case, Lemma B.2 implies that for every \(\alpha \in \mathcal{T}_{K(r)(V)}(\theta)\) with \(\|\alpha\| \leq 1\), there exist indices \(\ell_0, \ldots, \ell_k\) for which \(\sum_{i=0}^{k} \Gamma_i(\alpha, \ell_i)\) is small. Figure 6 displays two unit norm vectors in the tangent cone of a piecewise constant vector \(\theta\) (i.e., \(r = 1\)) and the corresponding indices \(\ell_0, \ldots, \ell_k\).

It is helpful here to observe that \(\Gamma_i(\alpha, \ell_i)\) is always nonnegative. Also, if \(\Gamma_i(\alpha, \ell_i) = 0\), then \(D^{(r-1)}\alpha\) is made of two monotone pieces in the interval from \(j_i\) to \(j_{i+1} - 1\) (one piece from \(j_i\) to \(\ell_i\) and the other from \(\ell_i\) to \(j_{i+1} - 1\)). When \(r = 1\), this means that \(\alpha\) is made of two monotone pieces in the interval from \(j_i\) to \(j_{i+1} - 1\). When \(r = 2\), this means that \(\alpha\) is made of two convex/concave pieces in the interval from \(j_i\) to \(j_{i+1} - 1\). For general \(r\), this
Fig 6: Let \( r = 1, n = 200 \) and let \( \theta \) be the vector obtained by sampling \( f_1^* \) at \( n \) equally spaced points with end points 0 and 1 (here \( f_1^* \) is the piecewise constant vector from Section 4 of the main paper). This vector \( \theta \) has \( k = 4 \) jumps at \( j_1 = 41, j_2 = 81, j_3 = 121 \) and \( j_4 = 161 \). These indices (and the indices \( j_i - 1, i = 1, 2, 3, 4 \)) are plotted in black lines in the above pair of plots along with vertical straight lines at \( j_0 = 1 \) and \( j_5 = 200 \). We then considered the tangent cone, \( T := T_{K^{(1)}(V)}(\theta) \), where \( V \) is the variation of \( \theta \) and plotted two vectors \( \alpha \) in \( T \) with \( \|\alpha\| = 1 \). For each of these two vectors \( \alpha \), we also plotted the integers \( \ell_0, \ldots, \ell_4 \) as blue vertical lines. Informally, in the five constant segments corresponding to \( \theta \), each vector \( \alpha \) is approximately made of two monotone segments. This suggests therefore that in order to prove (86), we need to prove bounds on the Gaussian suprema for vectors \( \alpha \in \mathbb{R}^n \) for which \( D^{(r-1)}\alpha \) is nearly monotone. This is the content of the next lemma which is another main ingredient for the proof of Theorem 2.2. 

**Lemma B.3.** Fix \( r \geq 1, n \geq r, 1 \leq l \leq n - r + 1, t > 0 \) and \( \delta \geq 0 \). For \( \theta \in \mathbb{R}^n \), let \( \Delta(\theta) = (\Delta_1(\theta), \ldots, \Delta_{n-r+1}(\theta)) := D^{(r-1)}\theta \). Also let \( \xi \sim N_n(0, \sigma^2 I_n) \). For every \( r_1, r_2 \in \{-1, 0, 1\} \), the quantity

\[
E \sup_{\theta \in \mathbb{R}^n, \|\theta\| \leq t} \{ \langle \xi, \theta \rangle : \Delta = \Delta(\theta), V(\Delta) \leq r_1(\Delta_l - \Delta_1) + r_2(\Delta_{n-r+1} - \Delta_l) + \delta \}
\]

is bounded from above as

\[
G \leq C_r \sigma \left( t + \delta n^{(2r-1)/2} \sqrt{\log(cn)} + C_r \sigma t^{(2r-1)/(2r)} n^{(2r-1)/(4r)} \delta^{1/(2r)} \right)
\]

for a positive constant \( C_r \) that depends on \( r \) alone.
The proof of Lemma B.3 is given in Subsection C.3 (in fact, in Subsection C.3, we prove Lemma C.7 which is a more accurate result compared to Lemma B.3 in the sense that Lemma C.7 gives a bound that depends on the actual values of \( r_1 \) and \( r_2 \)). The proof of Lemma C.7 uses results on expected Gaussian suprema for classes of shape-constrained vectors from Bellec [3].

We are now ready to prove Theorem 2.2.

**Proof of Theorem 2.2.** As the proof is rather long, we divide it into many steps.

**Step I:** We first note that the case when \( V = 0 \) is trivial. This is because the set \( \{ \theta \in \mathbb{R}^n : \|D^{(r)}\theta\|_1 = 0 \} \) is a subspace of dimension \( r \) so that \( \hat{\theta}^{(r)}_{V=0} \) becomes a linear projection onto a subspace. Thus,

\[
\frac{1}{n} \| \hat{\theta}^{(r)}_{V=0} - \theta^* \|^2 - \inf_{\theta \in \mathbb{R}^n : \|D^{(r)}\theta\|_1 = 0} \frac{1}{n} \| \theta - \theta^* \|^2 \sim \frac{\sigma^2}{n} \chi^2_r
\]

where \( \chi^2_r \) denotes the chi-squared distribution with \( r \) degrees of freedom. This and a standard tail bound for chi-squared random variables such as (see e.g., Laurent and Massart [27, Subsection 4.1])

\[
P\{ \chi^2_r \leq 2r + 3x \} \geq 1 - e^{-x} \quad \text{for every } x > 0
\]

prove inequalities (15) and (16) for \( V = 0 \); note that when \( \theta \in \mathbb{R}^n \) is such that \( V^{(r)}(\theta) = V = 0 \), we have \( k_r(\theta) = 0, \delta_r(\theta) = n^{(1/2)-r} \) and \( \Delta_r(\theta) = \frac{2}{n} \log(en) + \frac{1}{n} \).

We shall assume from now on that \( V > 0 \). Based on the discussion at the beginning of this subsection, it is enough to prove (86). We therefore fix \( \theta \in \mathbb{R}^n \) with \( V^{(r)}(\theta) = V \). We need to bound the quantity

\[
G := \mathbb{E} \left[ \sup_{\alpha \in T_{K^{(r)}(V)}(\theta) : \|\alpha\| \leq 1} \langle Z, \alpha \rangle \right]
\]

where \( Z \) is a standard \( n \)-dimensional Gaussian random vector. We bound \( G \) by breaking the set \( \{ \alpha \in T_{K^{(r)}(V)}(\theta) : \|\alpha\| \leq 1 \} \) into smaller subsets.

Let \( 2 \leq j_1 < \cdots < j_k \leq n - r + 1 \) denote all the \( r \)th order knots of \( \theta \). Also let \( t_1, \ldots, t_k \in \{-1, 1\} \) denote the signs of the knots. For convenience, we take \( j_0 = 1, j_{k+1} = n - r + 2 \) and \( t_0 = t_{k+1} = 0 \). Let \( n_0 = j_1 + r - 2 \) and \( n_u = j_{u+1} - j_u \) for \( u = 1, \ldots, k \). Check that \( \sum_{u=0}^k n_u = n \).

**Step II:** We shall prove (86) first under the simplifying assumption that

\[
n_i \leq \frac{2n}{k + 1} \quad \text{for every } i = 0, 1, \ldots, k.
\]

The goal of this step is to find a collection of sets whose union covers \( \{ \alpha \in T_{K^{(r)}(V)}(\theta) : \|\alpha\| \leq 1 \} \) (see (93)). For every vector \( \alpha \in \mathbb{R}^n \), let us define the vectors

\[
\alpha^{(0)} := (\alpha_{j_0}, \ldots, \alpha_{j_1+r-2})
\]
and
\[ \alpha^{(u)} := (\alpha_{j_u+r-1}, \ldots, \alpha_{j_{u+1}+r-2}) \quad \text{for } u = 1, \ldots, k. \]

Note that the vector \( \alpha^{(u)} \) has length exactly equal to \( n_u \) for \( u = 0, \ldots, k \).

Let \( \mathcal{M} \) denote the class of all vectors \( \mathbf{m} := (\mathbf{m}_0, \ldots, \mathbf{m}_k) \) where each \( \mathbf{m}_i \) is an integer with \( 1 \leq \mathbf{m}_i \leq k + 1 \) and such that \( \sum_{i=0}^k \mathbf{m}_i \leq 2(k+1) \). Because the number of \((k+1)\)-tuples of positive integers whose sum is equal to \( p \) equals \( \binom{p-1}{k} \), it is easy to see that \( \mathcal{M} \) is a finite set whose cardinality \( |\mathcal{M}| \) can be bounded as
\[
|\mathcal{M}| \leq \sum_{p=k+1}^{2k+2} \binom{p-1}{k} = \sum_{l=k}^{2k+1} \binom{2k+1}{l-k} \leq 2^{2k+1} \leq 4^{k+1}.
\]

Also let \( \mathcal{L} \) denote the class of all vectors \( \ell := (\ell_0, \ldots, \ell_k) \) where each \( \ell_i \) is an integer such that \( j_0 \leq \ell_0 \leq j_1 - 1 \) and \( j_u + r - 1 \leq \ell_u \leq j_{u+1} - 1 \) for \( u = 1, \ldots, k \). The cardinality \( |\mathcal{L}| \) of \( \mathcal{L} \) is clearly bounded from above by \( \prod_{u=0}^k n_u \).

Let
\[
\delta := C_r \sqrt{n_0^{-2r} + n_k^{-2r} + \sum_{i=1}^{k-1} n_i^{-2r} I\{r_i \neq r_{i+1}\}}
\]
where \( C_r \) is the constant given by Lemma B.2. Note that
\[
\delta \leq C_r \delta_r(\theta) \tag{92}
\]
where \( \delta_r(\theta) \) is as defined in (11). This is because \( n_i \geq n_{is} \) for every \( i \) and \( 1 - 2r < 0 \).

Also for \( \ell \in \mathcal{L} \), \( \alpha \in \mathbb{R}^n \) and \( 0 \leq i \leq k \), let \( \Gamma_i(\alpha, \ell_i) \) be defined (as in (89)) as
\[
V_{j_i,j_{i+1}-1}(D^{(r-1)}\alpha) - r_i+1 \{ (D^{(r-1)}\alpha)_{j_{i+1}-1} - (D^{(r-1)}\alpha)_{\ell_i} \} - r_i \{ (D^{(r-1)}\alpha)_{\ell_i} - (D^{(r-1)}\alpha)_{j_i} \}.
\]

For every \( \mathbf{m}, \mathbf{q} \in \mathcal{M} \) and \( \ell \in \mathcal{L} \), let \( T(\mathbf{m}, \mathbf{q}, \ell) \) denote the set of all \( \alpha \in \mathbb{R}^n \) with \( \|\alpha\| \leq 1 \) for which
\[
\|\alpha^{(i)}\|^2 \leq \frac{\mathbf{m}_i}{k+1} \quad \text{and} \quad \Gamma_i(\alpha, \ell_i) \leq \frac{\mathbf{q}_i \delta}{k+1}
\]
for every \( i = 0, 1, \ldots, k \). We then claim that
\[
\{ \alpha \in T_{K^{(\ell)}(V)}(\theta) : \|\alpha\| \leq 1 \} \subseteq \bigcup_{\mathbf{m}, \mathbf{q} \in \mathcal{M}, \ell \in \mathcal{L}} T(\mathbf{m}, \mathbf{q}, \ell). \tag{93}
\]

To see (93), note first that it follows from Lemma B.2 that for every \( \alpha \in T_{K^{(\ell)}(V)}(\theta) \) with \( \|\alpha\| \leq 1 \), there exists \( \ell \in \mathcal{L} \) such that \( \sum_{i=0}^k \Gamma_i(\alpha, \ell_i) \leq \delta \). This implies that for every
0 ≤ i ≤ k, the inequality 0 ≤ Γ_i(α, ℓ_i) ≤ δ and so there exists an integer 1 ≤ q_i ≤ k + 1 such that
\[ \frac{(q_i - 1)\delta}{k + 1} ≤ Γ_i(α, ℓ_i) ≤ \frac{q_i\delta}{k + 1}. \]
The integers q_0, ..., q_k would then have to satisfy
\[ \delta ≥ \sum_{i=0}^{k} Γ_i(α, ℓ_i) ≥ \delta \sum_{i=0}^{k} \frac{q_i - 1}{k + 1} \]
which is equivalent to \( \sum_{i=0}^{k} q_i ≤ 2(k + 1) \). Thus \( q = (q_0, q_1, ..., q_k) ∈ M \).
Similarly, for each 0 ≤ i ≤ k, the inequality 1 ≥ \|α\|^2 ≥ \|α(i)\|^2 holds so that there exists an integer 1 ≤ m_i ≤ k + 1 such that
\[ \frac{m_i - 1}{k + 1} ≤ \|α(i)\|^2 ≤ \frac{m_i}{k + 1}. \]
As \( \sum_{i=0}^{k} \|α(i)\|^2 ≤ 1 \), the integers m_i, 0 ≤ i ≤ k, satisfy \( \sum_{i=0}^{k} m_i ≤ 2(k + 1) \) which implies that \( m = (m_0, ..., m_k) ∈ M \). This completes the proof of (93).

Step III: In this step we find an upper bound of \( G \) (in (90)) that depends on the collection \( T(m, q, ℓ) \). Using (93), we can bound the quantity \( G \) in (90) via
\[ G ≤ E \left[ \max_{m, q ∈ M, ℓ ∈ L} \sup_{α ∈ T(m, q, ℓ)} \langle Z, α \rangle \right]. \]
Since \( Z \) is Gaussian, the first maximum in the right hand side above can be taken outside the expectation up to an additional correction term. We state this as a general result in Lemma D.1 (see the full statement and proof in Section D). Note that each set \( T(m, q, ℓ) \) contains the zero vector and also that every vector in \( T(m, q, ℓ) \) has norm bounded from above by 1. Therefore the quantity \( D \) in Lemma D.1 can be taken to be 1 (also take \( σ = 1 \) in Lemma D.1). Lemma D.1 thus gives
\[ G ≤ \max_{m, q ∈ M, ℓ ∈ L} E \left[ \sup_{α ∈ T(m, q, ℓ)} \langle Z, α \rangle \right] + \sqrt{4 \log |M| + 2 \log |L|} + \sqrt{\frac{\pi}{2}}. \]
As |\( M \)| ≤ 4^{k+1} and, by concavity of the logarithm,
\[ \log |L| ≤ \sum_{i=0}^{k} \log n_i = (k + 1) \frac{1}{k + 1} \sum_{i=0}^{k} \log n_i \]
\[ ≤ (k + 1) \log \left( \sum_{i=0}^{k} \frac{n_i}{k + 1} \right) \]
\[ = (k + 1) \log \frac{n}{k + 1} ≤ (k + 1) \log \frac{en}{k + 1}, \]
we obtain (using also the fact that $\sqrt{2 + 4 \log 4} < 3$) that
\[
G \leq \max_{m,q \in M, \ell \in L} \mathbb{E} \left[ \sup_{\alpha \in T(m,q,\ell)} \langle Z, \alpha \rangle \right] + 3\sqrt{(k+1) \log \frac{en}{k+1} + \frac{\pi}{2}}. \tag{94}
\]
We now fix $m, q \in M$ and $\ell \in L$ and attempt to bound
\[
G(m, q, \ell) := \mathbb{E} \left[ \sup_{\alpha \in T(m,q,\ell)} \langle Z, \alpha \rangle \right].
\]
We write $\langle Z, \alpha \rangle = \sum_{i=0}^{k} \langle Z^{(i)}, \alpha^{(i)} \rangle$ so that $G(m, q, \ell) \leq \sum_{i=0}^{k} G_i(m, q, \ell)$ where $G_i(m, q, \ell) := \mathbb{E} \sup_{\alpha \in T(m,q,\ell)} \langle Z^{(i)}, \alpha^{(i)} \rangle$. Let us now fix $0 \leq i \leq k$ and bound $G_i(m, q, \ell)$.

By the definition of $T(m, q, \ell)$,
\[
G_i(m, q, \ell) \leq \mathbb{E} \left[ \sup_{\alpha \in T^{(i)}(m,q,\ell)} \langle Z^{(i)}, \alpha^{(i)} \rangle \right] \tag{95}
\]
where
\[
T^{(i)}(m, q, \ell) := \left\{ \alpha \in \mathbb{R}^n : \|\alpha^{(i)}\|^2 \leq \frac{m_i}{k+1}, \Gamma_i(\alpha, \ell_i) \leq \frac{q_i \delta}{k+1} \right\}.
\]

**Step IV:** In this step, we describe how Lemma B.3 can be used to bound the right hand side in (95). Fix $0 \leq i \leq k$. We do this by rewriting the underlying set $T^{(i)}(m, q, \ell)$ in a form recognizable from Lemma B.3. For convenience, let
\[
\delta_i := \frac{q_i \delta}{k+1}.
\]
We claim that for every $0 \leq i \leq k$ and $\alpha \in T^{(i)}(m, q, \ell)$, we have
\[
V(\Delta(\alpha^{(i)})) \leq r_{i+1} \left( (\Delta(\alpha^{(i)}))_{n_i-r+1} - (\Delta(\alpha^{(i)}))_{\ell_i} \right) + r_i \left( (\Delta(\alpha^{(i)}))_{\ell_i} - (\Delta(\alpha^{(i)}))_{1} \right) + \delta_i \tag{96}
\]
where $\Delta(\alpha^{(i)}) := D^{(r-1)}(\alpha^{(i)})$ and $\ell_i'$ is related to $\ell_i$ via
\[
\ell_0 := \ell_0' \quad \text{and} \quad \ell_i = j_i + r - 2 + \ell_i' \quad \text{for} \quad 1 \leq i \leq k. \tag{97}
\]
Before proving (96), let us observe that the expected supremum of $\langle Z^{(i)}, \alpha^{(i)} \rangle$ over all $\alpha^{(i)}$ which satisfy the norm condition $\|\alpha^{(i)}\|^2 \leq m_i/(k+1)$ and which satisfy (96) can be controlled directly using Lemma B.3 (this is done in the next step). The argument for (96) goes as follows. Fix $\alpha \in T^{(i)}(m, q, \ell)$ and observe that, from the definition of $T^{(i)}(m, q, \ell)$, we have $\Gamma_i(\alpha, \ell_i) \leq \delta_i$. For $i = 0$, inequality (96) is exactly the same as $\Gamma_0(\alpha, \ell_0) \leq \delta_0$. For $1 \leq i \leq k$, note that
\[
(D^{(r-1)}(\alpha^{(i)}))_{\ell} = (D^{(r-1)}(\alpha))_{j_i+r-2+\ell} \quad \text{for every} \quad 1 \leq \ell \leq n_i - r + 1, \tag{98}
\]
which implies that \( V(D^{(r-1)}\alpha^{(i)}) = V_{j_i+r-1,j_i+1-1}(D^{(r-1)}\alpha) \) and that
\[
V_{j_i,j_i+1-1}(D^{(r-1)}\alpha) \geq V(D^{(r-1)}\alpha^{(i)}) + \mathfrak{r}_i \left( (D^{(r-1)}\alpha)_{j_i+r-1} - (D^{(r-1)}\alpha)_{j_i} \right).
\]
The notation \( V_{a,b}(\cdot) \) may be recalled from (87). The above inequality, together with \( \Gamma_i(\alpha,\ell_i) \leq \delta_i \), allows us to deduce that
\[
V(D^{(r-1)}\alpha^{(i)}) \leq \mathfrak{r}_{i+1} \left\{ (D^{(r-1)}\alpha)_{j_{i+1}-1} - (D^{(r-1)}\alpha)_{j_i} \right\} + \mathfrak{r}_i \left\{ (D^{(r-1)}\alpha)_{\ell_i} - (D^{(r-1)}\alpha)_{j_{i+1}-1} \right\} + \delta_i.
\]
Using (97) and (98), it is now easy to see that the above inequality is the same as (96). This proves (96).

**Step V:** Next, we use the characterization in (96) to bound \( G_i(m,q,\ell) \) using Lemma B.3. Indeed, we can take \( \sigma = 1, t = \sqrt{m_i/(k+1)} \leq 1, n = n_i, \ell = \ell_i \) and \( \delta = \delta_i \) in Lemma B.3 to obtain
\[
G_i(m,q,\ell) \leq C_r \left( \sqrt{\frac{m_i}{k+1}} + \delta_i n_i^{(2r-1)/2} \right) \sqrt{\log(en_i)}
\]
\[
+ C_r \left( \frac{m_i}{k+1} \right)^{(2r-1)/(4r)} n_i^{(2r-1)/(4r)} \delta_i^{1/(2r)}
\]
for all \( 0 \leq i \leq k \). Here \( C_r \) is a constant that depends on \( r \) alone. This inequality, together with \( G(m,q,\ell) \leq \sum_{i=0}^k G_i(m,q,\ell) \), gives the following upper bound for \( G(m,q,\ell)/C_r \):
\[
\sum_{i=0}^k \sqrt{\frac{m_i}{k+1}} \sqrt{\log(en_i)} + \sum_{i=0}^k \delta_i n_i^{(2r-1)/2} \sqrt{\log(en_i)}
\]
\[
+ \sum_{i=0}^k \left( \frac{m_i n_i}{k+1} \right)^{(2r-1)/(4r)} \delta_i^{1/(2r)}.
\]
(99)

We now bound separately each of the three terms above. For the first term, note that by the Cauchy-Schwarz inequality and the fact that \( \sum_{i=0}^k m_i \leq 2(k+1) \), we get
\[
\sum_{i=0}^k \sqrt{\frac{m_i}{k+1}} \sqrt{\log(en_i)} \leq \sqrt{\sum_{i=0}^k \frac{m_i}{k+1} \sum_{i=0}^k \log(en_i)} \leq \sqrt{2(k+1) \log \frac{en}{k+1}}
\]
where we have also used concavity of the logarithm function to claim that \( \sum_{i=0}^k \log(en_i) \leq (k+1) \log \frac{en}{k+1} \). For the second term in (99), we write
\[
\sum_{i=0}^k \delta_i n_i^{(2r-1)/2} \sqrt{\log(en_i)} \leq \max_{0 \leq i \leq k} \left[ n_i^{(2r-1)/2} \sqrt{\log(en_i)} \right] \sum_{i=0}^k \delta_i
\]
\[
\leq 2\delta \max_{0 \leq i \leq k} \left[ n_i^{(2r-1)/2} \sqrt{\log(en_i)} \right]
\]
where we have used that \( \sum_{i=0}^{k} \delta_i = \delta \sum_{i=0}^{k} q_i/(k+1) \leq 2\delta \). Assumption (91) now gives

\[
\max_{0 \leq i \leq k} \left[ n_i^{(2r-1)/2} \sqrt{\log(en_i)} \right] \leq 2^r \left( \frac{n}{k+1} \right)^{(2r-1)/2} \sqrt{\log \frac{en}{k+1}}.
\]

We thus obtain

\[
\sum_{i=0}^{k} \delta_i n_i^{(2r-1)/2} \sqrt{\log(en_i)} \leq 2^{1+r} \delta \left( \frac{n}{k+1} \right)^{(2r-1)/2} \sqrt{\log \frac{en}{k+1}}.
\]

For the third term in (99), we use the standard Holder’s inequality \( (\sum_i \alpha_i \beta_i \leq (\sum_i \alpha_i^p)^{1/p}(\sum_i \beta_i^q)^{1/q} \) with \( p = 2r/(2r-1) \) and \( q = 2r \) to obtain

\[
\sum_{i=0}^{k} \left( \frac{m_i n_i}{k+1} \right)^{(2r-1)/(4r)} \delta_i^{1/(2r)} \leq \left( \sum_{i=0}^{k} \frac{m_i n_i}{k+1} \right)^{(2r-1)/(2r)} \left( \sum_{i=0}^{k} \delta_i \right)^{1/(2r)}
\]

\[
\leq 2^{1/(2r)} \delta^{1/(2r)} \left( \sum_{i=0}^{k} \frac{m_i}{k+1} \right)^{(2r-1)/(2r)} \sum_{i=0}^{k} n_i \leq 2^{(2r+1)/(4r)} \delta^{1/(2r)} n^{(2r-1)/(4r)}
\]

where, in the second inequality above, we used \( \sum_{i=0}^{k} \delta_i \leq 2\delta \) and the Cauchy-Schwarz inequality and, in the final inequality, we used \( \sum_{i=0}^{k} m_i \leq 2(k+1) \) and \( \sum_{i=0}^{k} n_i = n \). Putting the bounds for the three terms in (99) together, we obtain

\[
\frac{G(m, q, \ell)}{C_r} \leq \sqrt{2(k+1) \log \frac{en}{k+1}} + 2^{1+r} \delta \left( \frac{n}{k+1} \right)^{(2r-1)/2} \sqrt{\log \frac{en}{k+1}}
\]

\[
+ 2^{(2r+1)/(4r)} \delta^{1/(2r)} n^{(2r-1)/(4r)}
\]

(100)

which gives (note also that \( \delta \leq C_r \delta_r(\theta) \) by (92))

\[
G(m, q, \ell) \leq c_r \sqrt{n\Delta_r(\theta)},
\]

for a suitable constant \( c_r \) depending only on \( r \); note that \( \sqrt{a} + \sqrt{b} + \sqrt{c} \leq \sqrt{3} \sqrt{a+b+c} \) for \( a, b, c > 0 \). Combined with (94), this completes the proof of (86) when assumption (91) is true.

**Step VI:** Now we work with the situation when the assumption (91) is violated. Our basic idea here is that we will add indices to the set of knots \( j_1, \ldots, j_k \) to create a new set of indices which contains all the knots of \( \theta \) and which satisfies an assumption similar to (91). Specifically for every \( i \geq 1 \) for which \( n_i \) is strictly larger than \( 2n/(k+1) \), we add the indices

\[
J_i = \left\lfloor \frac{2n}{k+1} \right\rfloor, j_i + 2 \left\lfloor \frac{2n}{k+1} \right\rfloor, \ldots, j_i + A_i \left\lfloor \frac{2n}{k+1} \right\rfloor
\]
to the original set of knots, where \(|x|\) denotes the largest integer less than or equal to \(x\).

Here \(A_i\) is the integer part of the ratio of \(n_i\) to \(\lfloor 2n/(k+1) \rfloor\) and hence

\[
A_i \leq n_i \left\lfloor \frac{2n}{k+1} \right\rfloor^{-1}.
\]

Similarly, if \(n_0 \geq 2n/(k+1)\), then we add the indices \(2 - r + \lfloor 2n/(k+1) \rfloor, 2 - r + 2 \lfloor 2n/(k+1) \rfloor, \ldots, 2 - r + A_0 \lfloor 2n/(k+1) \rfloor\) to the original set of knots \(j_1, \ldots, j_k\) where again \(A_0 = n_0 \lfloor 2n/(k+1) \rfloor^{-1}\). This construction will create a set of indices \(j'_1, \ldots, j'_{k'}\) that contains all the original knots and which satisfy

\[
n'_i \leq \frac{2n}{k+1} \quad \text{for every } i = 0, \ldots, k'.
\] (101)

where \(n'_i\) are defined with respect to \(j'_1 < \cdots < j'_{k'}\) as \(n'_0 = j'_1 + r - 2\) and \(n'_i := j'_{i+1} - j'_i\) for \(i = 1, \ldots, k'.\) We now note that the number of these new indices, \(k'\), satisfies

\[
k' \leq k + \sum_{i=0}^{k} A_i \leq k + \left\lfloor \frac{2n}{k+1} \right\rfloor^{-1} \sum_{i=0}^{k} n_i = k + n \left\lfloor \frac{2n}{k+1} \right\rfloor^{-1} \leq 2k + 1
\]

where we have used \(\left\lfloor \frac{2n}{k+1} \right\rfloor \geq \frac{2n}{k+1} - 1 \geq \frac{n}{k+1}\). The inequality \(k' \leq 2k + 1\), along with (101), implies that

\[
n'_i \leq \frac{4n}{k'+1} \quad \text{for every } i = 0, \ldots, k'.
\] (102)

For these indices \(j'_1, \ldots, j'_{k'}\), we shall assign signs \(v'_1, \ldots, v'_{k'} \in \{-1, 0, 1\}\) in the following way. If \(j'_i\) is a knot (i.e., it is one of \(j_1, \ldots, j_k\)), then \(v'_i\) equals the sign of the knot \(j_i\). If \(j'_i\) is not a knot, then we assign \(v'_i\) to be the sign of the nearest knot that is to the right of \(j'_i\) (if there is no knot to the right of \(j'_i\), then we take \(v'_i\) to be zero).

We now go through the previous proof (which was under the case when assumption (91) is satisfied) with the set of indices \(j'_1, \ldots, j'_{k'}\) and signs \(v'_1, \ldots, v'_{k'}\). Instead of (91), we use the inequality (102) which has a slightly worse constant 4 instead of 2. This argument will end with an inequality similar to (100) (but with slightly different constants). Thus we obtain the following upper bound for \(G\):

\[
\frac{G^2}{C_r} \leq (k' + 1) \log \frac{en}{k'+1} + (\delta')^2 \left( \frac{n}{k'+1} \right)^{2r-1} \log \frac{en}{k'+1} + (\delta')^{1/r} n^{(2r-1)/(2r)}
\] (103)

for a constant \(C_r\) where

\[
\delta' := \left( \sum_{i=0}^{k'} (n'_i)^{1-2r} I\{v'_i \neq v'_{i+1}\} \right)^{1/2}.
\]

Now because \(k \leq k' \leq 2k + 1\), we have \(k + 1 \leq k' + 1 \leq 2(k + 1)\) and thus we can replace the \(k'\) on the right hand side of (103) by \(k\) by enlarging the constant \(C_r\) slightly. Finally,
to complete the proof, it suffices to observe that because of the construction of the set of indices \( j'_1, \ldots, j'_{k'} \) and the choice of the signs, we have

\[
(\delta')^2 = \sum_{i=0}^{k} \left[ \min \left( n_i, \left\lfloor \frac{2n}{k+1} \right\rfloor \right) \right] I\{r_i \neq r_{i+1}\} \\
\leq \sum_{i=0}^{k} n_i^{1-2r} I\{r_i \neq r_{i+1}\} \\
= n_0^{1-2r} + n_k^{1-2r} + \sum_{i=1}^{k-1} n_i^{1-2r} I\{r_i \neq r_{i+1}\} = \delta_r^2(\theta).
\]

This, along with replacing \( k' \) by \( k \) in (103), completes the proof of Theorem 2.2. \( \square \)

### B.3. Proof of Corollary 2.3

We prove Corollary 2.3 as a consequence of Theorem 2.2 and Theorem 2.1. The following lemma (proved in Subsection D.5) will be needed for this.

**Lemma B.4.** Fix \( r \geq 1 \) and \( n \geq r + 1 \). For every \( \theta \in \mathbb{R}^n \), there exists \( \eta \in \mathbb{R}^n \) such that

\[
\|D(\eta)\|_1 = 0 \quad \text{and} \quad \|\theta - \eta\|^2 \leq n^{2r-1}\|D^{(r)}\theta\|_1^2.
\]  

(104)

**Remark B.2.** When \( r = 1 \), the inequality in (104) is equivalent to

\[
\sum_{i=1}^{n} (\theta_i - \bar{\theta})^2 \leq n\|D\theta\|^2 = nV^2(\theta)
\]

which relates the variance of \( \theta_1, \ldots, \theta_n \) to the variation (here \( \bar{\theta} := (\theta_1 + \cdots + \theta_n)/n \)). Therefore Lemma B.4 can be seen as a relation between variance and variation for general \( r \geq 1 \).

We are now ready to prove Corollary 2.3.

**Proof of Corollary 2.3.** Recall from (10) that \( V^{(r)}(\theta^*) = n^{r-1}\|D^{(r)}\theta^*\|_1 \). We first consider the case when \( V^{(r)}(\theta^*) > 0 \). In this case, we use Lemma B.4 to claim the existence of \( \eta^* \in \mathbb{R}^n \) such that \( V^{(r)}(\eta^*) = 0 \) and

\[
\|\theta^* - \eta^*\|^2 \leq n^{2r-1}\|D^{(r)}\theta^*\|_1^2 = n\left(V^{(r)}(\theta^*)\right)^2.
\]  

(105)

Let now \( \theta \in \mathbb{R}^n \) be defined as

\[
\theta := \eta^* + \frac{V}{V^{(r)}(\theta^*)} (\theta^* - \eta^*).
\]
As $V^{(r)}(\eta^*) = 0$, it follows that $V^{(r)}(\theta) = V$. We deduce therefore that

$$\inf_{\alpha \in \mathbb{R}^n, V^{(r)}(\alpha) = V} \left\{ \frac{1}{n} \|\theta^* - \alpha\|^2 + C_r\sigma^2\Delta_r(\alpha) \right\}$$

is bounded from above by

$$\frac{1}{n} \|\theta^* - \theta\|^2 + C_r\sigma^2\Delta_r(\theta) = \frac{1}{n} \|\theta^* - \eta^*\|^2 \left(1 - \frac{V}{V^{(r)}(\theta^*)}\right)^2 + C_r\sigma^2\Delta_r(\theta)$$

$$\leq \left(V - V^{(r)}(\theta^*)\right)^2 + C_r\sigma^2\Delta_r(\theta)$$

where the last inequality above follows from (105). We now note that, by construction, $\theta$ satisfies the minimum length condition (13) with the same constant $c$ because $\theta^*$ does so. As a consequence, we have from (14) that

$$\Delta_r(\theta) \leq C_r(c) \frac{k + 1}{n} \log \frac{en}{k + 1}$$

where $C_r(c)$ depends on $r$ and $c$ alone and $k = k_r(\theta) = k_r(\theta^*)$. We have thus shown that (106) is bounded from above by

$$\left(V - V^{(r)}(\theta^*)\right)^2 + C_r(c)\sigma^2\frac{k + 1}{n} \log \frac{en}{k + 1}.$$

Inequality (17) then directly follows from Theorem 2.2.

We now assume that $V^{(r)}(\theta^*) = 0$. Here we have $k_r(\theta^*) = 0$ so that the second term on the right hand side of (17) becomes $C_r\sigma^2\log(en)$. Note also that because $V^{(r)}(\theta^*) = 0$ and $V \geq 0$, we can use Theorem 2.1. To complete the proof, we therefore only need to prove that

$$C_r \max \left( \left(\frac{\sigma^2 V^{1/r}}{n}\right)^{2r/(2r+1)}, \frac{\sigma^2}{n} \log(en) \right) \leq V^2 + \frac{C_r(2r+1)/(2r)}{n} \frac{\sigma^2}{n} \log(en).$$

To prove the above inequality, we may assume that

$$\left(\frac{\sigma^2 V^{1/r}}{n}\right)^{2r/(2r+1)} > C_r^{1/(2r)}\frac{\sigma^2}{n} \log(en)$$

for otherwise (107) is trivial. It is now straightforward to check that the above inequality is equivalent to

$$\frac{\sigma^2}{n} < \frac{V^2}{(\log en)^{2r+1} C_r^{-(2r+1)/(2r)}}.$$ 

From here, it is easy to show that

$$C_r \left(\frac{\sigma^2 V^{1/r}}{n}\right)^{2r/(2r+1)} \leq \frac{V^2}{(\log en)^{2r}} \leq V^2$$

which proves (107). This completes the proof of (17) when $V^{(r)}(\theta^*) = 0$. Inequality (18) trivially follows from (17).
B.4. Proof of Lemma 2.4

Here we provide the proof of the Lemma 2.4 which implies that the log $\frac{c\sigma^2}{n\sigma^2 + 1}$ appearing in our risk bounds cannot be completely removed. This proof uses the fact (71) as well as the precise characterization of the tangent cones of the set $K^{(r)}(V)$ (defined in (63)) given in Lemma C.3 (in Subsection C.2).

**Proof of Lemma 2.4.** Let $\theta^* = (0, \ldots, 0, 1, \ldots, 1)$ where the jump appears at the index $j := \lceil n/2 \rceil$ (i.e., $\theta_j^* = 1$ and $\theta_{j-1}^* = 0$). The estimator $\hat{\theta}_{V=1}^{(1)}$ is simply the least squares projection of $Y$ onto the closed convex set $K^{(1)}(V)$ (defined in (63)) with $V = 1$. The identity (71) therefore gives

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R(\hat{\theta}_{V=1}^{(1)}, \theta^*) = \frac{1}{n} \delta(T_{K^{(1)}(V)}(\theta^*)) .$$

The characterization of $T := T_{K^{(1)}(V)}(\theta^*)$ from Lemma C.3 implies, for this specific $\theta^*$, that $T$ consists of all vectors $\alpha \in \mathbb{R}^n$ for which

$$V_{1,j-1}(\alpha) + V_{j,n}(\alpha) \leq \alpha_j - 1$$

(108)

where $V_{1,j-1}(\alpha)$ and $V_{j,n}(\alpha)$ are defined as in (87). Now let $\mathfrak{M}$ consist of all vectors $\alpha \in \mathbb{R}^n$ which satisfy:

$$0 = \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{j-1}$$

and

$$\alpha_j \leq \alpha_{j+1} \leq \cdots \leq \alpha_n = 0 .$$

Note that there is no relation between $\alpha_{j-1}$ and $\alpha_j$ in the definition of $\mathfrak{M}$. Then, it is easy to check directly that every vector $\alpha \in \mathfrak{M}$ satisfies (108) so that

$$\delta(T) \geq \delta(\mathfrak{M}) .$$

This follows from the fact that $\delta(C_1) \leq \delta(C_2)$ whenever $C_1$ and $C_2$ are two closed convex cones such that $C_1 \subseteq C_2$ (this fact is stated, for example, Amelunxen et al. [1, Subsection 3.1]). Now if

$$\mathfrak{M}_1 := \{(\alpha_1, \ldots, \alpha_{j-1}) : 0 = \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{j-1}\}$$

then it is clear that $\delta(\mathfrak{M}) \geq \delta(\mathfrak{M}_1)$ so that we have

$$\delta(T) \geq \delta(\mathfrak{M}_1) .$$

We now use the fact that $\delta(\mathfrak{M}_1)$ is precisely known to satisfy (see Amelunxen et al. [1, Equation (D.12) in Subsection D.4])

$$\delta(\mathfrak{M}_1) = \frac{1}{2} \left( 1 + \frac{1}{2} + \cdots + \frac{1}{j-1} \right) .$$

This therefore implies (via $1 + (1/2) + \cdots + (1/m) \geq \log(m+1)$) that

$$\lim_{\sigma \downarrow 0} \frac{n}{\sigma^2} R(\hat{\theta}_{V=1}^{(1)}, \theta^*) = \delta(T_{K^{(1)}(V)}(\theta^*)) \geq \delta(\mathfrak{M}_1) \geq \frac{1}{2} \log(j) \geq \frac{1}{2} \log(n/2)$$

which proves Lemma 2.4. 

\[\square\]
B.5. Proof of Theorem 2.6

The following proposition is the key to proving Theorem 2.6. This proposition provides an upper bound for $D(\lambda \partial g(\theta))$ for a convex function $g$ (recall that $D(\cdot)$ is defined as in (76) and also that $\text{dist}(z, C) := \inf_{x \in C} \|z - x\|$ for $z \in \mathbb{R}^n$ a subset $C \subseteq \mathbb{R}^n$) in terms of the smaller quantity $D(\text{cone}(\partial g(\theta)))$. It is a generalization of Foygel and Mackey [13, Proposition 1]. Indeed, this latter result of [13] is the special case of Proposition B.5 under the additional assumption that $v_0 \in \partial g(\theta)$ (this assumption does not necessarily hold for $g(\theta) := n^{r-1}\|D^{(r)}\theta\|_1$ when $r \geq 2$).

**Proposition B.5.** Suppose $g : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\theta \in \mathbb{R}^n$. Suppose that the vector $v_0$ defined by
\[
  v_0 := \arg\min_{v \in \text{aff}(\partial g(\theta))} \|v\|. \tag{109}
\]
is a non-zero vector in $\mathbb{R}^n$. Then for every $z \in \mathbb{R}^n$,
\[
  \lambda(z) := \arg\min_{\lambda \geq 0} \text{dist}(z, \lambda \partial g(\theta)) = \arg\min_{\lambda \geq 0} \inf_{v \in \partial g(\theta)} \|z - \lambda v\| \tag{110}
\]
exists uniquely and, moreover, $\mathbb{E}\lambda(Z) < \infty$ where the expectation is taken with respect to $Z \sim N(0, I_n)$.

Further, let
\[
  \lambda^* := \mathbb{E}\lambda(Z) + \frac{2}{\|v_0\|} \quad \text{where } Z \sim N(0, I_n).
\]

Then for every $\lambda \geq \lambda^*$ and $v^* \in \partial g(\theta)$, we have
\[
  D(\lambda \partial g(\theta)) \leq 4 + \left( \sqrt{D(\text{cone}(\partial g(\theta)))} + \frac{4\|v^*\|}{\|v_0\|} + 2 + (\lambda - \lambda^*) \|v^*\| \right)^2. \tag{111}
\]

Before proving Proposition B.5, let us first show how Theorem 2.6 follows from Proposition B.5. The fact (77) and the bound (74) (which was proved in Subsection B.2) will be used in the proof below.

**Proof of Theorem 2.6.** Let $f(\theta) := \|D^{(r)}\theta\|_1$ and $g(\theta) := n^{r-1}f(\theta)$. Because $\hat{\theta}_\lambda^{(r)}$ equals the penalized estimator (75), Theorem A.3 gives
\[
  R(\hat{\theta}_\lambda^{(r)}, \theta^*) \leq \frac{\sigma^2}{n} D(\lambda \partial g(\theta^*)). \tag{112}
\]

We now use inequality (111) in Proposition B.5 to bound the right hand side above. Note that under the assumption $D^{(r)}\theta^* \neq 0$, we observed (after (25)) that $v_0$ is non-zero so that Proposition B.5 is applicable. This gives
\[
  D(\lambda \partial g(\theta^*)) \leq 4 + \left( \sqrt{D(\text{cone}(\partial g(\theta^*)))} + \frac{4\|v^*(g)\|}{\|v_0(g)\|} + 2 + (\lambda - \lambda^*) \|v^*(g)\| \right)^2.
\]
for every $\lambda \geq \lambda^*(g)$ where

$$\lambda^*(g) := \mathbb{E} \arg \min_{\lambda \geq 0} \text{dist}(Z, \lambda \partial g(\theta^*)) + \frac{2}{\|v_0(g)\|},$$

$v_0(g)$ is defined as in (109) and

$$v^*(g) := \arg \min_{v \in \partial g(\theta^*)} \|v\|.$$ 

Note that (111) holds for every $v^* \in \partial g(\theta^*)$ so it holds, in particular, for $v^*(g)$. Also note that $v^*(g)$ and $v_0(g)$ above are different from $v_0$ and $v^*$ in Theorem 2.6 which are all defined in terms of $f$. Now the relation $g = n^{r-1}f$ implies that

$$v_0 = \frac{v_0(g)}{n^{r-1}}, \quad v^* = \frac{v^*(g)}{n^{r-1}} \quad \text{and} \quad \lambda^*(g) = \lambda^*.$$ 

Note that $\lambda^*(g) = \lambda^*$ holds without any scaling factor because of the presence of the $n^{1-r}$ factor in the definition of $\lambda^*$ in (27). We have therefore proved that for every $\lambda \geq \lambda^*$, we have

$$\mathbf{D}(\lambda \partial g(\theta^*)) \leq 4 + \left(\sqrt{\mathbf{D}(\text{cone}(\partial g(\theta^*)))} + \frac{4\|v^*\|}{\|v_0\|} + 2 + \frac{(\lambda - \lambda^*)}{n^{1-r}}\|v^*\|\right)^2$$

$$\leq 20 + 4\mathbf{D}(\text{cone}(\partial g(\theta^*))) + \frac{64\|v^*\|^2}{\|v_0\|^2} + \frac{4(\lambda - \lambda^*)^2}{n^{2-2r}}\|v^*\|^2$$

where, in the last inequality, we used the elementary fact $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$. Note now that

$$\mathbf{D}(\text{cone}(\partial g(\theta^*))) = \mathbf{D}(\text{cone}(\partial f(\theta^*)))$$

so that, by inequality (77), we deduce that

$$\mathbf{D}(\text{cone}(\partial g(\theta^*))) \leq 1 + w^2(T_{K(\nu)}(V^*(\theta^*))).$$

The bound (74) then gives

$$\mathbf{D}(\text{cone}(\partial g(\theta^*))) \leq 1 + C_r^2 n \Delta_r(\theta^*).$$

Putting the above pieces together (and the fact that $\Delta_r(\theta^*) \geq 1/n$), we obtain

$$\mathbf{D}(\lambda \partial g(\theta^*)) \leq C_r n \Delta_r(\theta^*) + \frac{64\|v^*\|^2}{\|v_0\|^2} + \frac{4(\lambda - \lambda^*)^2}{n^{2-2r}}\|v^*\|^2$$

for every $\lambda \geq \lambda^*$. Combining this with (112) gives (28) and completes the proof of Theorem 2.6.

We now give the proof of Proposition B.5.
Proof of Proposition B.5. Note first that $\partial g(\theta)$ cannot contain the zero vector because we assumed that $v_0$ (defined by (109)) is non-zero. As a result, it follows from Rockafellar [39, Corollary 9.6.1] that

$$\text{cone}(\partial g(\theta)) := \bigcup_{\lambda \geq 0} (\lambda \partial g(\theta))$$

is closed (and, of course, a convex cone). It follows therefore that

$$\Pi_{\text{cone}(\partial g(\theta))}(z) := \arg\min_{u \in \text{cone}(\partial g(\theta))} \|z - u\|$$

exists uniquely. Let $\Pi_{\text{cone}(\partial g(\theta))}(z) := \lambda_1 v_1$ for some $\lambda_1 \geq 0$ and $v_1 \in \partial g(\theta)$. Then it is clear that $\lambda_1$ minimizes $\text{dist}(z, \lambda \partial g(\theta))$ over $\lambda \geq 0$. To prove that $\lambda_1$ is the unique minimizer, assume, if possible, the existence of $\lambda_2 \geq 0$ and $v_2 \in \partial g(\theta)$ such that $\lambda_1 v_1 = \lambda_2 v_2$. Note now that because $\text{aff}(\partial g(\theta))$ is an affine set, the vector $v_0$ defined by (109) (which is the projection of the zero vector onto $\text{aff}(\partial g(\theta))$) satisfies the orthogonality property:

$$\langle v - v_0, v_0 \rangle = 0 \quad \text{for every } v \in \partial g(\theta). \quad (113)$$

In particular, we have $\langle v, v_0 \rangle = \|v_0\|^2$ for every $v \in \partial g(\theta)$. Applying this to $v = v_1$ and $v = v_2$, we obtain that

$$\lambda_1 \|v_0\|^2 = \langle \lambda_1 v_1, v_0 \rangle = \langle \lambda_2 v_2, v_0 \rangle = \lambda_2 \|v_0\|^2$$

which implies that $\lambda_1 = \lambda_2$. This proves therefore that there is a unique $\lambda_1 \geq 0$ for which $\Pi_{\text{cone}(\partial g(\theta))}(z) \in \lambda_1 \partial g(\theta)$ and this $\lambda_1$ clearly is equal to $\lambda(z)$ defined in (110).

To prove that $\mathbb{E}\lambda(Z) < \infty$ for $Z \sim N_n(0, I_n)$, we write $\Pi_{\text{cone}(\partial g(\theta))} z = \lambda(z) v(z)$ for some $v(z) \in \partial g(\theta)$ and use (113) to obtain

$$\lambda(z) = \frac{1}{\|v_0\|^2} \langle \lambda(z) v(z), v_0 \rangle = \frac{1}{\|v_0\|^2} \langle \Pi_{\text{cone}(\partial g(\theta))}(z), v_0 \rangle \leq \frac{\|\Pi_{\text{cone}(\partial g(\theta))}(z)\|}{\|v_0\|}$$

where the last inequality follows from the Cauchy-Schwarz inequality. The standard fact that the projection onto a closed convex cone reduces norm gives $\|\Pi_{\text{cone}(\partial g(\theta))}(z)\| \leq \|z\|$ so that $\lambda(z) \leq \|z\|/\|v_0\|$ which implies obviously that $\mathbb{E}\lambda(Z) < \infty$ when $Z \sim N_n(0, I_n)$.

Let us now proceed to prove (111). The first step for this is to observe that the map $z \mapsto \lambda(z) = \arg\min_{\lambda \geq 0} \text{dist}(z, \lambda \partial g(\theta))$ is Lipschitz with parameter $1/\|v_0\|$ i.e.,

$$|\lambda(z_1) - \lambda(z_2)| \leq \frac{\|z_1 - z_2\|}{\|v_0\|} \quad \text{for every } z_1, z_2 \in \mathbb{R}^n. \quad (114)$$

To see this, fix $z_1, z_2 \in \mathbb{R}^n$ and let $\Pi_{\text{cone}(\partial g(\theta))}(z_i) = \lambda(z_i) v_i$ for two vectors $v_1, v_2 \in \partial g(\theta)$. Then, by the contraction property for projections on closed convex cones, we have

$$\|z_1 - z_2\| \geq \|\lambda(z_1) v_1 - \lambda(z_2) v_2\|$$

$$= \|(\lambda(z_1) - \lambda(z_2)) v_0 + \lambda(z_1)(v_1 - v_0) - \lambda(z_2)(v_2 - v_0)\|$$

$$= \|(\lambda(z_1) - \lambda(z_2)) v_0\| + \|\lambda(z_1)(v_1 - v_0) - \lambda(z_2)(v_2 - v_0)\|$$
where the last equality follows from the orthogonality property (113). Because the last term above is nonnegative, the inequality (114) follows.

The Lipschitz property of $z \mapsto \lambda(z)$ proved above implies, by standard Gaussian concentration, that

$$\mathbb{P}\left\{ |\lambda(z) - \mathbb{E}\lambda(Z)| < \frac{2}{\|v_0\|} \right\} \geq 1 - 2e^{-2}.$$  

Let $E := \{ z \in \mathbb{R}^n : |\lambda(z) - \mathbb{E}\lambda(Z)| < 2/\|v_0\| \}$ so that $\mathbb{P}\{z \in E\} \geq 1 - 2e^{-2}$. Note that $0 \leq \lambda(z) < \lambda^*$ when $z \in E$. This implies that for every $\lambda \geq \lambda^*$ and vectors $v, v^* \in \partial g(\theta)$, we have (by convexity of the subdifferential $\partial g(\theta)$)

$$\frac{\lambda(z)}{\lambda} v + \left(1 - \frac{\lambda(z)}{\lambda}\right) v^* \in \partial g(\theta).$$

In particular, this is true with $v = v(z)$ where $\Pi_{\text{cone}(\partial g(\theta))}(z) := \lambda(z)v(z)$. As a result,

$$\text{dist}(z, \lambda \partial g(\theta)) \leq \|z - \lambda(z)v(z) - (\lambda - \lambda(z))v^*\|$$

$$\leq \|z - \lambda(z)v(z)\| + (\lambda - \lambda(z)) \|v^*\|$$

$$= \text{dist}(z, \text{cone}(\partial g(\theta))) + (\lambda - \lambda(z)) \|v^*\|.$$

Now, again for $z \in E$, we have $\lambda(z) > \mathbb{E}\lambda(Z) - 2/\|v_0\|$ so that

$$\lambda - \lambda(z) \leq \lambda - \mathbb{E}\lambda(Z) + \frac{2}{\|v_0\|} = \lambda - \lambda^* + \frac{4}{\|v_0\|}.$$

We have therefore proved that

$$\text{dist}(z, \lambda \partial g(\theta)) \leq \text{dist}(z, \text{cone}(\partial g(\theta))) + (\lambda - \lambda^*) \|v^*\| + \frac{4\|v_0\|}{\|v^*\|}$$

for $z \in E$ which further implies that the probability

$$\mathbb{P}\left\{ \frac{1}{2} \text{dist}(Z, \lambda \partial g(\theta)) - \frac{1}{2} \text{dist}(Z, \text{cone}(\partial g(\theta))) > \frac{2\|v^*\|}{\|v_0\|} + \frac{1}{2}(\lambda - \lambda^*) \|v^*\| \right\}$$

is bounded from above by $2e^{-2}$. We now use Foygel and Mackey [13, Lemma 4] to claim that

$$\mathbb{E}\text{dist}(Z, \lambda \partial g(\theta)) - \mathbb{E}\text{dist}(Z, \text{cone}(\partial g(\theta))) \leq (\lambda - \lambda^*) \|v^*\| + \frac{4\|v^*\|}{\|v_0\|}$$

$$+ 2\sqrt{-2 \log(1 - 2e^{-2})}$$

$$\leq (\lambda - \lambda^*) \|v^*\| + \frac{4\|v^*\|}{\|v_0\|} + 2. \quad (115)$$
To convert this into a bound on $\mathbb{E} \text{dist}^2(Z, \lambda \partial g(\theta))$, we use the fact that $z \mapsto \text{dist}(z, \lambda \partial g(\theta))$ is a 1-Lipschitz function so that again by standard Gaussian concentration, we have

$$\text{var}(\text{dist}(Z, \lambda \partial g(\theta))) = \int_0^\infty \mathbb{P}\left\{|\text{dist}(Z, \lambda \partial g(\theta)) - \mathbb{E}\text{dist}(Z, \lambda \partial g(\theta))| \geq \sqrt{t}\right\} dt \leq 2 \int_0^\infty e^{-t/2} dt = 4.$$ 

This gives

$$D(\lambda \partial g(\theta)) = \mathbb{E}\text{dist}^2(Z, \lambda \partial g(\theta)) = (\mathbb{E}\text{dist}(Z, \lambda \partial g(\theta)))^2 + \text{var}(\text{dist}(Z, \lambda \partial g(\theta))) \leq (\mathbb{E}\text{dist}(Z, \lambda \partial g(\theta)))^2 + 4,$$

which, combined with (115) and the elementary fact

$$\mathbb{E}\text{dist}(Z, \text{cone}(\partial g(\theta))) \leq \sqrt{\mathbb{E}\text{dist}^2(Z, \text{cone}(\partial g(\theta)))} = \sqrt{D(\text{cone}(\partial g(\theta)))},$$

completes the proof of Proposition B.5.

### B.6. Proofs of Corollary 2.8, Lemma 2.9 and Corollary 2.10

In this subsection, we shall provide the proofs of Corollary 2.8, Lemma 2.9 and Corollary 2.10.

**Proof of Corollary 2.8.** Corollary 2.8 is a simple consequence of Theorem 2.6 and Lemma 2.7. Indeed, Lemma 2.7 states that for $r = 1$, we have $v^* = v_0$ and that

$$\|v^*\|^2 = \frac{1}{n_0} + \frac{1}{n_k} + 4 \sum_{i=1}^{k-1} \frac{I\{r_i \neq r_{i+1}\}}{n_i} \leq 4 \sum_{i=0}^{k} \frac{I\{r_i \neq r_{i+1}\}}{n_i}.$$ 

Using this in the right hand side of (28), we get

$$R(\hat{\theta}^{(1)}_\lambda, \theta^*) \leq C_1 \sigma^2 \Delta_1(\theta^*) + \frac{64 \sigma^2}{n} + \frac{16 \sigma^2}{n} (\lambda - \lambda^*)^2 \sum_{i=0}^{k} \frac{I\{r_i \neq r_{i+1}\}}{n_i}$$

which implies (32) as $\Delta_1(\theta^*) \geq 1/n$. To prove (33), we further bound the right hand side above under the minimum length condition (13) by noting that $\Delta_1(\theta^*) \leq C(c) \frac{k+1}{n} \log \frac{en}{k+1}$ and also that

$$\sum_{i=0}^{k} \frac{I\{r_i \neq r_{i+1}\}}{n_i} \leq \frac{k + 1}{cn} \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\}.$$ 

\qed
Proof of Lemma 2.9. From the formula (27) for $\lambda^*$, it is clear that we need to bound both the terms $\mathbb{E}\lambda_{\theta^*}(Z)$ and $2/\|v_0\|$ from above in order to upper bound $\lambda^*$. For bounding $1/\|v_0\|$ from above, we use (29) to obtain
\[
\|v_0\|^2 = \frac{1}{n_0} + \frac{1}{n_k} + 4\sum_{i=1}^{k-1} \frac{I\{r_i \neq r_{i+1}\}}{n_i} \geq \frac{1}{n} \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\}
\]
where, in the last inequality above, we used $n_i \leq n$. This gives
\[
\frac{2}{\|v_0\|} \leq \sqrt{\frac{4n}{\sum_{i=0}^{k} I\{r_i \neq r_{k+1}\}}}.
\]
(116)

We shall now bound $\mathbb{E}\lambda_{\theta^*}(Z)$. Note that $\theta^* \in \mathbb{R}^n$ is such that $D\theta^* \neq 0$. Throughout this proof, $f(\theta) := \|D\theta\|_1$. As observed in the proof of Proposition B.5, $\text{cone}(\partial f(\theta^*))$ is a closed convex cone and for every $z \in \mathbb{R}^n$, we have
\[
\Pi_{\text{cone}(\partial f(\theta^*))}(z) = \lambda_{\theta^*}(z)v(z)
\]
(117)
for some vector $v(z) \in \partial f(\theta^*)$. Suppose now that $\theta^*$ has the $k$ jumps $2 \leq j_1 < \cdots < j_k \leq n$ with associated signs $r_1, \ldots, r_k$. Also let $j_0 = 1, j_{k+1} = n + 1$ and $r_0 = r_{k+1} = 0$. Then by the characterization of $\partial f(\theta^*)$ from Proposition 2.5, we have
\[
\sum_{u=j_i}^{n} v_u(z) = r_i \quad \text{for every } i = 0, \ldots, k + 1
\]
where $(v_1(z), \ldots, v_n(z))$ are the components of the vector $v(z)$. This implies, via (117), that
\[
\lambda_{\theta^*}(z) = \sum_{u=j_i}^{n} (\Pi z)_u
\]
where $(\Pi z)_1, \ldots, (\Pi z)_n$ denote the components of $\Pi z := \Pi_{\text{cone}(\partial f(\theta^*))}(z)$. As a consequence (by subtracting the above identity for $i$ from the corresponding identity for $i+1$), we obtain
\[
(r_i - r_{i+1}) \lambda_{\theta^*}(z) = (\Pi z)_{j_i} + \cdots + (\Pi z)_{j_{i+1} - 1}
\]
for every $i = 0, \ldots, k$. Multiplying both sides above by $(r_i - r_{i+1})$, we get
\[
(r_i - r_{i+1})^2 \lambda_{\theta^*}(z) = (r_i - r_{i+1}) ((\Pi z)_{j_i} + \cdots + (\Pi z)_{j_{i+1} - 1})
\]
for every $i = 0, \ldots, k$. Adding these for $i = 0, \ldots, k$, we obtain
\[
\lambda_{\theta^*}(z) \sum_{i=0}^{k} (r_i - r_{i+1})^2 = \sum_{i=0}^{k} (r_i - r_{i+1}) ((\Pi z)_{j_i} + \cdots + (\Pi z)_{j_{i+1} - 1})
\]
We now use the important identity (81) which gives
\[ \Pi z = \Pi_{\text{cone}(\partial f(\theta^*))}(z) = z - \Pi_{K^{(1)}(V^*)}(z) \]
where \( V^* := \|D\theta^*\|_1 \). This gives (below we write \( \Pi_T z \) as shorthand for \( \Pi_{K^{(1)}(V^*)}(z) \))
\[
\lambda_{\theta^*}(z) \sum_{i=0}^{k} (r_i - r_{i+1})^2 = \sum_{i=0}^{k} (r_i - r_{i+1}) \left( z_{j_i} + \cdots + z_{j_i+1-1} \right) \\
- \sum_{i=0}^{k} (r_i - r_{i+1}) \left( (\Pi_T z)_{j_i} + \cdots + (\Pi_T z)_{j_i+1-1} \right). 
\]
This equality holds for all vectors \( z \in \mathbb{R}^n \). Applying this to \( Z \sim N(0, I_n) \) and taking expectations on both sides with respect to \( Z \), we obtain
\[
\mathbb{E} \lambda_{\theta^*}(Z) \sum_{i=0}^{k} (r_i - r_{i+1})^2 = - \sum_{i=0}^{k} (r_i - r_{i+1}) \left( (\mathbb{E} \Pi_T Z)_{j_i} + \cdots + (\mathbb{E} \Pi_T Z)_{j_i+1-1} \right). 
\]
Using the Cauchy-Schwarz inequality on the right hand side above, we deduce
\[
\mathbb{E} \lambda_{\theta^*}(Z) \sum_{i=0}^{k} (r_i - r_{i+1})^2 \leq \mathbb{E} \|\Pi_T Z\|_2 \sqrt{\sum_{i=0}^{k} (r_i - r_{i+1})^2 n_i} \\
\leq 2 \mathbb{E} \Pi_T Z \sqrt{\sum_{i=0}^{k} n_i I\{r_i \neq r_{i+1}\}}.
\]
where we used the fact that \(|r_i - r_{i+1}| \leq 2\) when \( r_i \neq r_{i+1} \). This gives (also using \(|r_i - r_{i+1}| \geq 1\) when \( r_i \neq r_{i+1} \) on the left hand side)
\[
\mathbb{E} \lambda_{\theta^*}(Z) \leq 2 \mathbb{E} \Pi_T Z \sqrt{\frac{\sum_{i=0}^{k} n_i I\{r_i \neq r_{i+1}\}}{\sum_{i=0}^{k} I\{r_i \neq r_{i+1}\}}}.
\] (118)
To bound \( \|\Pi_T Z\| \), we use Jensen’s inequality and inequality (69) (recall the notions of statistical dimension and Gaussian width from Subsection A) to obtain
\[
\|\Pi_T Z\|^2 \leq \mathbb{E} \|\Pi_T Z\|^2 = \delta(T) \leq 1 + w^2(T) = 1 + w^2(T_{K^{(1)}(V^*)})
\]
Inequality (74) now gives
\[
w^2(T_{K^{(1)}(V^*)}) \leq C_1^2 n \Delta_1(\theta^*)
\]
for a positive constant \( C_1^2 \). This implies (note that \( \Delta_1(\theta^*) \geq 1/n \)) that
\[
\|\Pi_T Z\| \leq C \sqrt{n \Delta_1(\theta^*)}.
\]
Combining this with (118), we get
\[ \mathbb{E}_{\theta^*}(Z) \leq C\sqrt{n\Delta_1(\theta^*)} \frac{\sum_{i=0}^{k} n_i I\{r_i \neq r_{i+1}\}}{\left(\sum_{i=0}^{k} I\{r_i \neq r_{i+1}\}\right)^2}. \]

We now use the length condition (37). Under this condition, we know that
\[ n\Delta_1(\theta^*) \leq C(c_1)(k + 1) \log \left(\frac{en}{k + 1}\right). \]

Using this (and the fact that \( n_i \leq c_2 n/(k + 1) \)), we obtain
\[ \mathbb{E}_{\theta^*}(Z) \leq C(c_1, c_2) \sqrt{n \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\} \log \frac{en}{k + 1}}. \]

The proof of (38) is now completed by combining the above bound with (116).

Proof of Corollary 2.10. Suppose \( \lambda \) is as in (39) for \( \Gamma \geq C^*(c_1, c_2) \) (where \( C^*(c_1, c_2) \) comes from Lemma 2.9). Then, by Lemma 2.9, \( \lambda \geq \lambda^* \). We can therefore apply Corollary 2.8 (specifically, inequality (33) as \( \theta^* \) satisfies the length condition (37) which implies the minimum length condition with constant \( c_1 \)) to obtain
\[ R(\hat{\theta}^{(1)}_{\lambda}, \theta^*) \leq C(c_1)\sigma^2 \left( \frac{k + 1}{n} \log \frac{en}{k + 1} + (\lambda - \lambda^*)^2 \frac{k + 1}{n^2} \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\} \right) \]
\[ \leq C(c_1)\sigma^2 \left( \frac{k + 1}{n} \log \frac{en}{k + 1} + \lambda^2 \frac{k + 1}{n^2} \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\} \right) \]
(119)

for a constant \( C(c_1) \) depending only on \( c_1 \). In the last inequality above, we used the trivial fact that \((\lambda - \lambda^*)^2 \leq \lambda^2\). Plugging in the value of \( \lambda \) from (39) in the bound (119), we obtain (40).

We shall now prove (42) assuming that \( \lambda \) is as in (41) with \( \Gamma \geq C^*(c_1, c_2) \). For this, note first that (119) holds for this \( \lambda \) as well because \( \lambda \geq \lambda^* \). Plugging in \( \lambda = \Gamma \sqrt{n \log(en)} \) in (119), we obtain
\[ R(\hat{\theta}^{(1)}_{\lambda}, \theta^*) \leq C(c_1)\sigma^2 \left( \frac{k + 1}{n} \log \frac{en}{k + 1} + \Gamma^2 (\log(en)) \frac{k + 1}{n} \sum_{i=0}^{k} I\{r_i \neq r_{i+1}\} \right). \]

The trivial bound \( \log(en/(k + 1)) \leq \log(en) \) now gives (41). The proof of Corollary 2.10 is complete.
B.7. Proofs of Corollary 2.11, Lemma 2.12 and Corollary 2.13

In this subsection, we provide the proofs of Corollary 2.11, Lemma 2.12 and Corollary 2.13.

Proof of Corollary 2.11. Corollary 2.11 is a simple consequence of Theorem 2.6 and Lemma 2.7. Indeed, plugging in the lower bound on $\|v_0\|$ from (30) and the upper bound on $\|v^*\|$ from (31) in inequality (28), we obtain

$$R(\hat{\theta}_r^{(r)}, \theta^*) \leq C_r \sigma^2 \Delta_r(\theta^*) + C_r(c) \frac{\sigma^2}{n} (k+1)^{2r} + C_r(c) \sigma^2 (\lambda - \lambda^*)^2 \frac{(k+1)^2}{n^2}. $$

for a constant $C_r(c)$ depending only on $c$ ($c$ appears in the minimum length condition (13)). From here, inequality (46) immediately follows from the observation that $\Delta_r(\theta^*) \leq C_r(c) \frac{k+1}{n} \log \frac{en}{k+1}$ under the minimum length condition.

Proof of Lemma 2.12. Recall that

$$\lambda^* = n^{1-r} \left( \mathbb{E} \lambda_{\theta^*}(Z) + \frac{2}{\|v_0\|} \right)$$

with

$$\lambda_{\theta^*}(z) := \arg\min_{\lambda \geq 0} \inf_{v \in \partial f(\theta^*)} \|z - \lambda v\|$$

where $f(\theta) := \|D^{(r)} \theta\|_1$ and we have assumed that $D^{(r)} \theta^* \neq 0$. To bound $\lambda^*$ from above, we therefore need to bound both the terms $\mathbb{E} \lambda_{\theta^*}(Z)$ and $2/\|v_0\|$ from above. To bound $2/\|v_0\|$, we simply used inequality (30) which gives

$$\frac{2}{\|v_0\|} \leq C_r n^{1-r/2} \quad (120)$$

for a constant $C_r$. The main task therefore is to bound $\mathbb{E} \lambda_{\theta^*}(Z)$. We follow a strategy similar to that employed in the proof of Lemma 2.9. As observed in the proof of Proposition B.5, $\text{cone}(\partial f(\theta^*))$ is a closed convex cone (because $D^{(r)} \theta^* \neq 0$) and for every $z \in \mathbb{R}^n$, we can write

$$\Pi_{\text{cone}(\partial f(\theta^*))}(z) = \lambda_{\theta^*}(z)v(z) \quad (121)$$

for some vector $v(z) \in \partial f(\theta^*)$. Suppose now that $\theta^*$ has the $k$ knots (or order $r$): $2 \leq j_1 < \cdots < j_k \leq n - r + 1$ with associated signs $r_1, \ldots, r_k$. Then by the characterization of $\partial f(\theta^*)$ from Proposition 2.5 (specifically using (24) with $j = j_k + r - 1$), we obtain

$$\sum_{i=j_k+r-1}^{n} \binom{i-j_k}{r-1} v_i(z) = r_k$$

where $v_1(z), \ldots, v_n(z)$ are the components of the vector $v(z)$. This implies, via (121), that

$$\sum_{i=j_k+r-1}^{n} \binom{i-j_k}{r-1} (\Pi z)_i = r_k \lambda_{\theta^*}(z)$$

Proof of Corollary 2.11, Lemma 2.12 and Corollary 2.13
where \((\Pi z)_1, \ldots, (\Pi z)_n\) denote the components of \(\Pi z := \Pi_{\text{cone}(\partial f(\theta^*))}(z)\). Using (81), we can write
\[
\Pi z = \Pi_{\text{cone}(\partial f(\theta^*))}(z) = z - \Pi_{\text{T}_K^{(r)}}(z).
\]
We thus obtain (using \(\Pi T z\) as shorthand for \(\Pi_{\text{T}_K^{(r)}}(V^*)\))
\[
\tau_k \lambda^*_\theta(z) = \sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right) z_i - \sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right) (\Pi T z)_i.
\]
Applying this to \(Z \sim N_n(0, I_n)\), we get
\[
\tau_k E \lambda^*_\theta(Z) = \sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right) (E \Pi T z)_i.
\]
so that
\[
E \lambda^*_\theta(Z) = \left|\sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right) (E \Pi T z)_i\right|.
\]
By the Cauchy-Schwarz inequality, we now get
\[
(E \lambda^*_\theta(Z))^2 \leq \left[\sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right)^2\right] \left[\sum_{i=j_k+r-1}^{n} ((E \Pi T Z)_i)^2\right]
\leq \left[\sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right)^2\right] \|E \Pi T Z\|^2
\leq \left[\sum_{i=j_k+r-1}^{n} \left(\frac{i-j_k}{r-1}\right)^2\right] E \|\Pi T Z\|^2.
\]
Note now that for every \(i = j_k + r - 1, \ldots, n\), clearly
\[
\left(\frac{i-j_k}{r-1}\right) \leq \left(\frac{n}{r-1}\right) \leq n^{r-1}.
\]
As a result, we have
\[
(E \lambda^*_\theta(Z))^2 \leq n^{2r-2}(n-j_k-r+2)E \|\Pi T Z\|^2.
\]
Noting that \(n_k = n - r + 2 - j_k\), we have proved that
\[
(E \lambda^*_\theta(Z))^2 \leq n^{2r-2}n_k E \|\Pi T Z\|^2.
\]
Inequality (69) (recall the notions of statistical dimension and Gaussian width from Subsection A) now gives
\[
E \|\Pi T Z\|^2 = \delta(T) \leq 1 + w^2(T) = 1 + w^2(T_{K^{(r)}}(V^*)).
\]
Using inequality (74), we get
\[ w^2(T_{K(V)}^{\text{r}}) \leq C_r^2 n \Delta_r(\theta^*) \]
for a positive constant $C_r^2$. We have therefore proved that
\[ \mathbb{E} \lambda_{\theta^*}(Z) \leq n^{-1} \sqrt{n_k \mathbb{E} \| \Pi_r Z \|^2} \leq n^{-1} \sqrt{n_k (1 + C_r^2 n \Delta_r(\theta^*))}. \]
We now invoke the length condition (48). Under this condition, we first have
\[ n \Delta_r(\theta^*) \leq C_r (c_1, c_2) (k + 1) \log \frac{en}{k + 1} \]
and also $n_k \leq c_2 n / (k + 1)$ so that
\[ \mathbb{E} \lambda_{\theta^*}(Z) \leq C_r (c_1, c_2) n^{-1} \sqrt{\frac{n}{k + 1} (k + 1) \log \frac{en}{k + 1}} = C_r (c_1, c_2) n^{-1} \sqrt{\frac{en}{k + 1}}. \]
Combining this with (120), we get
\[ \lambda^* = n^{1-r} \left( \mathbb{E} \lambda_{\theta^*}(Z) + \frac{2}{\| \nu_0 \|} \right) \leq n^{1-r} \left( C_r (c_1, c_2) n^{-1} \sqrt{\frac{en}{k + 1}} + C_r n^{-1/2} \right) \leq C^*_r (c_1, c_2) \sqrt{n \log \frac{en}{k + 1}}. \]
This finishes the proof of Lemma 2.12.

**Proof of Corollary 2.13.** Suppose $\lambda$ is as in (50) for $\Gamma \geq C^*_r (c_1, c_2)$ (where $C^*_r (c_1, c_2)$ comes from Lemma 2.12). Then, by Lemma 2.12, $\lambda \geq \lambda^*$. We can therefore apply Corollary 2.11 (note that $\theta^*$ satisfies the length condition (48) which implies the minimum length condition with constant $c_1$) to obtain
\[ R(\hat{\theta}_1^{(l)}, \theta^*) \leq C_r (c_1) \sigma^2 \left( \frac{k + 1}{n} \log \frac{en}{k + 1} + \frac{(k + 1)^{2r}}{n} + (\lambda - \lambda^*)^2 \frac{(k + 1)^{2r}}{n^2} \right) \leq C_r (c_1) \sigma^2 \left( \frac{k + 1}{n} \log \frac{en}{k + 1} + \frac{(k + 1)^{2r}}{n} + \lambda^2 \frac{(k + 1)^{2r}}{n^2} \right) \]
for a constant $C_r (c_1)$ depending only on $r$ and $c_1$. In the last inequality above, we used the trivial fact that $(\lambda - \lambda^*)^2 \leq \lambda^2$. Plugging in the value of $\lambda$ from (50) in the bound above, we obtain
\[ R(\hat{\theta}_1^{(l)}, \theta^*) \leq C_r (c_1) \sigma^2 \left( \frac{k + 1}{n} \log \frac{en}{k + 1} + \frac{(k + 1)^{2r}}{n} + \frac{\Gamma^2 (k + 1)^{2r}}{n^2} \log \frac{en}{k + 1} \right) \leq C_r (c_1) \sigma^2 (2 + \Gamma^2) \frac{(k + 1)^{2r}}{n} \log \frac{en}{k + 1} \]
which proves (51). We shall now prove (53) assuming that $\lambda$ is as in (52) with $\Gamma \geq C^*(c_1, c_2)$. For this, note first that (122) holds for this $\lambda$ as well because $\lambda \geq \lambda^*$. Plugging in $\lambda = \Gamma \sqrt{n \log(en)}$ in (122), we obtain

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

$$\leq C_r(c_1) \sigma^2 (2 + \Gamma^2) \frac{(k + 1)^2r}{n} (\log(en))$$

which proves (52) and completes the proof of Corollary 2.13. \qed

\section*{B.8. Proof of Lemma 2.14}

The proof of Lemma 2.14, which deals with the case when $D^{(r)} \theta^* = 0$, is provided here.

\textbf{Proof of Lemma 2.14.} Let $f(\theta) := \|D^{(r)} \theta\|_1$ and $g(\theta) := n^{r-1} f(\theta)$. The estimator $\hat{\theta}_\lambda^{(r)}$ is then given by

$$\hat{\theta}_\lambda^{(r)} = \arg\min_{\theta \in \mathbb{R}^n} \left( \frac{1}{2} \| Y - \theta \|^2 + \sigma \lambda g(\theta) \right).$$

The risk result (A.3) gives

$$R(\hat{\theta}_\lambda^{(r)}, \theta^*) \leq \frac{\sigma^2}{n} D(\lambda \partial g(\theta^*)) = \frac{\sigma^2}{n} D(n^{r-1} \lambda \partial f(\theta^*)). \tag{123}$$

Because $D^{(r)} \theta^* = 0$, the subdifferential of $f$ at $\theta^*$ consists precisely of all vectors $v \in \mathbb{R}^n$ for which

$$\sum_{i=1}^n \left( \begin{array}{c} r + i - 1 \\ r - 1 \end{array} \right) v_i = 0 \quad \text{for } 1 \leq j \leq r \tag{124}$$

and

$$\max_{r < j \leq n} \left| \sum_{i=j}^n \left( \begin{array}{c} r + i - 1 \\ r - 1 \end{array} \right) v_i \right| \leq 1.$$

This is a consequence of the characterization of the subdifferential given in Proposition 2.5.

Now let $S_r$ denote the set consisting of all vectors $v \in \mathbb{R}^n$ such that (124) holds. Clearly $S_r$ is a subspace in $\mathbb{R}^n$ of dimension exactly equal to $n - r$. Let $\Pi_{S_r}$ denote the projection matrix onto $S_r$ and let

$$\lambda(z) := n^{1-r} \max_{r < j \leq n} \left| \sum_{i=j}^n \left( \begin{array}{c} r + i - 1 \\ r - 1 \end{array} \right) (\Pi_{S_r} z)_i \right| \quad \text{for } z \in \mathbb{R}^n.$$
For $Z \sim N(0, I_n)$, we can write
\[
D(n^{-1}\lambda \partial f(\theta^*)) = \mathbb{E}\text{dist}^2(Z, n^{-1}\lambda \partial f(\theta^*))
= \mathbb{E}\text{dist}^2(Z, n^{-1}\lambda \partial f(\theta^*))I\{\lambda(Z) \leq \lambda\}
+ \mathbb{E}\text{dist}^2(Z, n^{-1}\lambda \partial f(\theta^*))I\{\lambda(Z) > \lambda\}
\]
From the characterization of $\partial f(\theta^*)$ given above, it is clear that when $\lambda(Z) \leq \lambda$, the vector $\Pi_S Z$ belongs to $n^{-1}\lambda \partial f(\theta^*)$. On the other hand, the zero vector always belongs to $\partial f(\theta^*)$ (note that we are working under the assumption that $D^{(r)}\theta^* = 0$). This allows us to deduce that
\[
D(n^{-1}\lambda \partial f(\theta^*)) \leq \mathbb{E}\|Z - \Pi_S Z\|^2 + \mathbb{E}\|Z\|^2 I\{\lambda(Z) > \lambda\}.
\]
Because $S_r$ is a subspace of dimension $n - r$, the first term above equals $r$. For the second term, we use Cauchy-Schwarz inequality (and the elementary fact that $\mathbb{E}\|Z\|^4 = n^2 + 2n$) to obtain
\[
D(n^{-1}\lambda \partial f(\theta^*)) \leq r + \sqrt{n^2 + 2n} \mathbb{P}\{\lambda(Z) > \lambda\}.
\]
To bound $\mathbb{P}\{\lambda(Z) > \lambda\}$, we write (via the union bound)
\[
\mathbb{P}\{\lambda(Z) > \lambda\} \leq \sum_{r < j \leq n} \mathbb{P}\left\{\left|\sum_{i=j}^n \binom{r + i - j - 1}{r - 1} (\Pi_S Z)_i\right| > n^{-1}\lambda\right\}.
\]
For each fixed $r < j \leq n$, the random variable
\[
\sum_{i=j}^n \binom{r + i - j - 1}{r - 1} (\Pi_S Z)_i
\]
is easily seen to be normally distributed with mean zero and variance equal to $\|\Pi_S u\|^2$ where $u$ is the vector whose $i^{th}$ entry is $\binom{r + i - j - 1}{r - 1}$ for $i \geq j$ and 0 for $i < j$. Note that
\[
\|\Pi_S u\|^2 \leq \|u\|^2 = \sum_{i=j}^n \binom{r + i - j - 1}{r - 1}^2 
\leq n \binom{n - j + r - 1}{r - 1}^2 \leq n \times (n^{-1})^2 = n^{2r-1}.
\]
Using this (and the Gaussian tail bound: $\mathbb{P}\{|N(0, 1)| \geq t\} \leq \exp(-t^2/2)$), we obtain
\[
\mathbb{P}\{\lambda(Z) > \lambda\} \leq \sum_{r < j \leq n} \mathbb{P}\left\{\left|\sum_{i=j}^n \binom{r + i - j - 1}{r - 1} (\Pi_S Z)_i\right| > n^{-1}\lambda\right\}
\leq n \exp\left(-\frac{(n^{-1}\lambda)^2}{2n^{2r-1}}\right) = n \exp\left(-\frac{\lambda^2}{2n}\right).
\]
Combining the above inequality with (125), we obtain
\[ D(n^{r-1}λ∂f(θ^*)) ≤ r + \sqrt{n^3 + 2n^2} \exp \left( \frac{-λ^2}{4n} \right). \]
Now for \( λ ≥ \sqrt{6n \log(en)} \), we obtain
\[ D(n^{r-1}λ∂f(θ^*)) ≤ r + \sqrt{n^3 + 2n^2(\log(en))^{-3/2}} ≤ r + e^{-3/2} \sqrt{1 + \frac{2}{n}} ≤ C_r \]
where \( C_r \) only depends on \( r \). This bound and inequality (123) together complete the proof of Lemma 2.14.

Appendix C: Proofs of Key Technical Results

Our main proofs presented in Section B were crucially reliant on the following technical results: Lemma B.1 (used in the proof of Theorem 2.1), Lemma B.2 and Lemma B.3 (used in the proof of Theorem 2.2). The proofs of these results are given in this section. In addition, this section also contains the proofs of Proposition 2.5 and Lemma 2.7 from Section 2 of the main paper. The proofs of this section will further involve other technical results which (together with some other supporting results from the previous section such as Lemma B.4 which was used in the proof of Corollary 2.3) will be proved in Section D.

The organization of this section is as follows. We first prove Lemma B.1 in Subsection C.1. Next Lemma B.2 is proved in Subsection C.2 and this requires a precise understanding of the tangent cones \( T_K(\theta) \). Subsection C.3 is devoted to the proof of Lemma B.3. In Subsection C.4, we study the subdifferential of \( θ ↦ ∥D(θ)∥_1 \) and provide proofs of Proposition 2.5 and Lemma 2.7.

C.1. Proof of Lemma B.1

In this subsection, we shall provide the proof of Lemma B.1 (which was crucially used for the proof of Theorem 2.1). Our strategy is to use Dudley’s entropy bound to control the left hand side of (84) in terms of the metric entropy of \( S_r(V, t) \) (defined in (83)). Let us first formally define the notion of metric entropy. For a set \( K ⊂ \mathbb{R}^n \) and \( ε > 0 \), we define \( N(ε, K) \) to be the smallest integer \( m \) for which there exist points \( a_1, \ldots, a_m ∈ \mathbb{R}^n \) satisfying
\[ \sup_{a ∈ K} \inf_{1 ≤ i ≤ m} ∥a - a_i∥ ≤ ε \]
where, as usual, \( ∥·∥ \) denotes the Euclidean norm. The \( ε \)-metric entropy of \( K \) is the logarithm of \( N(ε, K) \).

Dudley’s entropy bound bounds the left hand side of (84) via \( \log N(ε, S_r(V, t)) \). The following theorem then provides upper bounds on \( \log N(ε, S_r(V, t)) \).
Theorem C.1. For $r \geq 1$, $t > 0$, $V > 0$ and $n \geq r$, let
\[ S_r(V, t) := \{ \theta \in \mathbb{R}^n : \| \theta \| \leq t, V(D^{(r-1)}\theta) \leq Vn^{1-r} \} . \]
Then for every $\epsilon > 0$, we have
\[
\log N(\epsilon, S_r(V, t)) \leq C_r \left( \frac{V\sqrt{n}}{\epsilon} \right)^{1/r} + r \log \left( 2 + \frac{2^r n^r t}{\epsilon \sqrt{n}} \right) + C_r \tag{126}
\]
for a constant $C_r$ that depends only on $r$.

Let us first complete the proof of Lemma B.1 assuming that Theorem C.1. The proof of Theorem C.1 will be provided following the proof of Lemma B.1.

Proof of Lemma B.1. Let $G$ denote the left hand side of (84). We use Dudley’s entropy bound to deduce that
\[
G \leq C\sigma \int_0^t \sqrt{\log N(\epsilon, S_r(V, t))} \, d\epsilon
\]
where the set $S_r(V, t)$ is defined as $\{ \theta \in \mathbb{R}^n : \| \theta \| \leq t, V(D^{(r-1)}\theta) \leq Vn^{1-r} \}$ and $N(\epsilon, S_r(V, t))$ denotes the $\epsilon$-covering number of $S_r(V, t)$ under the Euclidean metric. These covering numbers are bounded in Theorem C.1 which furnishes a constant $C_r$ such that
\[
\sqrt{\log N(\epsilon, S_r(V, t))} \leq C_r \left( \frac{V\sqrt{n}}{\epsilon} \right)^{1/(2r)} + \sqrt{r \log \left( 2 + \frac{2^r n^r t}{\epsilon \sqrt{n}} \right)} + C_r
\]
for every $\epsilon > 0$. Note that the square root of the right hand side of (126) is bounded from above by the right hand side above via the elementary inequality $\sqrt{a_1 + a_2 + a_3} \leq \sqrt{a_1} + \sqrt{a_2} + \sqrt{a_3}$ for $a_1, a_2, a_3 \geq 0$. It follows therefore that
\[
G \leq C_r \sigma t \left( \frac{V\sqrt{n}}{t} \right)^{1/(2r)} + C_r \sigma t + C_r \sigma \int_0^t \sqrt{\log \left( 2 + \frac{2^r n^r t}{\epsilon \sqrt{n}} \right)} \, d\epsilon.
\]
The last integral above can be controlled in the following way:
\[
\frac{1}{t} \int_0^t \sqrt{\log \left( 2 + \frac{2^r n^r t}{\epsilon \sqrt{n}} \right)} \, d\epsilon = \int_0^1 \sqrt{\log \left( 2 + \frac{2^r n^r}{u \sqrt{n}} \right)} \, du
\]
\[
= \int_0^{\sqrt{n}^{-r}} \sqrt{\log \left( 2 + \frac{2^r n^r}{u \sqrt{n}} \right)} \, du + \int_{\sqrt{n}^{-r}}^1 \sqrt{\log \left( 2 + \frac{2^r n^r}{u \sqrt{n}} \right)} \, du.
\]
For the second integral above, we use $u \geq \sqrt{n}^{-r}$ to argue that it is bounded from above by $\sqrt{\log(2 + 2^r n^{2r-1})} \leq C_r \sqrt{\log(en)}$. For the first integral, we use
\[
\log \left( 2 + \frac{2^r n^r}{u \sqrt{n}} \right) \leq 1 + \frac{2^r n^r}{u \sqrt{n}} \leq \frac{2^{r+1} n^r}{u \sqrt{n}}
\]
to obtain
\[ \int_0^{\sqrt{n} - r} \sqrt{\log \left( 2 + \frac{2^r n^r}{u \sqrt{n}} \right)} \, du \leq C_r. \]

We have therefore proved that
\[ G \leq C_r \sigma t \left( \frac{V \sqrt{n}}{t} \right)^{1/(2r)} + C_r \sigma t \sqrt{\log(en)} \]
for a constant $C_r$ which completes the proof of Lemma B.1.

Let us now provide the proof of Theorem C.1. For this, let us first introduce the following definition.

**Definition C.1.** For $r \geq 1$, $n \geq r$, real numbers $a_0, \ldots, a_{r-1}$ and non-negative real numbers $s_0, \ldots, s_{r-1}$, let $C_r(\{a_i\}, \{s_i\})$ denote the class of all $\theta \in \mathbb{R}^n$ for which $a_i \leq (D^{(i)} \theta)_1 \leq a_i + s_i$, $i = 0, 1, \ldots, r - 2$, and
\[ a_{r-1} \leq (D^{(r-1)} \theta)_1 \leq \cdots \leq (D^{(r-1)} \theta)_{n-r+1} \leq a_{r-1} + s_{r-1}. \]

**Remark C.1.** Note that when $r = 1$, the condition $a_i \leq (D^{(i)} \theta)_1 \leq a_i + s_i$, $i = 0, \ldots, r - 2$ is vacuous so that vectors in $C_1(\{a_i\}, \{s_i\})$ are required to only satisfy the inequality
\[ a_0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n \leq a_0 + s_0. \]

Our strategy for proving Theorem C.1 is to derive it from another result on the metric entropy of $C_r(\{a_i\}, \{s_i\})$. The following lemma gives an upper bound on the metric entropy of $C_r(\{a_i\}, \{s_i\})$. This is the most important ingredient for the proof of Theorem C.1. The proof of this lemma is given in Subsection D.6 and is based on an upper bound on the fat shattering dimension of the classes $C_r(\{a_i\}, \{s_i\})$ and a standard result (from Rudelson and Vershynin [41]) relating fat shattering dimension to metric entropy. See Subsection D.6 for full details including the definition of fat shattering dimension.

**Lemma C.2.** For every $\epsilon > 0$, $r \geq 1$, $n \geq r$, $a_0, \ldots, a_{r-1} \in \mathbb{R}$ and $s_0, \ldots, s_{r-1} \geq 0$, we have
\[ \log N(\epsilon, C_r(\{a_i\}, \{s_i\})) \leq C_r \left( \frac{\sqrt{n} \sum_{j=1}^{r} n^{j-1} s_{j-1}}{\epsilon} \right)^{1/r} \]
where $C_r$ is a positive constant that depends on $r$ alone.

We are now ready to prove Theorem C.1.

**Proof of Theorem C.1.** Fix $\delta > 0$ and let
\[ K_i := \max \left\{ u \geq 0 \text{ integer} : u \delta \leq 2^i t \right\} \quad \text{for } 0 \leq i < r. \]
It is then clear that $K_i \leq 2^t t / \delta < K_i + 1$ for every $0 \leq i < r$. Let $\mathcal{K}$ denote the class of all vectors $k := (k_0, \ldots, k_{r-1})$ where each $k_i$ is an integer satisfying $-(K_i + 1) \leq k_i \leq K_i$. For every $k = (k_0, \ldots, k_{r-1}) \in \mathcal{K}$, let

$$\mathcal{M}(k) := \{ \theta \in S_r(V, t) : k_i \delta \leq (D(i) \theta)_1 \leq (k_i + 1) \delta \text{ for } 0 \leq i < r \}.$$ 

As

$$|(D(i) \theta)_1| = \left| \sum_{j=1}^{i+1} (-1)^j \binom{i}{j-1} \theta_j \right| \leq \left( \binom{i}{0}^2 + \cdots + \binom{i}{i}^2 \right)^{1/2} \|	heta\| = \left( \frac{2i}{i} \right)^{1/2} \|	heta\| \leq 2^i t$$

for $\theta \in S_r(V, t)$ and $0 \leq i < r$, it follows that $S_r(V, t) \subseteq \cup_{k \in \mathcal{K}} \mathcal{M}(k)$. As a result

$$N(\epsilon, S_r(V, t)) \leq \sum_{k \in \mathcal{K}} N(\epsilon, \mathcal{M}(k)) \leq 2^r \prod_{i=0}^{r-1} (K_i + 1) \sup_{k \in \mathcal{K}} N(\epsilon, \mathcal{M}(k)).$$

Since $K_i \leq 2^t t / \delta \leq 2^{r-1} t / \delta$, we deduce

$$\log N(\epsilon, S_r(V, t)) \leq r \log \left( 2 + \frac{2^r t}{\delta} \right) + \sup_{k \in \mathcal{K}} \log N(\epsilon, \mathcal{M}(k)). \quad (127)$$

We now bound $\log N(\epsilon, \mathcal{M}(k))$ from above for a fixed $k \in \mathcal{K}$. For every $\theta \in \mathbb{R}^n$, let us define two vectors $\alpha(\theta) := (\alpha_1(\theta), \ldots, \alpha_n(\theta))$ and $\beta(\theta) := (\beta_1(\theta), \ldots, \beta_n(\theta))$ in $\mathbb{R}^n$ via

$$\alpha_i(\theta) := \sum_{j=1}^{i-r} \left( \frac{i-j-1}{r-1} \right) (D(r) \theta)^+_j + \sum_{j=1}^{r} \left( \frac{i-1}{j-1} \right) (D(j-1) \theta)^+_1 \quad (128)$$

and

$$\beta_i(\theta) := \sum_{j=1}^{i-r} \left( \frac{i-j-1}{r-1} \right) (D(r) \theta)^-_j + \sum_{j=1}^{r} \left( \frac{i-1}{j-1} \right) (D(j-1) \theta)^-_1$$

where $x^+ := \max(x, 0)$ and $x^- = x^+ - x$. It then follows from Lemma D.2 that $\theta = \alpha(\theta) - \beta(\theta)$ and, consequently,

$$\log N(\epsilon, \mathcal{M}(k)) \leq \log N(\epsilon/2, \mathcal{M}_\alpha(k)) + \log N(\epsilon/2, \mathcal{M}_\beta(k)) \quad (129)$$

where

$$\mathcal{M}_\alpha(k) := \{ \alpha(\theta) : \theta \in \mathcal{M}(k) \} \quad \text{and} \quad \mathcal{M}_\beta(k) := \{ \beta(\theta) : \theta \in \mathcal{M}(k) \}.$$ 

We now show how to control $\log N(\epsilon/2, \mathcal{M}_\alpha(k))$ below. The argument for $\log N(\epsilon/2, \mathcal{M}_\beta(k))$ will be similar. The main idea here (recall the definition of $C_r(\{a_i\}, \{s_i\})$ from Definition C.1) is to note that

$$\mathcal{M}_\alpha(k) \subseteq C_r(\{a_i\}, \{s_i\}) \quad (130)$$
with
\[ a_i = k_i^+ \delta \quad \text{for } i = 0, \ldots, r - 1, \]
and
\[ s_i = \delta \quad \text{for } i = 0, \ldots, r - 2 \text{ and } s_{r-1} = V n^{1-r} + \delta. \]
To see (130), first note that from the definition of \( \alpha(\theta) \) in (128), it is straightforward to check that
\[ (D^{(r)} \alpha(\theta))_j = (D^{(r)} \theta)_j^+ \quad \text{for } j = 1, \ldots, n - r \quad \text{(131)} \]
and
\[ (D^{(i)} \alpha(\theta))_1 = (D^{(i)} \theta)_1^+ \quad \text{for } 0 \leq i < r. \quad \text{(132)} \]
From these identities, it is easy to verify (130) in the following way. Let \( \theta \in \mathcal{M}(k) \) so that \( \alpha(\theta) \in \mathcal{M}_\alpha(k) \). Then \( k_i \delta \leq (D^{(i)} \theta)_1 \leq (k_i + 1) \delta \) for \( 0 \leq i < r \). This implies (because the map \( x \mapsto x^+ \) is non-decreasing and subadditive) via (132) that
\[ k_i^+ \delta \leq (D^{(i)} \alpha(\theta))_1 = (D^{(i)} \theta)_1^+ \leq k_i^+ \delta + \delta. \quad \text{(133)} \]
Also the identity (131) implies that \( D^{(r)} \alpha(\theta) \geq 0 \) which, together with (133), means that
\[
\frac{k_{r-1}^+ \delta}{V n^{1-r} + \delta - \frac{k_{r-1}^+ \delta + \delta}{V n^{1-r} + \delta}} = V(D^{(r-1)} \alpha(\theta))_1 \cdot \frac{D^{(r)} \alpha(\theta)}{V n^{1-r} + \delta} \leq V(D^{(r-1)} \alpha(\theta))_1 \leq V n^{1-r}.
\]
This follows since
\[
V(D^{(r-1)} \alpha(\theta)) = \|D^{(r)} \alpha(\theta)\|_1 = \|D^{(r)} \theta\|_1 = V(D^{(r-1)} \theta) \leq V n^{1-r}.
\]
This proves (130). We can thus use Lemma C.2 to bound \( N(\epsilon/2, \mathcal{M}_\alpha(k)) \) as
\[
\log N(\epsilon/2, \mathcal{M}_\alpha(k)) \leq C_r n^{1/(2r)} \left( \frac{\delta n^{r-1} + V}{\epsilon} \right)^{1/r}.
\]
Using the elementary inequality \( (a + b)^{1/r} \leq a^{1/r} + b^{1/r} \), we obtain the simpler inequality
\[
\log N(\epsilon/2, \mathcal{M}_\alpha(k)) \leq C_r \left( \frac{\delta n^{r-1} + V}{\epsilon} \right)^{1/r} + C_r \left( \frac{V \sqrt{n}}{\epsilon} \right)^{1/r}. \quad \text{(134)}
\]
Combining (127), (129) and (134), we obtain
\[
\log N(\epsilon, S_r(V, t)) \leq r \log \left( 2 + \frac{2^r t}{\delta} \right) + C_r \left( \frac{\delta n^{r-1} + V}{\epsilon} \right)^{1/r} + C_r \left( \frac{V \sqrt{n}}{\epsilon} \right)^{1/r}.
\]
Note that \( \delta > 0 \) above is arbitrary. Taking \( \delta = \epsilon \sqrt{n n^{-r}} \), we obtain (126) which completes the proof of Theorem C.1. \( \square \)
C.2. Study of the tangent cones $T_{K^{(r)}(V)}(\theta)$ and the proof of Lemma B.2

This section deals with the tangent cone (see (67) for the definition of tangent cone) of the convex set $K^{(r)}(V)$ (defined in (63)) at $\theta \in \mathbb{R}^n$ for which $V^{(r)}(\theta) = V$. This tangent cone is denoted by $T_{K^{(r)}(V)}(\theta)$. The ultimate goal of this subsection is to prove Lemma B.2 which was crucial for the proof of Theorem 2.2.

We start with the statement and proof of a lemma (Lemma C.3) which gives a precise characterization of $T_{K^{(r)}(V)}(\theta)$. Recall the notation $V_{a,b}(\alpha)$ (from (87)) for $1 \leq a \leq b \leq m$ and $\alpha \in \mathbb{R}^m$. Also recall, from Section 2, the notion of $r^{th}$ order knots (along with their signs) of vectors in $\mathbb{R}^n$.

**Lemma C.3.** Fix $r \geq 1$, $n \geq r + 1$ and let $K^{(r)}(V)$ be as in (63). Let $\theta \in K^{(r)}(V)$ be such that $V^{(r)}(\theta) = V$.

(i) Let $2 \leq j_1 < \cdots < j_k \leq n - r + 1$ denote all the $r^{th}$ order knots of $\theta$ along with associated signs $r_1, \ldots, r_k \in \{-1, 1\}$. Then

$$T_{K^{(r)}(V)}(\theta) = \left\{ \alpha \in \mathbb{R}^n : \sum_{i=0}^{k} V_{j_i, j_{i+1} - 1}(D^{(r-1)} \alpha) \leq \sum_{i=1}^{k} r_i \left((D^{(r-1)} \alpha)_{j_{i-1}} - (D^{(r-1)} \alpha)_{j_i}\right) \right\}$$

with the convention $j_0 = 1$ and $j_{k+1} = n - r + 2$.

(ii) Suppose $2 \leq j_1 < \cdots < j_k \leq n - r + 1$ denote any set of indices which contains all the $r^{th}$ order knots of $\theta$. Let $r_1, \ldots, r_k$ be such that $r_i$ is the sign of the knot corresponding to $j_i$ if $j_i$ is a knot and $r_i \in \{-1, 0, 1\}$ is arbitrary if $j_i$ is not a knot. Then

$$T_{K^{(r)}(V)}(\theta) \subseteq \left\{ \alpha \in \mathbb{R}^n : \sum_{i=0}^{k} V_{j_i, j_{i+1} - 1}(D^{(r-1)} \alpha) \leq \sum_{i=1}^{k} r_i \left((D^{(r-1)} \alpha)_{j_{i-1}} - (D^{(r-1)} \alpha)_{j_i}\right) \right\}$$

where again $j_0 = 1$ and $j_{k+1} = n - r + 2$.

**Remark C.2.** Lemma C.3 only deals with those $\theta \in K^{(r)}(V)$ for which $V^{(r)}(\theta) = V$. On the other hand, it is easy to see that when $V^{(r)}(\theta) < V$, the tangent cone $T_{K^{(r)}(V)}(\theta)$ equals $\mathbb{R}^n$.

**Remark C.3.** It must be clear from the right hand side of (136) that the tangent cone $T_{K^{(r)}(V)}(\theta)$ only depends on the knot indices $j_1, \ldots, j_k$ and the knot signs $r_1, \ldots, r_k$. For example, the exact values of $\theta$ at $j_1, \ldots, j_k$ are not relevant for the determination of the tangent cone.
Proof of Lemma C.3. We start with the proof of the first part of the lemma. Let $T$ denote the set on the right hand side of (135). Let us first prove that $T \subseteq T_{K^{(r)}(V)}(\theta)$. For this, we fix $\alpha \in T$ and argue that $\alpha \in T_{K^{(r)}(V)}(\theta)$, i.e., we show that there exists $c > 0$ such that $\theta + c\alpha \in K^{(r)}(V)$. For $c > 0$, first note that, by the definition of $V(\cdot)$, the variation $v := V(D^{(r-1)}(\theta + c\alpha))$ can be written as

$$v = \sum_{i=0}^{k} V_{j_i,j_{i+1}}(D^{(r-1)}(\theta + c\alpha))$$

$$+ \sum_{i=1}^{k} \left| (D^{(r-1)}(\theta + c\alpha))_{j_i} - (D^{(r-1)}(\theta + c\alpha))_{j_{i-1}} \right|$$

Because $\theta$ has no $r^{th}$ order knots except at $j_1, \ldots, j_k$, first term above can be simplified to obtain

$$v = c \sum_{i=0}^{k} V_{j_i,j_{i+1}}(D^{(r-1)}\alpha)$$

$$+ \sum_{i=1}^{k} \left| (D^{(r-1)}(\theta + c\alpha))_{j_i} - (D^{(r-1)}(\theta + c\alpha))_{j_{i-1}} \right|.$$ 

Now when $c > 0$ is sufficiently small, we can rewrite the above as

$$v = c \sum_{i=0}^{k} V_{j_i,j_{i+1}}(D^{(r-1)}\alpha)$$

$$+ \sum_{i=1}^{k} r_i \left\{ (D^{(r-1)}(\theta + c\alpha))_{j_i} - (D^{(r-1)}(\theta + c\alpha))_{j_{i-1}} \right\}$$

$$= V(D^{(r-1)}\theta)$$

$$+ c \left\{ \sum_{i=0}^{k} V_{j_i,j_{i+1}}(D^{(r-1)}\alpha) - \sum_{i=1}^{k} r_i \left( (D^{(r-1)}\alpha)_{j_i} - (D^{(r-1)}\alpha)_{j_{i-1}} \right) \right\}$$

$$\leq Vn^{1-r}$$

where the last step follows from the fact that $\alpha \in T$ and $V(D^{(r-1)}\theta) = Vn^{1-r}$. This proves $T \subseteq T_{K^{(r)}(V)}$.

We shall now verify that $T_{K^{(r)}(V)} \subseteq T$. As $T$ is a closed convex cone, it is enough to show that $\alpha - \theta \in T$ for every $\alpha \in K^{(r)}(V)$. For this, as $D^{(r-1)}(\alpha - \theta) = D^{(r-1)}\alpha - D^{(r-1)}\theta$, we need to show that

$$\sum_{i=0}^{k} V_{j_i,j_{i+1}}(D^{(r-1)}(\alpha - \theta)) + \sum_{i=1}^{k} r_i \left( (D^{(r-1)}\alpha)_{j_i} - (D^{(r-1)}\alpha)_{j_{i-1}} \right)$$

(137)
is not larger than
\[
\sum_{i=1}^{k} r_i \left( (D^{(r-1)} \theta)_{j_i} - (D^{(r-1)} \theta)_{j_i-1} \right).
\]
(138)

This is easy because (138) equals \(V(D^{(r-1)} \theta) = Vn^{1-r}\) and (137) is clearly bounded from above by \(V(D^{(r-1)} \alpha) \leq Vn^{1-r}\). This proves the first part of the lemma.

The second part is an easy consequence of the first part of the lemma and the following trivial observation. If \(j_i\) and \(j_{i+1}\) denote two consecutive knots of \(\theta\) and if \(j'_i\) is any integer with \(j_i < j'_i < j_{i+1}\), then
\[
V_{j_i,j_{i+1}-1}(\Delta) \geq V_{j_i,j'_i-1}(\Delta) + V_{j'_i,j_{i+1}-1}(\Delta) + r'_i \left( \Delta_{j'_i} - \Delta_{j_i} \right)
\]
for every \(\alpha \in \mathbb{R}^n\) and \(r'_i \in \{-1, 0, 1\}\) where \(\Delta := D^{(r-1)} \alpha\).

The following corollary to Lemma C.3 gives a simple necessary condition for a vector \(\alpha\) to belong to \(T_{K^{(r)}(V)}(\theta)\).

**Corollary C.4.** Fix \(r \geq 1\) and let \(K^{(r)}(V)\) be as in (63). Let \(\theta\) be any point in \(K^{(r)}(V)\) for which \(V^{(r)}(\theta) = V\). Let \(2 \leq j_1 < \cdots < j_k \leq n - r + 1\) and \(r_1, \ldots, r_k \in \{-1, 0, 1\}\) be as in Lemma C.3(ii). For every \(0 \leq i \leq k\), let \(\ell_i\) denote an arbitrary index lying in the set \(\{j_i, \ldots, j_{i+1} - 1\}\). Then for every \(\alpha \in T_{K^{(r)}(V)}(\theta)\) we have (with the convention that \(j_0 = 1, j_{k+1} = n - r + 2, r_0 = 0\) and \(r_{k+1} = 0\))
\[
\sum_{i=0}^{k} \Gamma_i(\alpha, \ell_i) \leq \sum_{i=0}^{k} (r_{i+1} - r_i) (D^{(r-1)} \alpha)_{\ell_i}
\]
(139)

where
\[
\Gamma_i(\alpha, \ell_i) := V_{j_i,j_{i+1}-1}(\Delta) - r_{i+1} \left( \Delta_{j_{i+1}-1} - \Delta_{\ell_i} \right) - r_i \left( \Delta_{\ell_i} - \Delta_{j_i} \right)
\]

with \(\Delta = (\Delta_1, \ldots, \Delta_{n-r+1}) := D^{(r-1)} \alpha\).

**Proof of Corollary C.4.** Fix \(\alpha \in T_{K^{(r)}(V)}(\theta)\). Lemma C.3 gives that
\[
\sum_{i=0}^{k} V_{j_i,j_{i+1}-1}(\Delta) \leq \sum_{i=1}^{k} r_i \left( \Delta_{j_{i-1}} - \Delta_{j_i} \right).
\]
(140)

Writing
\[
V_{j_i,j_{i+1}-1}(\Delta) = \Gamma_i(\alpha, \ell_i) + r_{i+1} \left( \Delta_{j_{i+1}-1} - \Delta_{\ell_i} \right) + r_i \left( \Delta_{\ell_i} - \Delta_{j_i} \right)
\]
in (140), we deduce that \(\sum_{i=0}^{k} \Gamma_i(\alpha, \ell_i)\) is bounded from above by
\[
\sum_{i=1}^{k} r_i \left( \Delta_{j_{i-1}} - \Delta_{j_i} \right) - \sum_{i=0}^{k} r_{i+1} \left( \Delta_{j_{i+1}-1} - \Delta_{\ell_i} \right) - \sum_{i=0}^{k} r_i \left( \Delta_{\ell_i} - \Delta_{j_i} \right).
\]

It is now trivial to check that the expression above equals the right hand side of (139) which completes the proof of Corollary C.4. \(\square\)
We next show that under the assumption that $\|\alpha\| \leq 1$, the right hand side of (139) can be made small by choosing $\ell_0, \ldots, \ell_k$ appropriately. This is the content of the next lemma. Let $2 \leq j_1 < \cdots < j_k \leq n - r + 1$ and $r_1, \ldots, r_k \in \{-1, 0, 1\}$ be as in Lemma C.3(ii). Also let $j_0 = 1$, $j_{k+1} = n - r + 2$ and $r_0 = r_{k+1} = 0$. The indices $j_0, j_1, \ldots, j_k, j_{k+1}$ can be used to define a partition of $\{1, \ldots, n\}$ in the following way: $I_0 := \{j_0, \ldots, j_1 + r - 2\}$ and $I_i = \{j_i + r - 1, \ldots, j_{i+1} + r - 2\}$ for $i = 1, \ldots, k$.

Observe that the length of $I_i$ equals $n_i$ where $n_0 := j_1 + r - 2$ and $n_i := j_{i+1} - j_i$ for $1 \leq i \leq k$.

**Lemma C.5.** Let $\theta \in \mathbb{R}^n$ and let $2 \leq j_1 < \cdots < j_k \leq n - r + 1$ and $r_1, \ldots, r_k \in \{-1, 0, 1\}$ be as in Lemma C.3(ii). Also let $j_0 = 1$, $j_{k+1} = n - r + 2$ and $r_0 = r_{k+1} = 0$. Further let $I_0, \ldots, I_k$ and $n_0, \ldots, n_k$ be as described above. For every $\alpha \in \mathbb{R}^n$ with $\|\alpha\| \leq 1$, there exist indices $\ell_0 \in I_0, \ldots, \ell_k \in I_k$ such that

$$
\sum_{i=0}^{k} (r_{i+1} - r_i)(D^{(r-1)}\alpha)_{\ell_i} \leq C_r \sqrt{\sum_{i=0}^{k} n_i^{1-2r} I\{r_i \neq r_{i+1}\}}
$$

(141)

where $C_r$ is a positive constant that depends only on $r$.

Note that the role of $\theta \in \mathbb{R}^n$ in the above lemma is just to define the $j_i$’s and the $r_i$’s as in Lemma C.3(ii).

The proof of Lemma C.5 is given next. A crucial role in this proof is played by the following result on the magnitude of $\min_{1 \leq i \leq n - r + 1}(D^{(r-1)}\theta)_i$ for a vector $\theta$ with $\|\theta\| \leq 1$. This result (proved in Subsection D.4) might be of independent interest.

**Lemma C.6.** Fix $r \geq 1$. There exists a positive constant $C_r$ depending only on $r$ such that for every $n \geq 2r$, $t > 0$ and $\theta \in \mathbb{R}^n$ with $\|\theta\| \leq t$, there exist indices $\ell_1, \ell_2 \in \{1, \ldots, n - r + 1\}$ such that

$$(D^{(r-1)}\theta)_{\ell_1} \leq C_r n^{(1/2) - r} t \quad \text{and} \quad (D^{(r-1)}\theta)_{\ell_2} \geq -C_r n^{(1/2) - r} t.$$  

(142)

**Remark C.4.** Lemma C.6 is trivial for $r = 1$ (when it holds with $C_1 = 1$) but the extension to $r \geq 2$ is non-trivial. Also, for general $r \geq 2$, the two indices $\ell_1$ and $\ell_2$ will be different and it will be incorrect to claim that for every $\theta \in \mathbb{R}^n$ with $\|\theta\| \leq 1$, there exists a single index $\ell \in \{1, \ldots, n - r + 1\}$ for which $|D^{(r-1)}\theta|_\ell \leq C_r n^{(1/2) - r} t$. One may define $\ell_1$ and $\ell_2$ as

$$
\ell_1 := \argmin_{1 \leq j \leq n - r + 1} (D^{(r-1)}\theta)_j \quad \text{and} \quad \ell_2 := \argmax_{1 \leq j \leq n - r + 1} (D^{(r-1)}\theta)_j.
$$

We are now ready to prove Lemma C.5.

**Proof of Lemma C.5.** The proof of Lemma C.5 is crucially reliant on Lemma C.6 (proved in Section D.4) which essentially says that

$$
\sup_{\alpha \in \mathbb{R}^n : \|\alpha\| \leq 1} \min_{1 \leq i \leq n - r + 1} (D^{(r-1)}\alpha)_i \leq C_r n^{(1/2) - r} t.
$$
for every \( t > 0 \) and \( n \geq r \).

Fix \( \alpha \in \mathbb{R}^n \). Define
\[
\alpha^{(0)} := (\alpha_{j_0}, \ldots, \alpha_{j_1+r-2})
\]
and
\[
\alpha^{(u)} := (\alpha_{j_u+r-1}, \ldots, \alpha_{j_{u+1}+r-2})
\]
for \( u = 1, \ldots, k \). Note that the vector \( \alpha^{(u)} \) has length exactly equal to \( n_u \), for \( u = 0, \ldots, k \).

Fix \( 0 \leq u \leq k \) and let \( t_u := \|\alpha^{(u)}\| \). By Lemma C.6, there exists an index \( \ell_u' \in \{1, \ldots, n_u - r + 1\} \) such that
\[
(r_{u+1} - r_u)(D^{(r-1)}\alpha^{(u)})_{\ell_u'} \leq 2C_r n_u^{1/2-r} t_u I\{r_u \neq r_{u+1}\}
\]
for a constant \( C_r \) depending on \( r \) alone. Taking
\[
\ell_0 := \ell_0' \quad \text{and} \quad \ell_u := j_u + r - 2 + \ell_u' \quad \text{for } 1 \leq u \leq k,
\]
and using the fact that \( (D^{(r-1)}\alpha^{(u)})_{\ell_u'} = (D^{(r-1)}\alpha)_{\ell_u} \), we deduce from (143) that
\[
(r_{u+1} - r_u)(D^{(r-1)}\alpha)_{\ell_u} \leq 2C_r n_u^{1/2-r} t_u I\{r_u \neq r_{u+1}\}
\]
for every \( u = 0, 1, \ldots, k \). The left hand side of (141) can therefore be bounded as
\[
\sum_{i=0}^{k} (r_{i+1} - r_i)(D^{(r-1)}\alpha)_{\ell_i} \leq 2C_r \sum_{i=0}^{k} n_i^{1/2-r} t_i I\{r_i \neq r_{i+1}\}
\]
\[
\leq 2C_r \sqrt{\sum_{i=0}^{k} n_i^{1-2r} I\{r_i \neq r_{i+1}\}} \sqrt{\sum_{i=0}^{k} t_i^2}
\]
\[
\leq 2C_r \sqrt{\sum_{i=0}^{k} n_i^{1-2r} I\{r_i \neq r_{i+1}\}}
\]
where we have used Cauchy-Schwarz inequality and the fact that \( \sum_{i=0}^{k} t_i^2 = \|\alpha\|^2 \leq 1 \).

This completes the proof of Lemma C.5.

We now have all the ingredients to complete the proof of Lemma B.2.

Proof of Lemma B.2. The result clearly follows by combining Corollary C.4 and Lemma C.5.

\[\square\]
C.3. Proof of Lemma B.3

The goal of this subsection is to prove Lemma B.3 which was crucial for the proof of Theorem 2.2. We shall actually prove the following more precise result from which Lemma B.3 easily follows.

Lemma C.7. Fix \( r \geq 1, n \geq r, 1 \leq \ell \leq n - r + 1, t > 0 \) and \( \delta \geq 0 \). For \( \theta \in \mathbb{R}^n \), let \( \Delta(\theta) = (\Delta_1(\theta), \ldots, \Delta_{n-r+1}(\theta)) := D^{(r-1)}\theta \). For every \( r_1, r_2 \in \{-1, 0, 1\} \), the quantity

\[
G := \mathbb{E} \left[ \sup \left\{ \langle \xi, \theta \rangle : \theta \in \mathbb{R}^n, \|\theta\| \leq t, \quad V(\Delta(\theta)) \leq r_1(\Delta_{\ell}(\theta) - \Delta_1(\theta)) \\
+ r_2(\Delta_{n-r+1}(\theta) - \Delta_\ell(\theta)) + \delta \right\} \right]
\]

is bounded from above in the following way. When \( r_1 = r_2 = 0 \), we have

\[
G \leq C_r \sigma (2r-1)/(2r) \delta^{1/(2r)} n^{(2r-1)/(4r)} + C_r \sigma t \sqrt{\log(en)}.
\]

When \( r_1 = 0, r_2 \neq 0 \), we have

\[
G \leq C_r \sigma \left\{ t^{(2r-1)/(2r)} \ell_1^{(2r-1)/4r} \delta^{1/(2r)} + t \sqrt{\log(\ell_1)} \right\}
\]

\[
+ C_r \sigma \left( t + \delta \ell_2^{(2r-1)/2} \right) \sqrt{\log(\ell_2)}.
\]

When \( r_1 \neq 0, r_2 = 0 \), we have

\[
G \leq C_r \sigma \left( t + \delta \ell_1^{(2r-1)/2} \right) \sqrt{\log(\ell_1)}
\]

\[
+ C_r \sigma \left\{ t^{(2r-1)/(2r)} \ell_2^{(2r-1)/4r} \delta^{1/(2r)} + t \sqrt{\log(\ell_2)} \right\}.
\]

Finally when \( r_1 \neq 0, r_2 \neq 0 \), we have

\[
G \leq C_r \sigma \left( t + \delta \ell_1^{(2r-1)/2} \right) \sqrt{\log(\ell_1)} + C_r \sigma \left( t + \delta \ell_2^{(2r-1)/2} \right) \sqrt{\log(\ell_2)}.
\]

In each case, \( \ell_1 := \ell + r - 1, \ell_2 := n - \ell - r + 1 \) and \( C_r \) is a constant depending on \( r \) alone.

Remark C.5. It is easy to see that Lemma C.7 implies Lemma B.3. This is a consequence of the fact that the integers \( \ell_1 \) and \( \ell_2 \) appearing in Lemma C.7 are both bounded from above by \( n \).

The rest of this subsection is dedicated to the proof of Lemma C.7. As described in Remark above, Lemma C.7 implies Lemma B.3. Before proceeding to prove Lemma C.7, we prove an auxiliary result below which will considerably simplify the proof of Lemma C.7.
Lemma C.8. For every $r \geq 1$, $n \geq r$, $t > 0$ and $\delta \geq 0$, we have that

$$
\mathbb{E}
\left[
\sup_{\theta \in \mathbb{R}^n: \|\theta\| \leq t}
\langle \xi, \theta \rangle
\right]
\leq
\left\|
\frac{\langle \beta - \alpha, \theta \rangle}{\text{max}(\text{Gaussian width of the cone})}
\right\|
\leq
\frac{C_r \sigma (t + \delta n^{r-1/2}) \sqrt{\log(en)}}{}
$$

is bounded from above by

$$
C_r \sigma (t + \delta n^{r-1/2}) \sqrt{\log(en)}
$$

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

Lemma C.8 is proved below. This proof will use Lemma D.2 (stated and proved in Subsection D.2) which provides a formula for an arbitrary vector $\theta$ in terms of $D^{(r)} \theta$ and Bellec [3, Theorem 1 in the supplementary material] which provides a bound for the statistical dimension of the cone of all $\gamma \in \mathbb{R}^n$ which satisfy $\min_{1 \leq i \leq n-r} (D^{(r)} \gamma)_i \geq 0$.

Proof of Lemma C.8. We can assume without loss of generality that $t = 1$ (which is ensured by scaling and replacing $\delta$ by $\delta/t$). The idea of this proof is to write $\theta$ as the difference of two vectors $\alpha(\theta)$ and $\beta(\theta)$ which satisfy $\min_{1 \leq i \leq n-r} (D^{(r)} \alpha(\theta))_i \geq 0$ and $\min_{1 \leq i \leq n-r} (D^{(r)} \beta(\theta))_i \geq 0$. Bellec [3, Theorem 1 in the supplementary material] will then be used to control the Gaussian width of the cone of all $\gamma \in \mathbb{R}^n$ which satisfy $\min_{1 \leq i \leq n-r} (D^{(r)} \gamma)_i \geq 0$.

To construct the sequences $\alpha(\theta)$ and $\beta(\theta)$, we use Lemma D.2 which gives the following formula for expressing a vector $\theta \in \mathbb{R}^n$ in terms of $D^{(r)} \theta$ and $(D^{(i)} \theta)_1$ for $i = 0, \ldots, r - 1$:

$$
\theta_i = \sum_{j=1}^{i-r} \binom{i-j-1}{r-1} (D^{(r)} \theta)_j + \sum_{j=1}^{r} \binom{i-1}{j-1} (D^{(j-1)} \theta)_1
$$

where we take the convention that $\binom{a}{b} = 0$ for $b > a$, $\binom{a}{0} = 1$ so that the first term in the right hand side is zero unless $i > r$. Motivated by the above expression, we define $\alpha(\theta) := (\alpha_1(\theta), \ldots, \alpha_n(\theta))$ and $\beta(\theta) := (\beta_1(\theta), \ldots, \beta_n(\theta))$ in the following way:

$$
\alpha_i(\theta) := \sum_{j=1}^{i-r} \binom{i-j-1}{r-1} (D^{(r)} \theta)_j^+ + \sum_{j=1}^{r} \binom{i-1}{j-1} (D^{(j-1)} \theta)_1^-
$$

and

$$
\beta_i(\theta) := \sum_{j=1}^{i-r} \binom{i-j-1}{r-1} (D^{(r)} \theta)_j^-
$$

where $x^+ := \max(x, 0)$ and $x^- := x^+ - x$. It is easy then to observe the following: (a) $\theta = \alpha(\theta) - \beta(\theta)$, (b) $(D^{(r)} \alpha(\theta))_i = (D^{(r)} \theta)_i^+$, $(D^{(r)} \beta(\theta))_i = (D^{(r)} \theta)_i^-$, (c) both vectors $\alpha(\theta)$ and $\beta(\theta)$ belong to $S^{[r]}_n$ where

$$
S^{[r]}_n := \left\{ \gamma \in \mathbb{R}^n : \min_{1 \leq i \leq n-r} (D^{(r)} \gamma)_i \geq 0 \right\}
$$

$$
= \left\{ \gamma \in \mathbb{R}^n : (D^{(r-1)} \gamma)_1 \leq \cdots \leq (D^{(r-1)} \gamma)_{n-r+1} \right\}.
$$
and (d) \((D^{(j-1)}\beta(\theta))_1 = 0\) for \(1 \leq j \leq r\). From these, it follows that

\[
V(D^{(r-1)}\theta) = \|D^{(r)}\theta\|_1 = \sum_{i=1}^{n-r} |(D^{(r)}\theta)_i| \\
= \sum_{i=1}^{n-r} (D^{(r)}\alpha(\theta))_i + \sum_{i=1}^{n-r} (D^{(r)}\beta(\theta))_i \\
= (D^{(r-1)}\alpha(\theta))_{n-r+1} - (D^{(r-1)}\alpha(\theta))_1 \\
+ (D^{(r-1)}\beta(\theta))_{n-r+1} - (D^{(r-1)}\beta(\theta))_1 \\
= (D^{(r-1)}\alpha(\theta))_{n-r+1} - (D^{(r-1)}\alpha(\theta))_1 \\
+ (D^{(r-1)}\beta(\theta))_{n-r+1}.
\]

From the above (and the fact that \(D^{(r-1)}\theta = D^{(r-1)}\alpha(\theta) - D^{(r-1)}\beta(\theta)\)), it is straightforward to observe that the condition

\[
V(D^{(r-1)}\theta) \leq (D^{(r-1)}\theta)_{n-r+1} - (D^{(r-1)}\theta)_1 + \delta
\]

is equivalent to

\[
(D^{(r-1)}\beta(\theta))_{n-r+1} \leq \frac{\delta}{2}.
\]

(144)

Now for \(\beta(\theta) \in S^{[r]}_n\), \((D^{(j-1)}\beta(\theta))_1 = 0\) for \(1 \leq j \leq r\), and satisfying (144), we can use Lemma D.2 (with \(r\) replaced by \(r - 1\)) to observe that

\[
0 \leq \beta_i(\theta) \leq \frac{\delta}{2} \sum_{j=1}^{i-r+1} \left(\frac{i-j-1}{r-2}\right) = \frac{\delta}{2} \frac{(i-1)}{r-1} \leq \frac{\delta}{2} i^{r-1}
\]

(145)

where we have used the following elementary identity involving binomial coefficients: for every two integers \(a\) and \(b\) with \(0 \leq b < a\), we have

\[
\binom{b}{b} + \binom{b+1}{b} + \cdots + \binom{a}{b} = \binom{a+1}{b+1}.
\]

(146)

Note the presence of the term \(r - 2\) in some of the binomial coefficients in (145) which will be negative when \(r = 1\). But the inequality \(0 \leq \beta_i(\theta) \leq \delta/2\) is also true for \(r = 1\) which can directly be seen from \(\beta_1(\theta) \leq \delta/2\) (inequality (144) for \(r = 1\)), the fact that \(S^{[1]}_n\) consists of monotone sequences (so that \(\beta_i(\theta) \leq \beta_n(\theta)\)) and the fact that \((D^{(j-1)}\beta(\theta))_1 = 0\) for \(1 \leq j \leq r\) (which for \(r = 1\) gives \(\beta_1(\theta) = 0\)).

A consequence of (145) is that

\[
\|\beta(\theta)\|^2 \leq \frac{\delta^2}{4} \sum_{i=1}^{n} i^{2r-2} \leq \frac{\delta^2}{4} n^{2r-1}
\]
or \(\|\beta(\theta)\| \leq \delta n^{r-1/2}/2\). Because \(\|\theta\| \leq 1\), we further deduce that

\[
\|\alpha(\theta)\| \leq \|\theta\| + \|\beta(\theta)\| \leq 1 + \frac{\delta}{2} n^{r-1/2}.
\]

Based on these observations, if

\[
G := \mathbb{E} \left[ \sup_{\theta \in \mathbb{R}^n: \|\theta\| \leq t} \langle \xi, \theta \rangle \right],
\]

we can write

\[
G \leq \mathbb{E} \left[ \sup_{\alpha \in S_n^{[r]}: \|\alpha\| \leq 1 + \delta n^{r-1/2}/2} \langle \xi, \alpha \rangle \right] + \mathbb{E} \left[ \sup_{\beta \in S_n^{[r]}: \|\beta\| \leq \delta n^{r-1/2}/2} \langle \xi, -\beta \rangle \right].
\]

By an elementary scaling property and the fact that \(\xi\) and \(-\xi\) have the same distribution, we deduce that

\[
G \leq (1 + \delta n^{r-1/2}) \ w(S_n^{[r]})
\]

where \(w(S_n^{[r]})\) is the Gaussian width of \(S_n^{[r]}\) (defined in (68)). The right hand side above can be bounded using Bellec [3, Theorem 1 in the supplementary material] which implies that

\[
w(S_n^{[r]}) \leq C_r \sigma \sqrt{\log(en)}
\]

for a constant \(C_r\). To be precise, Bellec [3, Equation (5) in the supplementary material] gives a bound for \(\delta(S_n^{[r]})\). The connection (69) between Gaussian width and statistical dimension then leads to the above stated bound. We therefore have

\[
G \leq C_r \left(1 + \delta n^{r-1/2}\right) \sqrt{\log(en)}.
\]

which completes the proof of Lemma C.8. \(\square\)

We are now ready to prove Lemma C.7.

**Proof of Lemma C.7.** The case when \(r_1 = r_2 = 0\) follows directly from Lemma B.1 so we assume that at least one of \(r_1\) and \(r_2\) is non-zero.

For \(\theta \in \mathbb{R}^n\), let \(\theta^{(1)} := (\theta_1, \ldots, \theta_{\ell+r-1})\) and \(\theta^{(2)} := (\theta_{\ell+r}, \ldots, \theta_n)\). We analogously define \(\xi^{(1)}\) and \(\xi^{(2)}\). Recall that \(\Delta \equiv \Delta(\theta) = (\Delta_1(\theta), \ldots, \Delta_{n-r+1}(\theta)) := D^{(r-1)}(\theta)\). We first claim that under the assumption \(V(\Delta) \leq r_1(\Delta_\ell - \Delta_1) + r_2(\Delta_{n-r+1} - \Delta_\ell) + \delta\), we have

\[
V(D^{(r-1)}(\theta^{(1)})) = V(\Delta_1, \ldots, \Delta_\ell) \leq r_1(\Delta_\ell - \Delta_1) + \delta \quad (147)
\]

and

\[
V(D^{(r-1)}(\theta^{(2)})) = V(\Delta_{\ell+r}, \ldots, \Delta_{n-r+1}) \leq r_2(\Delta_{n-r+1} - \Delta_{\ell+r}) + \delta. \quad (148)
\]
Inequality (147) is a consequence of
\[ r_1(\Delta_\ell - \Delta_1) + r_2(\Delta_{n-r+1} - \Delta_\ell) + \delta \geq V(\Delta) \geq V(\Delta_1, \ldots, \Delta_\ell) + r_2(\Delta_{n-r+1} - \Delta_\ell) \]
while (148) is a consequence of
\[ r_1(\Delta_\ell - \Delta_1) + r_2(\Delta_{n-r+1} - \Delta_\ell) + \delta \geq V(\Delta) \geq r_1(\Delta_\ell - \Delta_1) + V(\Delta_{\ell+r}, \ldots, \Delta_{n-r+1}) + r_2(\Delta_{\ell+r} - \Delta_\ell). \]

From inequalities (147) and (148), and the fact that \( \langle \xi, \theta \rangle = \sum_{i=1}^{2} \langle \xi^{(i)}, \theta^{(i)} \rangle \), it follows that \( G \leq G_1 + G_2 \) where
\[
G_1 := \mathbb{E} \left[ \sup \left\{ \langle \xi^{(1)}, \theta^{(1)} \rangle : \|\theta^{(1)}\| \leq t, \right. \\
V(\mathcal{D}^{(r-1)}\theta^{(1)}) \leq r_1((\mathcal{D}^{(r-1)}\theta^{(1)})_\ell - (\mathcal{D}^{(r-1)}\theta^{(1)})_1) + \delta \} \right]
\]
and
\[
G_2 := \mathbb{E} \left[ \sup \left\{ \langle \xi^{(2)}, \theta^{(2)} \rangle : \|\theta^{(2)}\| \leq t, \right. \\
V(\mathcal{D}^{(r-1)}\theta^{(2)}) \leq r_2((\mathcal{D}^{(r-1)}\theta^{(2)})_{n-\ell-2r+2} - (\mathcal{D}^{(r-1)}\theta^{(2)})_1) + \delta \} \right].
\]

Note now that when \( r_1 = 0 \), we have
\[
G_1 \leq C_r \sigma \left\{ t^{(2r-1)/(2r)}(\ell + r - 1)^{(2r-1)/(4r)}\delta^{1/(2r)} + t\sqrt{\log(e(\ell + r - 1))} \right\}
\]
as this bound simply follows from Lemma B.1. On the other hand, when \( r_1 \neq 0 \), we have
\[
G_2 \leq C_r \sigma \left( t + \delta(\ell + r - 1)^{(2r-1)/2} \right) \sqrt{\log(e(\ell + r - 1))}.
\]
This follows from Lemma C.8 when \( r_1 = 1 \). When \( r_1 = -1 \), we can switch from \( \theta^{(1)} \) to \(-\theta^{(1)}\) so that the above bound will again follow from Lemma C.8. An identical argument also gives that
\[
G_2 \leq C_r \sigma \left\{ t^{(2r-1)/(2r)}(n - \ell - r + 1)^{(2r-1)/(4r)}\delta^{1/(2r)} + t\sqrt{\log(e(n - \ell - r + 1))} \right\}
\]
when \( r_2 = 0 \) and
\[
G_2 \leq C_r \sigma \left( t + \delta(n - \ell - r + 1)^{(2r-1)/2} \right) \sqrt{\log(e(n - \ell - r + 1))}
\]
when \( r_2 \neq 0 \). By putting together the above bounds for \( G_1 \) and \( G_2 \) the proof of Lemma C.7 is complete. \( \square \)
C.4. Subdifferential of $\theta \mapsto \|D^{(r)}\theta\|_1$ and proof of Lemma 2.7

This subsection provides a study of the subdifferential $\partial f(\theta)$ where $f(\theta) := \|D^{(r)}\theta\|_1$ with an aim to prove Proposition 2.5 and Lemma 2.7 in Section 2. We start by proving Proposition 2.5 which gives a precise characterization of the subdifferential.

Proof of Proposition 2.5. Let us first construct an $n \times n$ matrix $M$ such that for every $\beta \in \mathbb{R}^n$, we have

$$ (M\beta)_i = \begin{cases} (D^{(i-1)}\beta)_1 & \text{for } i = 1, \ldots, r \\ (D^{(r)}\beta)_{i-r} & \text{for } i = r + 1, \ldots, n. \end{cases} $$

This is of course possible because $\beta \mapsto (D^{(i)}\beta)_j$ is a linear mapping. More specifically, it can be checked that $M = (M_{ij})$ defined by

$$ M_{ij} = \begin{cases} (-1)^{i-j} (\binom{i-1}{j}) I\{1 \leq j \leq i \leq n\} & \text{for } 1 \leq i \leq r, 1 \leq j \leq n \\ (-1)^{i-j} (\binom{r}{i-j}) I\{i-r \leq j \leq i\} & \text{for } r + 1 \leq i \leq n, 1 \leq j \leq n \end{cases} $$

satisfies the requirement. This is a consequence of the expression:

$$ (D^{(r)}\beta)_j = \sum_{k=j}^{j+r} (-1)^{j+r-k} \binom{r}{k-j} \beta_k \quad \text{for } 1 \leq j \leq n - r. $$

It is easy to see from the formula for $M$ that it is lower triangular with positive diagonal entries and hence invertible.

Now a vector $v \in \mathbb{R}^n$ is in $\partial f(\theta)$ if and only if it satisfies

$$ f(\theta + \beta) - f(\theta) \geq \langle v, \beta \rangle \quad \text{for every } \beta \in \mathbb{R}^n. \quad (149) $$

The left hand side above can be written as

$$ f(\theta + \beta) - f(\theta) = \sum_{j=1}^{n-r} [|(D^{(r)}\theta)_j + (M\beta)_{j+r}| - |(D^{(r)}\theta)_j|]. \quad (150) $$

The right hand side in (149) can be written using Lemma D.2 as

$$ \langle v, \beta \rangle = \sum_{i=1}^{n} v_i \beta_i $$

$$ = \sum_{i=1}^{n} v_i \sum_{j=1}^{i-r} \binom{i-j-1}{r-1} (D^{(r)}\beta)_j + \sum_{i=1}^{n} v_i \sum_{j=1}^{r} \binom{i-1}{j-1} (D^{(j-1)}\beta)_1 $$

$$ = \sum_{j=1}^{n} (D^{(r)}\beta)_j \sum_{i=j+r}^{n} \binom{i-j-1}{r-1} v_i + \sum_{j=1}^{r} (D^{(j-1)}\beta)_1 \sum_{i=j}^{n} \binom{i-1}{j-1} v_i $$

$$ = \sum_{j=1}^{n-r} a_{r+j}(M\beta)_{r+j} + \sum_{j=1}^{r} b_j(M\beta)_j $$
where

\[ b_j := \sum_{i=1}^{n} \left( \frac{i-1}{j-1} \right) v_i \quad \text{for} \ 1 \leq j \leq r \]

and

\[ a_{r+j} := \sum_{i=r+j}^{n} \left( \frac{i-j-1}{r-1} \right) v_i \quad \text{for} \ 1 \leq j \leq n - r. \]

We now set \( \beta = \pm M^{-1} e_j \) for \( 1 \leq j \leq r \), where \( e_j \) is the \( j \)’th standard basis vector of \( \mathbb{R}^n \). Then, using (150), \( f(\theta + \beta) - f(\theta) = 0 \), so we must have \( \langle v, \beta \rangle = b_j = 0 \). Now set \( \beta = \lambda M^{-1} e_{r+j} \) for \( 1 \leq j \leq n - r \). If \( (D(\theta \beta))_j > 0 \), then \( f(\theta + \beta) - f(\theta) = \lambda \) for \( \lambda \geq -(D(\theta \beta))_j \), and \( \langle v, \beta \rangle = \lambda a_{r+j} \). In particular, \( a_{r+j} \leq 1 \) by taking \( \lambda > 0 \), and \( a_{r+j} \geq 1 \) by taking \( 0 > \lambda \geq -(D(\theta \beta))_j \), so we must have \( a_{r+j} = 1 \). Similarly, if \( (D(\theta \beta))_j < 0 \), then we must have \( a_{r+j} = -1 \). If \( (D(\theta \beta))_j = 0 \), then \( f(\theta + \beta) - f(\theta) = |\lambda| \), so we must have \( a_{r+j} \in [-1, 1] \). We have thus proved that if \( v \in \partial f(\theta) \), then \( b_j = 0 \) for \( 1 \leq j \leq r \) and

\[ a_{r+j} = \begin{cases} \operatorname{sgn}((D(\theta \beta))_j) & \text{if} \ (D(\theta \beta))_j \neq 0 \\ \in [-1, 1] & \text{otherwise} \end{cases} \]

for \( 1 \leq j \leq n - r \). On the other hand, it is easy to see that if these two conditions are satisfied, then \( v \in \partial f(\theta) \). The proof of Lemma 2.5 will then be complete by the observation that \( b_j = 0 \) for \( 1 \leq j \leq r \) is equivalent to \( a_j = 0 \) for \( 1 \leq j \leq r \), where \( a_j \) is the left hand side of (23). To see this, just note that

\[
\sum_{k=j}^{r} \left( \frac{r-j}{r-k} \right) b_k = \sum_{k=j}^{r} \left( \frac{r-j}{r-k} \right) \sum_{i=k}^{n} \left( \frac{i-1}{k-1} \right) v_i = \sum_{i=j}^{n} v_i \sum_{k=j}^{r} \left( \frac{r-j}{r-k} \right) \left( \frac{i-1}{k-1} \right)
\]

\[ = \sum_{i=j}^{n} v_i \sum_{k=1}^{r-j} \left( \frac{i-1}{k-1} \right) = \sum_{i=j}^{n} v_i \left( \frac{r+i-j-1}{r-1} \right) = a_j. \]

so that \( a_j \) is related to \( b_j \) by a triangular linear system. This completes the proof of Proposition 2.5. \( \square \)

We are now ready to prove Lemma 2.7.

**Proof of Lemma 2.7.** We start with proof of the assertions for \( r = 1 \) (including inequality (29)) and then proceed to the proofs of inequalities (30) and (31).

**Proofs for \( r = 1 \).** Assume that \( r = 1 \) and that \( D \theta^* \neq 0 \). Let \( 2 \leq j_1 < \cdots < j_k \leq n \) denote the jumps (first order knots) of \( \theta \) with signs are \( v_1, \ldots, v_k \). Also let \( j_0 = 1 \), \( j_{k+1} = n + 1 \) and \( v_0 = v_{k+1} = 0 \). Then \( n_i := j_{i+1} - j_i \) for \( 0 \leq i \leq k \) denote the lengths of the \( k + 1 \) constant pieces of \( \theta^* \).

Define the vector \( v_0 = (v_{01}, \ldots, v_{0m}) \in \mathbb{R}^m \) in the following way. For \( 1 \leq i \leq n \), let \( 0 \leq l \leq k \) be the unique integer such that \( j_l \leq i < j_{l+1} \). Then we take \( v_{0i} := (v_l - v_{l+1})/n_l. \)
We first claim that $v_0 \in \partial f(\theta^*)$ where $f(\theta) := \| D\theta \|_1$. By the characterization of $\partial f(\theta^*)$ given in Proposition 2.5, to prove that $v_0 \in \partial f(\theta^*)$, we need to prove that

$$v_0 + \cdots + v_n = 0,$$

and

$$v_0 + \cdots + v_n = 0 \quad \text{for } 1 \leq j \leq n$$

and

$$v_{0j} + \cdots + v_{0n} = r_u \quad \text{for } u = 1, \ldots, k.$$  

Each of these conditions follows from the calculation below. Fix $1 \leq j \leq n$ and let $0 \leq l \leq k$ be the unique integer such that $j_l \leq i < j_{l+1}$. Then

$$\sum_{i=j}^n v_{oi} = \sum_{i=j}^{j_{l+1}-1} v_{0i} + \sum_{i=j}^{j_{l+1}-1} v_{0i} = \sum_{i=j}^{j_{l+1}-1} \frac{v_r - v_{r+1}}{n_l} + \sum_{i=j}^{j_{l+1}-1} \frac{r_u - r_{u+1}}{n_u}$$

$$= \frac{v_r - v_{r+1}}{n_l} (j_{l+1} - j) + \sum_{u=l+1}^k (r_u - r_{u+1})$$

$$= \frac{v_r - v_{r+1}}{n_l} (j_{l+1} - j) + v_{r+1} (\frac{j_{l+1} - j}{n_l}) + v_{r+1} (\frac{j - j_l}{n_l}).$$

This proves $v_0 \in \partial f(\theta^*)$. We shall next prove that $v_0$ minimizes $\|v\|$ over $v \in \text{aff}(\partial f(\theta^*))$. This will automatically (because $v_0 \in \partial f(\theta^*)$) also prove that $v_0$ minimizes $\|v\|$ over $v \in \partial f(\theta^*)$ so that $v_0 = v^*$. Because $\text{aff}(\partial f(\theta^*))$ is an affine set and $v_0 \in \partial f(\theta^*)$, the fact that $v_0$ minimizes $\|v\|$ over $\text{aff}(\partial f(\theta^*))$ is equivalent to the condition:

$$\langle v - v_0, v \rangle = 0 \quad \text{for every } v \in \partial f(\theta^*). \quad (151)$$

Therefore we only need to verify (151). For this, write

$$\langle v - v_0, v \rangle = \sum_{u=0}^{k} \sum_{i=j_u}^{j_{u+1}-1} \left( v_i - \frac{r_u - r_{u+1}}{n_u} \right) \left( \frac{r_u - r_{u+1}}{n_u} \right)$$

$$= \sum_{u=0}^{k} \frac{r_u - r_{u+1}}{n_u} \left( \sum_{i=j_u}^{j_{u+1}-1} v_i \right) - \sum_{u=0}^{k} \frac{(r_u - r_{u+1})^2}{n_u}.$$  

The quantity above equals zero because, by the characterization of the subdifferential $\partial f(\theta^*)$, we have $\sum_{i=j_u}^{j_{u+1}-1} v_i = r_u - r_{u+1}$ for every $v \in \partial f(\theta^*)$ and $0 \leq u \leq k$. This proves that the condition (151) holds.
We now prove inequality (29). For this, simply write
\[
\|v_0\|^2 = \sum_{u=0}^{k} \sum_{i=j_u}^{j_u+1} \left( \frac{r_u - r_{u+1}}{n_u} \right)^2 \]
\[
= \sum_{u=0}^{k} \left( \frac{r_u - r_{u+1}}{n_u} \right)^2 = \frac{1}{n_0} + \frac{1}{n_k} + 4 \sum_{u=1}^{k-1} I\{r_u \neq r_{u+1}\}
\]
because \((r_u - r_{u+1})^2 = 1\) for \(u = 0, k\) and \(4I\{r_u \neq r_{u+1}\}\) for all other \(u\). This proves (29) and completes the proof of the first part of Lemma 2.7 (for \(r = 1\)).

**Proof of inequality (30).** Fix \(\theta^* \in \mathbb{R}^n\) with \(D(r)\theta^* \neq 0_{n-r}\). Note that \(v_0\) is the projection of the zero vector \(0_n\) onto \(\text{aff}(\partial f(\theta^*))\).

Because \(\partial f(\theta^*)\) is given by a finite number of linear inequalities (i.e., it is a polyhedron), its affine hull is given by the intersection of the inequalities which are actually equalities (see, for example, Schrijver [43, Chapter 8]). Therefore, \(\text{aff}(\partial f(\theta^*))\) is given by the vectors \(v \in \mathbb{R}^n\) for which (23) holds and for which
\[
a_j = \sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r - 1} \right) v_i = \text{sgn}((D(r)\theta^*)_{j-r})
\]
for \(r < j \leq n\) such that \((D(r)\theta^*)_{j-r} \neq 0\). Let the number of \(r^{th}\) order knots of \(\theta^*\) be \(k\) so that the number of equalities in \(\text{aff}(\partial f(\theta^*))\) is \(k + r\). We can represent these equalities in matrix form as \(Bv = b\) where \(B\) is \((k + r) \times n\) and \(b \in \mathbb{R}^{k+r}\) with \(\|b\|_1 = k\). Note also that \(\max_{i,j} |B_{ij}| \leq \binom{n+r-2}{r-1}\) so that
\[
\|B\|_1 := \sup_{x \neq 0} \frac{\|Bx\|_1}{\|x\|_1} = \max_{1 \leq j \leq n} \sum_{i=1}^{k+r} |B_{ij}| \leq (k + r) \binom{n + r - 2}{r - 1} \leq \frac{(r + 1)k}{(r - 1)!} (2n)^{r-1}.
\]
As a result, because the vector \(v_0\) satisfies \(Bv_0 = b\), we obtain
\[
\|v_0\| \geq \frac{\|v_0\|_1}{\sqrt{n}} \geq \frac{\|b\|_1}{\sqrt{n} \|B\|_1} \geq \frac{k}{\sqrt{n} (r-1)! (r+1)k(2n)^{r-1}} = \frac{(r-1)!}{(r+1)2^{r-1} n^{-r+1/2}}.
\]
This proves (30).

**Proof of Inequality (31).** This proof is rather long. Fix \(\theta^* \in \mathbb{R}^n\) with \(D(r)\theta^* \neq 0_{n-r}\). Let \(2 \leq j_1 < \cdots < j_k \leq n - r + 1\) be the \(r^{th}\) order knots of \(\theta^*\) along with associated signs \(r_1, \ldots, r_k \in \{-1, 1\}\). Also let \(j_0 = 1, j_{k+1} = n - r + 2\) and \(v_0 = r_{k+1} = 0\). It will be convenient below to take \(m_i := j_l + r - 1\) for \(l = 0, \ldots, k\). Also let \(n_0 = j_1 + r - 2\) and \(n_i = j_{i+1} - j_i\) for \(i = 1, \ldots, k\).

Because it is assumed that the minimum length condition (13) holds for \(\theta^*\) with constant \(c\), it follows that \(n_i \geq cn/(k + 1)\) whenever \(r_i \neq r_{i+1}\).

Let \(g : \mathbb{R} \to \mathbb{R}\) be a smooth (i.e., \(C^\infty\)) function such that
1. \( g(0) = 0, g(1) = 1 \).
2. \( g^{(j)}(0) = g^{(j)}(1) = 0 \) for \( j \geq 1 \).
3. \( g(t) \in [0, 1] \) for \( t \in [0, 1] \).

where \( g^{(j)} \) is the \( j^{th} \) order derivative of \( g \). For example, the function \( g(x) := \int_0^x \phi(t)dt \) where

\[
\phi(t) = \begin{cases} 
\gamma \exp\left( \frac{-1}{t(1-t)} \right) & \text{for } t \in (0,1) \\
0 & \text{otherwise}
\end{cases}
\]

where \( \gamma \) is chosen so that \( \int_0^1 \phi(t)dt = 1 \) will satisfy the requirements for \( g \).

Let us now define a function \( S : [1, n+r] \to \mathbb{R} \) as follows:

\[
S(t) = \begin{cases} 
0 & \text{for } t \in [1, r] \cup [n+1, n+r] \\
\mathbf{v}_i \left( 1 - g \left( \frac{t-m_i}{n_i} \right) \right) + \mathbf{v}_{i+1} g \left( \frac{t-m_i}{n_i} \right) & \text{for } m_i \leq t \leq m_{i+1}, 0 \leq i \leq k
\end{cases}
\]

By an abuse of notation, we shall also denote by \( S \), the \( n+r \)-dimensional vector \((S(1), \ldots, S(n+r))\). It will be clear from the context whether we are referring to the vector \( S \) or the function \( S \). From the properties of \( g \), it is easy to deduce that \( S(m_i) = \mathbf{v}_i \) and \( S^{(j)}(m_i) = 0 \) for all \( j \) and \( 0 \leq i \leq k+1 \). Also \( \sup_{t \in [1, n+r]} |S(t)| \leq 1 \).

The first key observation is that the vector \( \mathbf{v}^* \in \mathbb{R}^n \) defined by

\[
\mathbf{v}_j^* := (-1)^r (D^{(r)}S)_j = \sum_{k=j}^{j+r} (-1)^{k-j} \binom{r}{k-j} S_k
\]

for \( 1 \leq j \leq n \) belongs to the subdifferential \( \partial f(\theta^*) \). To see this, we need to use Proposition 2.5. Note that for \( 1 \leq j \leq n \),

\[
a_j^* := \sum_{i=j}^{n} \binom{r+i-j-1}{r-1} \mathbf{v}_i^* = \sum_{i=j}^{n+r} (-1)^{i-r} \binom{r}{k-i} S_k
\]

where the last equality follows because \( S_k = 0 \) for \( k = n+1, \ldots, n+r \). Now let

\[
\beta_i := \binom{r+i-j-1}{r-1} \quad \text{for } i = \ldots, -2, -1, 0, 1, 2, \ldots
\]

where the binomial coefficient is taken to be zero if \( r+i-j-1 < r-1 \). Then

\[
a_j^* = \sum_{k=j}^{n} S_k \sum_{i=k-r}^{k} (-1)^{k-i} \binom{r}{k-i} \beta_i = \sum_{k=j}^{n} S_k (D^{(r)}\beta)_{k-r}.
\]
It is now easy to see that $\beta_i$ is a polynomial in $i$ for $i \geq j + 1 - r$ which implies that $(D^{(r)}\beta)_{k-r} = 0$ for $k \geq j + 1$. It can also be checked that $(D^{(r)}\beta)_{j-r} = 1$. This therefore gives $a_j^* = S_j$ for $j = 1, \ldots, n$. Proposition 2.5 and the fact that $S_j = S(j) = 0$ for $1 \leq j \leq r$, $S(m_i) = \tau_i$ and $|S(t)| \leq 1$ for all $t$ proves that $v^* \in \partial f(\theta^*)$.

We shall now bound $\|v^*\|$ by writing

$$
\|v^*\|^2 = \sum_{l=0}^{k+1} \sum_{j=m_l-r+1}^{m_l-1} v_j^2 + \sum_{l=0}^{k} \sum_{j=m_l}^{m_{l+1}-r} v_j^2
$$

$$
= \sum_{l=0}^{k+1} \sum_{j=m_l-r+1}^{m_l-1} ( (D^{(r)}S)_j )^2 + \sum_{l=0}^{k} \sum_{j=m_l}^{m_{l+1}-r} ( (D^{(r)}S)_j )^2
$$

Let

$$
M_r := \sup_{t \in [0,1]} |g^{(r)}(t)| ,
$$

and note that

$$
|S^{(r)}(t)| \leq |\tau_{l+1} - \tau_l| M_r n^{-r} \leq 2M_r n^{-r}_min
$$

for $t \in [m_l, m_{l+1}]$ and $0 \leq l \leq k$ where

$$
n_{min} := \min_{0 \leq i \leq k; r \neq t_i+1} n_i .
$$

Then for $m_l - r < j < m_l + r$ and $0 \leq l \leq k + 1$ we have

$$
|S(j)| \leq \frac{2M_r n^{-r}_{min} |j - m_l|^r}{r!} \leq \frac{2r^r}{r!} M_r n^{-r}_{min} \leq \frac{2e^r}{\sqrt{2\pi r}} M_r n^{-r}_{min}
$$

by $(r-1)$-th order Taylor expansion about $m_l$ and Stirling’s approximation. (The bound trivially holds if $j < r$ or $j > n$; if $j \notin (m_{l-1}, m_{l+1})$, then the bound holds by expansion about the nearest $m_i$. Thus for $m_l - r < j < m_l$ and $0 \leq l \leq k + 1$, again by Stirling’s approximation, we have

$$
|(D^{(r)}S)_j| \leq \sum_{i=0}^{r} \binom{r}{i} \frac{2e^r}{\sqrt{2\pi r}} M_r n^{-r}_{min} \leq \frac{2^{r+1}e^r}{\sqrt{2\pi r}} M_r n^{-r}_{min}
$$

and so

$$
\sum_{l=0}^{k+1} \sum_{j=m_l-r+1}^{m_l-1} ( (D^{(r)}S)_j )^2 \leq \sum_{l=0}^{k+1} (r - 1) \frac{2^{2r+2}e^{2r}}{\pi} M_r^2 n^{-2r}_{min}
$$

$$
\leq \frac{2(k+1)}{\pi} (2e)^{2r} M_r^2 n^{-2r}_{min}
$$

$$
\leq (2e)^{2r} M_r^2 (k+1) n^{-2r}_{min} .
$$

We now proceed to the second term for bounding $\|v^*\|$. For this, let

$$
N_r = \sup_{t \in [0,1]} |g^{(r+1)}(t)| ,
$$
and note that
\[ |S^{(r+1)}(t)| \leq |v_{l+1} - v_l| N_r n_t^{-r-1}. \]
for \( t \in [m_l, m_{l+1}] \) and \( 0 \leq l \leq k \). Then for \( m_l \leq j \leq m_{l+1} - r \) and \( 0 \leq l \leq k \),
\[
|(-1)^r (D^{(r)} S)_j - S^{(r)}(j)| \leq \sum_{i=0}^{r} \binom{r}{i} \frac{|v_{l+1} - v_l| N_r n_t^{-r-1}}{(r+1)!} r^{r+1} \\
\leq \frac{2^r r^{r+1}}{(r+1)!} |v_{l+1} - v_l| N_r n_t^{-r-1} \\
\leq \frac{2^r e^{r+1}}{\sqrt{2\pi (r+1)}} |v_{l+1} - v_l| N_r n_t^{-r-1} \\
\leq \frac{2^{r-1} e^{r+1}}{\sqrt{\pi}} |v_{l+1} - v_l| N_r n_t^{-r-1},
\]
by \( r \)-th order Taylor expansion about \( j \) and Stirling’s approximation, using the fact that
the \( r \)-th order forward difference approximates the \( r \)-th derivative up to an error depending
on the \((r+1)\)-th derivative (i.e. all lower order terms in the Taylor expansion cancel).
Then the trivial inequality \(|a^2 - b^2| \leq (a - b)^2 + 2|a - b|\) gives, for
\[
T_j := \left| \left( (D^{(r)} S)_j \right)^2 - \left( S^{(r)}(j) \right)^2 \right|,
\]
the upper bound
\[
T_j \leq \left| (-1)^r (D^{(r)} S)_j - S^{(r)}(j) \right|^2 + 2 \left| S^{(r)}(j) \right| \left| (-1)^r (D^{(r)} S)_j - S^{(r)}(j) \right| \\
\leq \frac{2^{2r-2} e^{2r+2}}{\pi} (v_{l+1} - v_l)^2 N_r^2 n_t^{-2r-2} + \frac{2^r e^{r+1}}{\sqrt{\pi}} (v_{l+1} - v_l)^2 M_r N_r n_t^{-2r-1} \\
\leq (2e)^{2r} (v_{l+1} - v_l)^2 (M_r + N_r) N_r n_t^{-2r-1}.
\]
So for \( 0 \leq l \leq k \) we have,
\[
\left| \sum_{j=m_l}^{m_{l+1}-r} (D^{(r)} S)_j \right|^2 - \sum_{j=m_l}^{m_{l+1}-r} \left( S^{(r)}(j) \right)^2 \\
\leq (n_l - r + 1)(2e)^{2r} (v_{l+1} - v_l)^2 (M_r + N_r) N_r n_t^{-2r-1} \\
\leq (2e)^{2r} (v_{l+1} - v_l)^2 (M_r + N_r) N_r n_t^{-2r} \leq 4(2e)^{2r} (M_r + N_r) N_r n_{\min}^{-2r}
\]
(the above bound trivially holds if \( n_l < r \)). Thus
\[
\sum_{l=0}^{k} \sum_{j=m_l}^{m_{l+1}-r} (D^{(r)} S)_j^2 \leq \sum_{l=0}^{k} \sum_{j=m_l}^{m_{l+1}-r} \left( S^{(r)}(j) \right)^2 \\
+ 4(2e)^{2r} (M_r + N_r) N_r (k+1) n_{\min}^{-2r} \tag{153}
\]
Now let
\[
K_r = \sup_{t \in [0,1]} \left| \frac{d}{dt} \left( (g^{(r)}(t))^2 \right) \right|,
\]
and note that
\[
\left| \frac{d}{dt} \left( (S^{(r)}(t))^2 \right) \right| \leq (r_{l+1} - r_l)^2 K_r n_t^{-2r-1}
\]
for \( t \in [m_l, m_{l+1}] \) and \( 0 \leq l \leq k \), regarding the derivative as one-sided at the endpoints. Then for \( m_l \leq j \leq m_{l+1} - r \) and \( 0 \leq l \leq k \),
\[
\left| (S^{(r)}(j))^2 - \int_j^{j+1} (S^{(r)}(t))^2 \, dt \right| \leq \int_j^{j+1} (r_{l+1} - r_l)^2 K_r n_t^{-2r-1}(t - j) \, dt
\]
by a zeroth order Taylor expansion about \( j \). So for \( 0 \leq l \leq k \) we have
\[
\left| \sum_{j=m_l}^{m_{l+1}-r} (S^{(r)}(j))^2 - \int_{m_l}^{m_{l+1}-r+1} (S^{(r)}(t))^2 \, dt \right| 
\leq (n_l - r + 1) \frac{1}{2} (r_{l+1} - r_l)^2 K_r n_l^{-2r-1}
\leq \frac{1}{2} (r_{l+1} - r_l)^2 K_r n_l^{-2r} \leq 2K_r n_{\min}^{-2r}
\]
(the bound trivially holds if \( n_l < r \).) Thus
\[
\sum_{l=0}^{k} \sum_{j=m_l}^{m_{l+1}-r} (S^{(r)}(j))^2 \leq \sum_{l=0}^{k} \int_{m_l}^{m_{l+1}-r+1} (S^{(r)}(t))^2 \, dt + 2K_r (k+1) n_{\min}^{-2r}.
\]
Let
\[
I_r = \int_0^1 \left( g^{(r)}(t) \right)^2 \, dt,
\]
and note that for \( 0 \leq l \leq k \),
\[
\int_{m_l}^{m_{l+1}-r+1} (S^{(r)}(t))^2 \, dt \leq \int_{m_l}^{m_{l+1}} (S^{(r)}(t))^2 \, dt
\]
\[
= \int_{m_l}^{m_{l+1}} (r_{l+1} - r_l)^2 \left( g^{(r)} \left( \frac{t - m_l}{n_l} \right) \right)^2 n_t^{-2r} \, dt
\]
\[
= (r_{l+1} - r_l)^2 I_r n_l^{-2r} \leq 4I_r n_{\min}^{-2r+1}.
\]
Thus
\[
\sum_{l=0}^{k} \int_{m_l}^{m_{l+1}-r+1} (S^{(r)}(t))^2 \, dt \leq 4I_r (k+1) n_{\min}^{-2r+1}.
\]
Combining bounds (152), (153), (154), and (155), we have
\[
\| v^* \|^2 \leq (2e)^{2r} M_r^2 (k+1) n_{\min}^{-2r} + 4(2e)^{2r} (M_r + N_r) (k+1) n_{\min}^{-2r}
+ 2K_r (k+1) n_{\min}^{-2r} + 4I_r (k+1) n_{\min}^{-2r+1}
\leq (2e)^{2r} (M_r + 2N_r)^2 + 2K_r + 4I_r) (k+1) n_{\min}^{-2r+1}.
\]
This proves (31) with \( C_r = \sqrt{(2e)^2(M_r + 2N_r)^2 + 2K_r + 4I_r} \) (because of the fact that \( n_{\min} \geq cn/(k + 1) \) under assumption (13)).

Appendix D: Additional technical results and proofs

D.1. A result on Gaussian suprema

The following result was used in the proof of Theorem 2.2.

Lemma D.1. Suppose \( p, n \geq 1 \) and let \( \Theta_1, \ldots, \Theta_p \) be subsets of \( \mathbb{R}^n \) each containing the origin and each contained in the closed Euclidean ball of radius \( D \) centered at the origin. Then, for \( \xi \sim N(0, \sigma^2 I) \), we have

\[
E \left( \max_{1 \leq i \leq p} \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \right) \leq \max_{1 \leq i \leq p} E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle + D \sigma \left( \sqrt{2 \log p} + \sqrt{\frac{\pi}{2}} \right). \tag{156}
\]

Proof of Lemma D.1. For every \( t \geq 0 \), by the union bound

\[
\mathbb{P} \left\{ \max_{1 \leq i \leq p} \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \geq \max_{1 \leq i \leq p} E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle + t \sigma \right\} \leq \sum_{i=1}^p \mathbb{P} \left\{ \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \geq E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle + t \sigma \right\}.
\]

Now by hypothesis, every vector in \( \Theta_i \) has norm bounded by \( D \). As a result, the map \( \xi \mapsto \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \) is Lipschitz with constant \( D \). By the Gaussian concentration inequality, we deduce therefore that

\[
\mathbb{P} \left\{ \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \geq E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle + t \sigma \right\} \leq \exp \left( -\frac{t^2}{2D^2} \right)
\]

for every \( 1 \leq i \leq p \). Consequently,

\[
\mathbb{P} \left\{ \max_{1 \leq i \leq p} \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \geq \max_{1 \leq i \leq p} E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle + t \sigma \right\} \leq \min \left\{ p \exp \left( -\frac{t^2}{2D^2} \right), 1 \right\}
\]

for every \( t \geq 0 \). Integrating both sides of this inequality from \( t = 0 \) to \( t = \infty \), we obtain

\[
E \left( \max_{1 \leq i \leq p} \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle - \max_{1 \leq i \leq p} E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \right)^+ \leq \sigma \int_0^\infty \min \left\{ p \exp \left( -\frac{t^2}{2D^2} \right), 1 \right\} dt.
\]

The trivial inequality \( a \leq b + (a - b)^+ \) therefore gives

\[
E \left( \max_{1 \leq i \leq p} \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle \right) \leq \max_{1 \leq i \leq p} E \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle + \sigma \int_0^\infty \min \left\{ p \exp \left( -\frac{t^2}{2D^2} \right), 1 \right\} dt.
\]
We will now bound the integral from above. For this, we simply write
\[
\int_0^\infty \min \left\{ p \exp \left( -\frac{t^2}{2D^2} \right), 1 \right\} dt \leq \int_0^{\sqrt{2\log p}} 1 \, dt
\]
\[
+ p \int_{\sqrt{2\log p}}^\infty \exp \left( -\frac{t^2}{2D^2} \right) dt
\]
\[
= D\sqrt{2\log p} + \sqrt{2\pi p} D \left( 1 - \Phi(\sqrt{2\log p}) \right).
\]

We now complete the proof of \((156)\) via the Gaussian tail bound \(1 - \Phi(x) \leq \exp(-x^2/2)/2\) for \(x = \sqrt{2\log p}\) (see e.g., Dumbgen [11]).

\[\Box\]

**D.2. A formula for \(\theta\) in terms of \(D^{(r)}\theta\)**

The following result provides a formula for expressing a vector \(\theta \in \mathbb{R}^n\) in terms of \(D^{(r)}\theta\) and \((D^{(i)}\theta)_1\) for \(i = 0, \ldots, r-1\). This result is quite useful and we have used it in multiple places in our proofs.

**Lemma D.2.** Fix \(r \geq 1\) and \(n \geq r\). For every \(\theta \in \mathbb{R}^n\) and \(1 \leq i \leq n\), we have
\[
\theta_i = \sum_{j=1}^{i-r} \binom{i-j-1}{r-1} (D^{(r)}\theta)_j + \sum_{j=1}^{r} \binom{i-1}{j-1} (D^{(j-1)}\theta)_1
\]
(157)

where we take the convention that \(\binom{a}{b} = 0\) for \(b > a\), \(\binom{0}{0} = 1\) and that the first term in the right hand side is zero unless \(i > r\).

**Proof of Lemma D.2.** We shall use induction on \(r \geq 1\). For \(r = 1\), the formula \((157)\) becomes
\[
\theta_i = \sum_{j=1}^{i-1} (D\theta)_j + \theta_1
\]
(158)

which is trivial because \((D\theta)_j = \theta_{j+1} - \theta_j\).

Let us now assume that \((157)\) is true for some \(r = \ell \geq 1\) and we shall then prove it for \(r = \ell + 1\). Because \((157)\) is true for \(r = \ell\), we have
\[
\theta_i = \sum_{j=1}^{i-\ell} \binom{i-j-1}{\ell-1} (D^{(\ell)}\theta)_j + \sum_{j=1}^{\ell} \binom{i-1}{j-1} (D^{(j-1)}\theta)_1.
\]
(159)

Inequality \((158)\) for \(\theta\) replaced by \(D^{\ell}\theta\) gives
\[
(D^{\ell}\theta)_j = (D^{\ell}\theta)_1 + \sum_{k=1}^{j-1} (D^{(\ell+1)}\theta)_k.
\]
Using the above identity in (159), we obtain

\[
\theta_i = \sum_{j=1}^{i-\ell} \binom{i-j-1}{\ell-1} (D^\ell \theta)_1 + \sum_{k=1}^{j-1} (D^{(\ell+1)} \theta)_k + \sum_{j=1}^{\ell} \binom{i-1}{j-1} (D^{(\ell-1)} \theta)_1 \\
= \sum_{j=1}^{i-\ell} \sum_{k=1}^{j-1} \binom{i-j-1}{\ell-1} (D^{(\ell+1)} \theta)_k + (D^\ell \theta)_1 \sum_{j=1}^{i-\ell} \binom{i-j-1}{\ell-1} \\
+ \sum_{j=1}^{\ell} \binom{i-1}{j-1} (D^{(\ell-1)} \theta)_1
\]

(160)

We now use the elementary identity (146) involving binomial coefficients to obtain

\[
\sum_{j=k+1}^{i-\ell} \binom{i-j-1}{\ell-1} = \binom{i-k-1}{\ell-1} \quad \text{and} \quad \sum_{j=1}^{i-\ell} \binom{i-j-1}{\ell-1} = \binom{i-1}{\ell}.
\]

From the above and (160), we deduce that

\[
\theta_i = \sum_{k=1}^{i-\ell-1} (D^{(\ell+1)} \theta)_k \binom{i-k-1}{\ell} + \sum_{j=1}^{\ell+1} \binom{i-1}{j-1} (D^{(\ell-1)} \theta)_1
\]

which is exactly (157) for \( r = \ell + 1 \). This completes the proof of Lemma D.2. \( \square \)

**D.3. Strong Sparsity and Discrete Splines**

The following result gives a connection between sparsity of the vector \( D^{(r)} \theta \) and discrete splines.

**Proposition D.3.** Suppose \( \theta \in \mathbb{R}^n \) with \( \|D^{(r)} \theta\|_0 = k \). Then \( \theta \) equals \( (p(1/n), \ldots, p((n-1)/n), p(1)) \) for a discrete spline \( p \) that is made of \( k + 1 \) polynomials each of degree \( (r-1) \).

The proof of Proposition D.3 is given below. Note that the result is trivial when \( r = 1 \). So it may well be assumed that \( r \geq 2 \) in the rest of this subsection. In fact, the argument below will also hold for \( r = 1 \) provided the involved binomial coefficients are interpreted correctly for \( r = 1 \).

The following lemma will be used in the proof of Proposition D.3.
Lemma D.4. Let \( r \geq 1 \), \( n \geq r \) and \( 1 \leq a \leq b - 1 \leq n - r + 1 \). Suppose that
\[
(D^{(r-1)}\theta)_a = \cdots = (D^{(r-1)}\theta)_{b-1} = c. \quad (161)
\]
Then
\[
\theta_i = c\binom{i-a}{r-1} + \sum_{j=1}^{r-1} \binom{i-a}{j-1} (D^{(j-1)}\theta)_a
\]
\[
(162)
\]
for every \( i = a, \ldots, r + b - 2 \).

Proof of Lemma D.4. Let \( \alpha \) be the \( b - a + r - 1 \)-dimensional vector defined by
\[
\alpha = (\theta_a, \theta_{a+1}, \ldots, \theta_{b+r-2}).
\]
Then \( (D^{(r-1)}\alpha)_u = (D^{(r-1)}\theta)_{a+u-1} \) for \( u = 1, \ldots, b - a \) and hence we have \( (D^{(r-1)}\alpha)_1 = \cdots = (D^{(r-1)}\alpha)_{b-a} = c \) because of (161). An application of Lemma D.2 now gives
\[
\alpha_u = c\sum_{j=1}^{u-r+1} \binom{u-j-1}{r-2} + \sum_{j=1}^{r-1} \binom{u-1}{j-1} (D^{(j-1)}\alpha)_1
\]
for \( u = 1, \ldots, r - 1 + b - a \). The elementary inequality (146) applied to \( a = u - 2 \) and \( b = r - 2 \) allows us to deduce
\[
\alpha_u = c\binom{u-1}{r-1} + \sum_{j=1}^{r-1} \binom{u-1}{j-1} (D^{(j-1)}\alpha)_1
\]
for \( u = 1, \ldots, r - 1 + b - a \). Applying the above to \( u = i + 1 - a \), we obtain inequality (162). This completes the proof of Lemma D.4.

We now prove Proposition D.3.

Proof of Proposition D.3. Suppose \( \theta \in \mathbb{R}^n \) and let \( 2 \leq j_1 < \cdots < j_k \leq n - r + 1 \) denote all the \( r^{th} \) order knots of \( \theta \) with \( j_0 = 1 \) and \( j_{k+1} = n - r + 2 \). We then have
\[
(D^{(r-1)}\theta)_{j_u} = \cdots = (D^{(r-1)}\theta)_{j_{u+1}-1} = c_u \quad \text{for } u = 0, \ldots, k
\]
for some real numbers \( \{c_u, 0 \leq u \leq k\} \).

Lemma D.4 applied to \( a = j_u \) and \( b = j_{u+1} \) then implies that for every \( 0 \leq u \leq k \) and \( i = j_u, \ldots, r + j_{u+1} - 2 \), we have
\[
\theta_i = c_u \binom{i-j_u}{r-1} + \sum_{j=1}^{r-1} \binom{i-j_u}{j-1} (D^{(j-1)}\theta)_{j_u}
\]
\[
(163)
\]
Now, for each $0 \leq u \leq k$, let $p_u$ denote the polynomial in $x$ defined by

$$p_u(x) := \frac{c_u}{(r - 1)!} (nx - j_u) \ldots (nx - j_u - r + 2) + \sum_{j=1}^{r-1} \frac{(nx - j_u) \ldots (nx - j_u - j + 2)(D^{(j-1)}\theta)_{j_u}}{(j - 1)!}.$$

It is clear that $p_u(x)$ is a polynomial in $x$ of degree $(r - 1)$. Also the identity (163) is equivalent to

$$\theta_i = p_u(i/n) \quad \text{for } 0 \leq u \leq k \text{ and } j_u \leq i \leq r + j_{u+1} - 2. \quad (164)$$

We now define a function $p$ via

$$p(x) = \begin{cases} 
p_0(x) & \text{for } x < \frac{r + j_1 - 2}{n} \\
p_u(x) & \text{for } \frac{r + j_u - 2}{n} \leq x < \frac{r + j_{u+1} - 2}{n}, \, u = 1, \ldots, k - 1 \\
p_k(x) & \text{for } x \geq \frac{r + j_k - 2}{n}.
\end{cases}$$

Clearly $p$ is a piecewise polynomial of degree $(r - 1)$. Also, it is trivial to see from (164) that $p(i/n) = \theta_i$ for every $1 \leq i \leq n$. Moreover, using (164), it is easy to show that one has

$$p_{u-1}\left(\frac{i}{n}\right) = p_u\left(\frac{i}{n}\right) \quad \text{for } 1 \leq u \leq k \text{ and } j_u \leq i \leq r + j_u - 2 \quad (165)$$

for $r \geq 2$. Thus if $x_u := (r + j_u - 2)/n$ denotes the knots of the piecewise polynomial $p$, then we have

$$p_{u-1}\left(x_u - \frac{i}{n}\right) = p_u\left(x_u - \frac{i}{n}\right) \quad \text{for } i = 0, 1, \ldots, r - 2. \quad (166)$$

This means that the function $p$ is a discrete spline of degree $(r - 1)$ having $k+1$ polynomial pieces which proves Proposition D.3. \qed

**D.4. A result on the magnitude of $\min_{1 \leq i \leq n-r+1} (D^{(r-1)}\theta)_i$ when $\|\theta\| \leq 1$**

This section is devoted to the proof of the Lemma C.6 which was crucially used in the proof of Lemma C.5.

**Proof of Lemma C.6.** We only need to prove the first inequality in (142). The second inequality follows by applying the first inequality to $-\theta$.

Via Lemma D.2, we can write the following for every $\theta \in \mathbb{R}^n$ with $\|\theta\| \leq t$:

$$t^2 \geq \|\theta\|^2 = \sum_{i=1}^{n} \left( \sum_{j=1}^{i-r} \left( i - j - 1 \right) s_j + \sum_{j=1}^{r} \left( i - 1 \right) (D^{(j-1)}\theta)_1 \right)^2.$$
where \( s_j := (D^{(r)} \theta)_j \) for \( j = 1, \ldots, n - r \). It follows from here that
\[
t^2 \geq \inf_{\beta_1, \ldots, \beta_r \in \mathbb{R}} \sum_{i=1}^{n} \left( \sum_{j=1}^{i-r} \left( \frac{i-j-1}{r-1} \right) s_j - \sum_{j=1}^{r} \left( \frac{i-1}{j-1} \right) \beta_j \right)^2.
\]
We now define two matrices. Let \( X \) be the \( n \times r \) matrix whose \((i, j)\)th entry equals \( \binom{i-1}{j-1} \). Let \( S \) be the \( n \times (n - r) \) matrix whose \((i, j)\)th entry equals \( \binom{i-j-1}{r-1} \). Throughout we use the convention that \( \binom{a}{b} = 0 \) when \( a < b \). Also let \( s := (D^{(r)} \theta) = (s_1, \ldots, s_{n-r})^T \) and \( \beta := (\beta_1, \ldots, \beta_r) \). It is then easy to see from the previous inequality that
\[
t^2 \geq \inf_{\beta_1, \ldots, \beta_r \in \mathbb{R}} \| Ss - X\beta \|^2 = s^T S^T (I - P_X) Ss \tag{167}
\]
where \( P_X = X(X^TX)^{-1}X^T \) is the projection matrix on to the column space of \( X \).

We now need the following two facts about the matrix \( A := S^T(I - P_X)S \). These facts (whose proofs are long) are proved in Proposition D.7 and Proposition D.8 respectively.

1. If \( \mathbf{1} \) denotes the \( n - r \) vector consisting of ones, then \( \mathbf{1}^T A \mathbf{1} \geq C_r n^{2r+1} \) for a constant \( C_r \) depending on \( r \) alone.

2. Every entry of the matrix \( A \) is positive.

We shall now complete the proof of Lemma C.6 assuming the above two facts about the matrix \( A \). Let \( \delta := \min_{1 \leq j \leq n-r} s_j \). Our goal is to prove that \( \delta \leq C_r n^{r-1/2} \) so we can assume that \( \delta \geq 0 \) for otherwise there is nothing to prove. In that case, inequality (167) and the second fact about \( A \) together imply
\[
t^2 \geq \delta^2 \mathbf{1}^T S^T (I - P_X) S \mathbf{1} = \delta^2 \mathbf{1}^T A \mathbf{1}.
\]
The first fact about \( A \) then gives \( t^2 \geq C_r \delta^2 n^{2r+1} \) and this completes the proof of Lemma C.6.

The remainder of this subsection is devoted to proving the two facts about the matrix \( A := S^T(I - P_X)S \) stated in the proof of Lemma C.6. These proofs are tedious and long. We adopt the convention that \( \binom{n}{k} = \frac{(n)_k}{k!} \) if \( k \geq 0 \) and 0 otherwise, where \( (n)_k \) is the falling factorial, extending the definition of the binomial coefficient to integer arguments. We will make judicious use of the identities \( \binom{n}{k} = \binom{n-k}{k} \) and \( \binom{n}{k} = (-1)^k \binom{k-n-1}{k} \), as well as the Chu-Vandermonde identity, \( \binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k} \binom{n}{r-k} \), in its equivalent form \( \binom{m+n}{r-s} = \sum_{k=s}^{r} \binom{m}{k-s} \binom{n}{r-k} \).

Recall that \( X \) is the \( n \times r \) matrix with \( X_{ij} = \binom{i-j-1}{r-1} \), \( S \) is the \( n \times (n - r) \) matrix with \( S_{ij} = \binom{i-j-1}{r-1} \) if \( i - j \geq r \) and 0 otherwise, and \( A = S^T(I - P_X)S \) where \( P_X \) is the projection onto the column space of \( X \). Our first step is to compute the inverse of the matrix \( A \) explicitly. This is the content of the following Proposition.

**Proposition D.5.** Let \( T \) be the \((n-r) \times (n-r)\) matrix with \( T_{ij} = (-1)^{i-j} \binom{2r}{r+i-j} \). Then \( T = A^{-1} \).
In order to prove Proposition D.5, we need the following lemma.

**Lemma D.6.** Let $Y$ be the $r \times (n - r)$ matrix with $Y_{ij} = (-1)^{r+i-j}(r+i-1)_{i-j}$, and let $U$ be $n \times (n - r)$ matrix with $U_{ij} = (-1)^{r+i-j}(r+i-1)_{i-j}$. Then $XY + ST = U$.

**Proof of Lemma D.6.** We have

$$(XY + ST)_{ij} = \sum_{k=1}^{r} X_{ik}Y_{kj} + \sum_{l=1}^{n-r} S_{il}T_{lj}$$

$$= \sum_{k=j}^{r} (-1)^{r+k-j}\binom{i-1}{i-k}\binom{r+k-1}{k-j} + \sum_{l=1}^{i-r} (-1)^{l-j}\binom{i-1}{i-l-1}\binom{2r}{r+l-j}$$

$$= (-1)^r \sum_{k=j}^{i} \binom{i-1}{i-k}\binom{-r-j}{k-j} + (-1)^{r+i-j} \sum_{l=1}^{i-r} \binom{-r}{i-l-r}\binom{2r}{r+l-j}.$$ 

If $i < j$, then at least one of $i - k$, $k - j$ is negative, since $(i - k) + (k - j) = i - j < 0$. Hence $(XY)_{ij} = 0$, and similarly $(ST)_{ij} = 0$, so $(XY + ST)_{ij} = 0 = U_{ij}$. Otherwise, there are three cases. If $j \leq i \leq r$, then $(ST)_{ij} = 0$ since the sum is empty and

$$(XY)_{ij} = (-1)^r \sum_{k=j}^{i} \binom{i-1}{i-k}\binom{-r-j}{k-j}$$

$$= (-1)^r \binom{-r+i-j-1}{i-j} = (-1)^{r+i-j}\binom{r}{i-j} = U_{ij}.$$ 

If $r < j \leq i$, then $(XY)_{ij} = 0$ since the sum is empty, and

$$(ST)_{ij} = (-1)^{r+i-j} \sum_{l=j-r}^{i-r} \binom{-r}{i-l-r}\binom{2r}{r+l-j}$$

$$= (-1)^{r+i-j}\binom{r}{i-j} = U_{ij}.$$
Finally, if $j \leq r < i$, then

\[
(U - XY)_{ij} = U_{ij} - (-1)^r \sum_{k=j}^{i} \binom{i-1}{i-k} \binom{-r-j}{k-j}
\]

\[
+ (-1)^r \sum_{k=r+1}^{i-r} \binom{i-1}{i-k} \binom{-r-j}{k-j}
\]

\[
= (-1)^r \sum_{k=1}^{i-r} \binom{i-1}{i-k-r} \binom{-r-j}{r+k-j}
\]

\[
= \sum_{k=1}^{i-r} (-1)^{k-j} \binom{i-1}{i-k-r} \binom{2r+k-1}{r+k-j}
\]

\[
= \sum_{k=1}^{i-r} (-1)^{k-j} \binom{i-1}{i-k-r} \binom{2r+k-1}{r+j-1}
\]

\[
= \sum_{k=1}^{i-r} (-1)^{k-j} \binom{i-1}{i-k-r} \sum_{l=1}^{r+j} \binom{2r}{r+j-l} \binom{k-1}{l-1}
\]

\[
= \sum_{k=1}^{i-r} (-1)^{k-j} \binom{i-1}{i-k-r} \sum_{l=1}^{i-r} \binom{2r}{r+l-j} \binom{k-1}{k-l}
\]

\[
= \sum_{l=1}^{i-r} (-1)^{l-j} \binom{2r}{r+l-j} \sum_{k=l}^{i-r} \binom{i-1}{i-k-r} \binom{-l}{k-l}
\]

\[
= \sum_{l=1}^{i-r} (-1)^{l-j} \binom{2r}{r+l-j} \binom{i-l-1}{i-l-r} = (ST)_{ij}
\]

where the sixth equality above follows from the fact that \( \binom{2r}{r+j-l} = 0 \) for \( l > r + j \) and \( \binom{k-1}{l-1} = 0 \) for \( l > k \). \( \square \)

We are now ready to prove Proposition D.5.

**Proof of Proposition D.5.** Let $Y$ and $U$ be defined as in Lemma D.6. Note that

\[
(X^T U)_{ij} = \sum_{k=1}^{n} X_{k} U_{kj} = \sum_{k=1}^{r+j} (-1)^{r+k-j} \binom{k-1}{r+k-j} \binom{r}{i}
\]

\[
= (-1)^{r+i-j} \sum_{k=i}^{r+j} \binom{-i}{k-i} \binom{r}{r+j-k}.
\]
If \( r + j < i \), then \((X^T U)_{ij} = 0\) since the sum is empty. Otherwise,

\[
(X^T U)_{ij} = (-1)^{r+i-j} \binom{r-i}{r+j-i} = 0
\]
since \(0 \leq r - i < r + j - i\) for \(1 \leq i \leq r\). That is, \(X^T U = 0_{r \times (n-r)}\); then each column of \(U\) is in \(N(X^T) = C(X)^\perp\), so \((I - P_X)U = U\). Lemma D.6 gives \((I - P_X)ST = (I - P_X)(U - XY) = U\). Also,

\[
(U^T U)_{ij} = \sum_{k=1}^{n} U_{ki} U_{kj}
\]

\[
= (-1)^{i-j} \sum_{k=j}^{r+i} \left( \binom{r}{k-i} \binom{r}{k-j} \right)
\]

\[
= (-1)^{i-j} \sum_{k=j}^{r+i} \left( \binom{r}{r+i-k} \binom{r}{k-j} \right).
\]

If \( r + i < j \), then \((U^T U)_{ij} = 0 = T_{ij}\) since the sum is empty. Otherwise,

\[
(U^T U)_{ij} = (-1)^{i-j} \binom{2r}{r+i-j} = T_{ij}.
\]

That is, \(U^T U = T\). Then \(TAT = TST^T(I - P_X)(I - P_X)ST = ((I - P_X)ST)^T(I - P_X)ST = U^T U = T\), and \(T\) is invertible since \(U\) is lower triangular with full column rank \(n - r\) and \(rank(U^T U) = rank(U)\). Thus \(AT = I_{n-r}\), i.e. \(T = A^{-1}\).

We shall now prove the first fact about \(A\) in the proof of Lemma C.6: the bound on \(1^T A 1\). In fact, the result below gives a precise formula for this quantity from which the stated bound trivially follows.

**Proposition D.7.** \(1^T_{n-r} A 1_{n-r} = \left(\begin{array}{c} 2r \\ r \end{array} \right)^{-1} \left(\begin{array}{c} n+r \\ r \end{array} \right) = \left(\begin{array}{c} 2r \\ r \end{array} \right)^{-1} \left(\begin{array}{c} n+r \\ n-r-1 \end{array} \right).\)

**Proof of Proposition D.7.** Let us first complete the proof of Proposition D.7 assuming that the following claim is true. We shall subsequently give the proof of this claim.

\[
A 1_{n-r} = b
\]

(168)

where \(b\) is the \((n - r)\)-dimensional vector with \(b_i = \left(\begin{array}{c} 2r \\ r \end{array} \right)^{-1} \left(\begin{array}{c} n-i \\ r \end{array} \right) \left(\begin{array}{c} r+i-1 \\ r \end{array} \right).\)

By (168), for the claimed expression of \(1^T_{n-r} A 1_{n-r}\), it is equivalent to show that \(\left(\begin{array}{c} 2r \\ r \end{array} \right) 1^T_{n-r} b =\)
To see this, write
\[
\binom{2r}{r} T^* b = \binom{2r}{r} \sum_{i=1}^{n-r} b_i = \sum_{i=1}^{n-r} \binom{n-i}{r} \binom{r+i-1}{r} = \sum_{i=1}^{n-r} \binom{n-i}{n-r-i} \binom{r+i-1}{i-1} = (-1)^{n-r-1} \sum_{i=1}^{n-r} \binom{n-r-i}{i-1} \binom{-r-1}{i-1} = (-1)^{n-r-1} (-2r-2) = \binom{n+r}{n-r-1}
\]
which proves Proposition D.7 assuming that (168) is true. We shall now prove (168). By Proposition D.5, it is equivalent to show that \( \binom{2r}{r} T b = \binom{2r}{r} 1_{n-r} \). We have
\[
\binom{2r}{r} (Tb)_i = \binom{2r}{r} \sum_{j=1}^{n-r} T_{ij} b_j = \sum_{j=1}^{n-r} (-1)^{i-j} \binom{2r}{r+1-i-j} \binom{n-j}{r} \binom{r+j-1}{r} = \sum_{j=1}^{n-r} (-1)^{i-j} \binom{2r}{r+j-i} \binom{n-j}{n-r-j} \binom{r+j-1}{r} = \sum_{j=r+1}^{n} (-1)^{r+i-j} \binom{2r}{j-i} \binom{n+r-j}{n-j} \binom{j-1}{r} = (-1)^{n-r+i} \sum_{j=r+1}^{n} \binom{2r}{j-i} \binom{-r-1}{n-j} \binom{j-1}{r}.
\]

Since \( \binom{j-1}{r} \) is a degree \( r \) polynomial with leading coefficient \( \frac{1}{r!} \) and \( \binom{j-i}{r} \binom{j-1}{r} \) is a basis
for degree $r$ polynomials, we can write $(j^{-1})^r = \sum_{k=0}^r c_k(j - i)_k$ with $c_r = \frac{1}{r!}$. Then

$$\binom{2r}{r}(Tb)_i = (-1)^{n-r+i} \sum_{j=i}^n \binom{2r}{j-i}(n-j)^r \sum_{k=0}^r c_k(j-i)_k$$

$$= (-1)^{n-r+i} \sum_{j=i}^n c_k(2r)_k \binom{2r-k}{j-i-k}(n-j)^r$$

$$= (-1)^{n-r+i} \sum_{k=0}^r c_k(2r)_k \binom{r-k-1}{n-i-k}$$

$$= (-1)^{n-r+i} c_r(2r)_r \binom{-1}{n-r-i} = \frac{(2r)_r}{r!} \binom{n-r-i}{r} = \binom{2r}{r}.$$

The first equality follows from the fact that $\binom{2r}{j-i} = 0$ for $j < i$ and $\binom{i-1}{r} = 0$ for $j \leq r$. The second equality follows from the identity $\binom{2r}{j-i}(j-i)_k = (2r)_k(j-i)_k$. The third equality follows from the fact that $\binom{2r-k}{j-i-k} = 0$ for $j < i + k$. This completes the proof of (168).

We now turn to the second claimed fact about $A$ in the proof of Lemma C.6. This is the content of the following proposition.

**Proposition D.8.** Every entry of the matrix $A$ is positive.

We need the following lemma for the proof of Proposition D.8.

**Lemma D.9.** Let $x$ be the $(n-r)$-dimensional vector with $i^{th}$ component: $x_i = \binom{n+r-1}{n-i}^{-1} \binom{r+i-2}{r-1} \binom{n-i}{r-i}$. Then $x$ is the first column of $A$.

**Proof of Lemma D.9.** By Proposition D.5, it is equivalent to show that $Tx = e_1$, where
$e_1$ is the first standard basis vector of $\mathbb{R}^{n-r}$. We have

$$
\binom{n + r - 1}{n - 1} (Tx)_i = \binom{n + r - 1}{r} \sum_{j=1}^{n-r} T_{ij} x_j
$$

$$
= \sum_{j=1}^{n-r} (-1)^{i-j} \binom{2r}{r+i-j} \binom{r+j-2}{r-1} \binom{n-j}{n-r-j}
$$

$$
= (-1)^{n-r-i} \sum_{j=1}^{n-r} \binom{2r}{r+j-i} \binom{r+j-2}{r-1} \binom{n-r-1}{r-1}
$$

$$
= (-1)^{n-r-i} \sum_{j=r+1}^{n} \binom{2r}{j-i} \binom{j-2}{r-1} \binom{n-r}{n-j}
$$

$$
= (-1)^{n-r-i} \binom{2r}{1-i} \binom{-1}{r-1} \binom{-1}{n-1} \delta_{i,1},
$$

where $\delta_{ij}$ is the Kronecker delta. The last equality follows from the fact that $\binom{2r}{j-i} = 0$ for $j < i$ and $\binom{j-2}{r-1} = 0$ for $2 \leq j \leq r$. Now

$$
-(-1)^{n-r-i} \binom{2r}{1-i} \binom{-1}{r-1} \binom{-1}{n-1} = (-1)^{i-1} \binom{r-1}{r-1} \binom{n+r-1}{n-1} \delta_{i,1}
$$

$$
= \binom{n+r-1}{n-1} \delta_{i,1}.
$$

Writing $\binom{j-2}{r-1} = \sum_{k=0}^{r-1} c_k (j-i)_k$, similarly to the proof of Proposition D.7,

$$
\sum_{j=i}^{n} \binom{2r}{j-i} \binom{j-2}{r-1} \binom{-1}{n-j} = \sum_{j=i}^{n} \binom{2r}{j-i} \binom{-1}{n-j} \sum_{k=0}^{r-1} c_k (j-i)_k
$$

$$
= \sum_{j=i}^{n} \sum_{k=0}^{r-1} c_k (2r)_k \binom{2r-k}{j-i-k} \binom{-1}{n-j}
$$

$$
= \sum_{k=0}^{r-1} c_k (2r)_k \sum_{j=i+k}^{n} \binom{2r-k}{j-i-k} \binom{-1}{n-j}
$$

$$
= \sum_{k=0}^{r-1} c_k (2r)_k \binom{r-k-1}{n-i-k} = 0.
$$

The second and third equalities follow from the same reasoning as in the proof of Proposition D.7. The last equality follows from the fact that $0 \leq r - k - 1 < n - i - k$ for $i \leq n - r$. Thus $(Tx)_i = \delta_{i,1}$, i.e. $Tx = e_1$. \qed
We are now ready to prove Proposition D.8.

**Proof of Proposition D.8.** Let \( x \) be defined as in Lemma D.9. Observe that

\[
\frac{x_{k+1}}{x_k} = \frac{r + k - 1}{k} \cdot \frac{n - r - k}{n - k}
\]

and

\[
\frac{x_{n-r-k+1}}{x_{n-r-k}} = \frac{n - k}{n - r - k + 1} \cdot \frac{k - 1}{r + k - 1},
\]

so

\[
\frac{x_{k+1}}{x_k} \cdot \frac{x_{n-r-k+1}}{x_{n-r-k}} = \frac{k - 1}{k} \cdot \frac{n - r - k}{n - r - k + 1} < 1.
\]

Then for \( i \leq \frac{n-r+1}{2}, \)

\[
\frac{x_{n-r-i+1}}{x_i} = \prod_{k=i}^{n-r-i} \frac{x_{k+1}}{x_k} = \prod_{k=i}^{n-r-i} \frac{x_{n-r-k+1}}{x_{n-r-k}} = \sqrt[2]{\prod_{k=i}^{n-r-i} \frac{x_{k+1}}{x_k} \frac{x_{n-r-k+1}}{x_{n-r-k}} \leq 1}
\]

is increasing in \( i \) since the number of terms in the product decreases as \( i \) increases. Let \( 1 \leq i \leq j \leq n - r \) such that \( i + j \leq n - r + 1 \). If \( j \leq \frac{n-r+1}{2} \), then

\[
\frac{x_{n-r-i+1}}{x_i} \cdot \frac{x_{n-r-j+1}}{x_j} \leq 1.
\]

Otherwise, let \( j' = n - r - j + 1 \), so that \( i \leq j' \leq \frac{n-r+1}{2} \). Then

\[
\frac{x_{n-r-i+1}}{x_i} \cdot \frac{x_{n-r-j+1}}{x_j} = \frac{x_{n-r-i+1}}{x_i} \cdot \left( \frac{x_{n-r-j'+1}}{x_{j'}} \right)^{-1} \\
\leq \frac{x_{n-r-j'+1}}{x_{j'}} \cdot \left( \frac{x_{n-r-j'+1}}{x_{j'}} \right)^{-1} = 1.
\]

Thus

\[
x_i x_j - x_{n-r-i+1} x_{n-r-j+1} \geq 0.
\]

Observe that \( T \) is a symmetric Toeplitz matrix. By Lemma D.9, \( x \) is the first column of \( A \), so the symmetric Gohberg-Semencul formula (see, for example, Gohberg and Semencul [16]) gives

\[
A = \frac{1}{x_1} \begin{bmatrix}
    x_1 & 0 & \cdots & 0 \\
    x_2 & x_1 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n-r} & x_{n-r-1} & \cdots & x_1 \\
\end{bmatrix}
\begin{bmatrix}
    x_1 & x_2 & \cdots & x_{n-r} \\
    0 & x_1 & \cdots & x_{n-r-1} \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & x_1 \\
\end{bmatrix}
\begin{bmatrix}
    0 & x_{n-r} & \cdots & x_2 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & x_{n-r} \\
    0 & 0 & \cdots & 0 \\
\end{bmatrix}
\]
or

\[
A_{ij} = \frac{1}{x_1} \left[ \sum_{k=1}^{i} x_kx_{k+j-i} - \sum_{k=1}^{i-1} x_{n-r-k+1}x_{n-r+i-j-k+1} \right] = \frac{1}{x_1} \left[ x_i x_j + \sum_{k=1}^{i-1} (x_k x_{k+j-i} - x_{n-r-k+1}x_{n-r+i-j-k+1}) \right].
\]

Since \( T \) is symmetric Toeplitz, in particular it is symmetric persymmetric; by Proposition D.5, \( T = A^{-1} \), so \( A \) is symmetric persymmetric as well. It suffices then to consider \( A_{ij} \) for \( 1 \leq i \leq j \leq n-r \) satisfying \( i+j \leq n-r+1 \). Now \( 1 \leq k \leq k+j-i \leq n-r \) and \( k+(k+j-i) \leq n-r+1 \) for \( 1 \leq k \leq i-1 \), hence \( x_k x_{k+j-i} - x_{n-r-k+1}x_{n-r+i-j-k+1} \geq 0 \) and \( A_{ij} \geq \frac{x_i x_j}{x_1} > 0 \). Thus \( A_{ij} > 0 \) for all \( 1 \leq i, j \leq n-r \) which completes the proof of Proposition D.8.

\[ \square \]

D.5. A Result on Variance and Variation (Lemma B.4)

In this subsection, we provide the proof of Lemma B.4 which was used in the proof of Corollary 2.3.

**Proof of Lemma B.4.** Note that first that for \( r = 1 \), the result follows by taking \( \eta = \bar{\theta} 1_n \) (where \( \bar{\theta} := (\theta_1 + \cdots + \theta_n)/n \)) and using the inequality

\[
\sum_{i=1}^{n} (\theta_i - \bar{\theta})^2 \leq n\|D\theta\|_1^2 = nV^2(\theta), \tag{169}
\]

which is a consequence of the fact that \( |\theta_i - \bar{\theta}| \leq \max_{k,l} |\theta_k - \theta_l| \leq V(\theta) \) for every \( 1 \leq i \leq n \).

Let us therefore assume that \( r \geq 2 \). We may assume without loss of generality that the vector \( D^{(r-1)}\theta \) has mean zero (if not, we will work with \( \tilde{\theta} \) instead of \( \theta \) where \( \tilde{\theta} \) is created by subtracting a suitable polynomial sequence of degree \( r-1 \) from \( \theta \); this will ensure that \( D^{(r-1)}\tilde{\theta} \) has mean zero and that \( D^{(r)}\theta = D^{(r)}\tilde{\theta} \). Let \( X \) be the \( n \times (r-1) \) matrix whose \((i,j)^{th}\) entry equals \( (i-1) \). Let \( S \) be the \( n \times (n-r+1) \) matrix whose \((i,j)^{th}\) entry equals \( (i-j+1)/(r-1) \). Throughout we use the convention that \( \binom{a}{b} = 0 \) when \( a < b \). Let \( \eta \) denote the projection of \( \theta \) on to the column space of \( X \). We shall prove that the conditions of Lemma B.4 are satisfied for this choice of \( \eta \).

Note first that \( \eta \) belongs to the column space of \( X \) which implies that the entries \( \eta_i \) of \( \eta \) will be given by a polynomial in \( i \) of degree at most \( r-2 \) so that \( D^{(r-1)}\eta = 0_{n-r+1} \). The reader may observe that \( D^{(r-1)}\eta = 0_{n-r+1} \) is stronger than the statement of Lemma B.4 which reads \( D^{(r)}\eta = 0_{n-r} \). This is because we have assumed that \( D^{(r-1)}\theta \) has mean zero. When this condition is not true, we would need to add a polynomial sequence of degree \( r-1 \) to \( \eta \) so that then \( D^{(r-1)}\eta \) will have a constant mean which is same as saying that \( D^{(r)}\eta = 0_{n-r} \).
Note from Lemma D.2 that $SD^{(r-1)} \theta$ differs from $\theta$ by a polynomial of degree at most $r - 2$ so that

$$\theta - \eta = (I - P_X) \theta = (I - P_X) SD^{(r-1)} \theta$$

where $P_X$ is the projection matrix onto the column space of $X$. As a result

$$\|\theta - \eta\|^2 = \|(I - P_X) SD^{(r-1)} \theta\|^2 \leq \|(I - P_X) S\|^2 \|D^{(r-1)} \theta\|^2$$

where $\|(I - P_X) S\|$ denotes the operator norm of the matrix $(I - P_X) S$. It is clear that the square of the operator norm of $(I - P_X) S$ equals the operator norm of $A := S^T (I - P_X) S$ so that

$$\|\theta - \eta\|^2 \leq \|A\| \|D^{(r-1)} \theta\|^2.$$

Note now that because $A$ is symmetric, its operator norm is bounded by its $\|\cdot\|_\infty$ norm (see, for example, Golub and Loan [18, Corollary 2.3.2]) defined by

$$\|A\|_\infty := \max_{1 \leq i \leq n} \sum_{j=1}^{n-r+1} |a_{ij}|$$

and hence we have

$$\|\theta - \eta\|^2 \leq \|A\|_\infty \|D^{(r-1)} \theta\|^2.$$  \hfill (170)

It may be noted that the matrix $A$ is the same matrix that appeared in the previous section (for example, in Proposition D.8 and Proposition D.7) with $r$ replaced by $r - 1$. Therefore because all entries of $A$ are positive (Proposition D.8), we deduce that $\|A\|_\infty = \|A 1_{n-r+1}\|_\infty$ (this latter $\|\cdot\|_\infty$ norm refers to the usual $L_\infty$ norm for vectors). In the proof of Proposition D.7, we gave a precise expression for $A 1_{n-r+1}$ (see equation (168)). Using this, we deduce that (note that $r$ needs to be replaced by $r - 1$ in (168))

$$\|A\|_\infty = \max_{1 \leq i \leq n-r+1} \frac{(n-1) (r+2-i) (r+1)}{2 (r-1)} \leq \frac{n^{2(r-2)}}{2 (r-1)} \leq n^{2(r-2)}.$$

Using the above with inequality (170), we obtain

$$\|\theta - \eta\|^2 \leq n^{2(r-2)} \|D^{(r-1)} \theta\|^2.$$

To bound the right hand side above further, we use (169) (note that the mean of the vector $D^{(r-1)} \theta$ is taken to be zero) to deduce that

$$\|\theta - \eta\|^2 \leq n^{2r-1} \|D^{(r)} \theta\|_1^2$$

which completes the proof of Lemma B.4.
**D.6. Proof of the metric entropy bound for $C_r(\{a_i\}, \{s_i\})$ (Lemma C.2)**

We shall provide the proof of Lemma C.2 in this subsection. For this, we need to bound the metric entropy $\log N(\epsilon, C_r(\{a_i\}, \{s_i\}))$ of the class $C_r(\{a_i\}, \{s_i\})$ defined in (C.1). Our strategy for this involves the notion of fat shattering dimension. This is a standard concept from the theory of empirical processes (see e.g., Pollard [37], Rudelson and Vershynin [41]) and is recalled below for the convenience of the reader.

**Definition D.1 (Fat Shattering Dimension).** Let $K$ be a subset of $\mathbb{R}^n$. For $t \geq 0$, we say that a subset $\{i_1, \ldots, i_m\}$ of $\{1, \ldots, n\}$ is $t$-shattered by $K$ if there exist real numbers $h_{i_1}, \ldots, h_{i_m}$ such that for every subset $S \subseteq \{i_1, \ldots, i_m\}$, there exists a vector $\theta \in K$ for which $\theta_{i_k} \leq h_{i_k}$ if $i_k \in S$ and $\theta_{i_k} \geq h_{i_k} + t$ if $i_k \notin S$. The fat shattering dimension of $K$, denoted by $v(K, t)$, is defined as the maximum cardinality of a set $\{i_1, \ldots, i_m\} \subseteq \{1, \ldots, n\}$ that is $t$-shattered by $K$.

A deep connection between fat shattering dimension and metric entropy is given by the following result due to Rudelson and Vershynin [41, Corollary 6.4] which bounds the metric entropy using the fat shattering dimension.

**Theorem D.10 (Rudelson and Vershynin).** Let $K$ be a subset of $\mathbb{R}^n$. Assume that there exists a decreasing function $v : (0, \infty) \to (0, \infty)$ and a real number $a > 2$ such that

$$v(K, s) \leq v(s) \quad \text{and} \quad v(as) \leq \frac{1}{2}v(s) \quad \text{for all } s > 0. \quad (171)$$

Then there exists a constant $C$ depending on $a$ alone such that

$$\log N(\epsilon, K) \leq Cv \left( \frac{\epsilon}{C\sqrt{n}} \right). \quad (172)$$

In order to use Theorem D.10 to prove Lemma C.2, it is clear that we need to bound the fat shattering dimension $v(C_r(\{a_i\}, \{s_i\}), t)$ of $C_r(\{a_i\}, \{s_i\})$. The following lemma bounds the fat shattering dimension of the class $C_r(a, V)$ defined as:

$$C_r(a, V) := \{\theta \in \mathbb{R}^n : a \leq (D^{r-1}\theta)_1 \leq \cdots \leq (D^{r-1}\theta)_{n-r+1} \leq a + V\} \quad (173)$$

for $a \in \mathbb{R}$ and $V \geq 0$. Note that $C_r(\{a_i\}, \{s_i\}) \subseteq C_r(a_{r-1}, s_{r-1})$ so that the fat shattering dimension of $C_r(\{a_i\}, \{s_i\})$ is bounded from above by that of $C_r(a_{r-1}, s_{r-1})$.

**Lemma D.11.** For every $V > 0$, $a \in \mathbb{R}$, $r \geq 1$, $n \geq r$ and $t > 0$, we have

$$v(C_r(a, V), t) \leq r + \frac{V^{1/r}n^{1-(1/r)}}{t^{1/r}}C_r \quad (174)$$

for a positive constant $C_r$ that depends solely on $r$.

Let us first prove Lemma C.2 assuming that Lemma D.11 is true. The proof of Lemma D.11 will be provided following the next proof.
Proof of Lemma C.2. It turns out that it is enough to prove the following bound on the fat shattering dimension of $C_r(\{a_i\}, \{s_i\})$:

$$v(C_r(\{a_i\}, \{s_i\}), t) \leq C_r \left( \frac{\sum_{j=1}^{r} n^{j-1} s_{j-1}}{t} \right)^{1/r}.$$ (175)

Indeed, Lemma C.2 is a direct consequence of the above inequality along with Theorem D.10. To see this, note that if inequality (175) is true, one can simply take the function $v(\cdot)$ in Theorem D.10 to be

$$v(s) = C_r \left( \frac{\sum_{j=1}^{r} n^{j-1} s_{j-1}}{s} \right)^{1/r}.$$ 

Then the condition (171) in Theorem D.10 is true with $a = 2^r$ and Lemma C.2 is therefore a consequence of inequality (172).

The key therefore is to prove (175). For this, note first the identity (which is a consequence of Lemma D.2 applied with $r-1$ instead of $r$)

$$\theta_i = \sum_{j=1}^{i-r+1} \binom{i-j-1}{r-2} (D^{(r-1)}\theta)_j + \sum_{j=1}^{r-1} \binom{i-1}{j-1} (D^{(j-1)}\theta)_1.$$ 

This identity obviously implies the following lower and upper bounds on $\theta_i$ for every $\theta \in C_r(\{a_i\}, \{s_i\})$:

$$\theta_i \geq \sum_{j=1}^{i-r+1} \binom{i-j-1}{r-2} a_{r-1} + \sum_{j=1}^{r-1} \binom{i-1}{j-1} a_{j-1}$$ 

and

$$\theta_i \leq \sum_{j=1}^{i-r+1} \binom{i-j-1}{r-2} a_{r-1} + \sum_{j=1}^{r-1} \binom{i-1}{j-1} a_{j-1} + \sum_{j=1}^{r-1} \binom{i-1}{j-1} s_{j-1} + \sum_{j=1}^{i-r+1} \binom{i-j-1}{r-2} s_{r-1}.$$ 

The last two terms in the expression above can be combined into one term as follows:

$$\sum_{j=1}^{r-1} \binom{i-1}{j-1} s_{j-1} + \sum_{j=1}^{i-r+1} \binom{i-j-1}{r-2} = \sum_{j=1}^{r} \binom{i-1}{j-1} s_{j-1}.$$ 

This is a consequence of the fact that

$$\sum_{j=1}^{i-r+1} \binom{i-j-1}{r-2} = \binom{i-1}{r-1}.$$
which itself follows from (146) applied to \( a = i - 2 \) and \( b = r - 2 \). We thus have

\[
\theta_i \leq \sum_{j=1}^{i-r+1} \left( \frac{i-j}{r-2} \right) a_{r-1} + \sum_{j=1}^{r-1} \left( \frac{i-1}{j-1} \right) a_{j-1} + \sum_{j=1}^{r} \left( \frac{i-1}{j-1} \right) s_{j-1}.
\]

Combining the upper and lower bounds for \( \theta_i \) derived above, we deduce that

\[
\max_{\theta \in \mathcal{C}_r(\{a_i\}, \{s_i\})} \theta_i - \min_{\theta \in \mathcal{C}_r(\{a_i\}, \{s_i\})} \theta_i \leq \sum_{j=1}^{r} \left( \frac{n-1}{j-1} \right) s_{j-1} \leq \sum_{j=1}^{r} n^{j-1} s_{j-1}.
\]

The presence of \( r - 2 \) in the binomial coefficients above might seem to make the above statement true only for \( r \geq 2 \). However for \( r = 1 \), this directly follows from the fact that every vector \( \theta \) in \( \mathcal{C}_1(\{a_i\}, \{s_i\}) \) satisfies \( a_0 \leq \theta_1 \leq \cdots \leq \theta_n \leq a_0 + s_0 \).

As a consequence, it turns out that \( v(\mathcal{C}_r(\{a_i\}, \{s_i\}), t) = 0 \) if \( t > \Gamma := \sum_{j=1}^{r} n^{j-1} s_{j-1} \) and hence inequality (175) is trivially true when \( t > \Gamma \). We can therefore assume that \( t \leq \Gamma \). In this case, because \( \mathcal{C}_r(\{a_i\}, \{s_i\}) \subseteq \mathcal{C}_r(a_{r-1}, s_{r-1}) \), Lemma D.11 gives

\[
v(\mathcal{C}_r(\{a_i\}, \{s_i\}), t) \leq v(\mathcal{C}_r(a_{r-1}, s_{r-1}), t)
\leq r + C_r \left( \frac{n^{r-1} s_{r-1}}{t} \right)^{1/r}
\leq r \left( \frac{\Gamma}{t} \right)^{1/r} + C_r \left( \frac{\Gamma}{t} \right)^{1/r} = (C_r + r) \left( \frac{\Gamma}{t} \right)^{1/r},
\]

which proves (175) when \( t \leq \Gamma \). The completes the proof of Lemma D.6. \( \square \)

We now prove Lemma D.11. For this, we use the notion of divided differences (see, for example, Kuczma [26, Chapter 15]). For \( k \geq 1 \), indices \( 1 \leq \ell_1 < \cdots < \ell_k \leq n \) and real numbers \( \alpha_{\ell_1}, \ldots, \alpha_{\ell_k} \), the divided difference \([\ell_1, \ldots, \ell_k; \alpha] \) is defined as

\[
[\ell_1, \ldots, \ell_k; \alpha] := \sum_{i=1}^{k} \frac{\alpha_{\ell_i}}{\prod_{j \neq i} (\ell_i - \ell_j)}
\]

As examples, note that \([\ell_1; \alpha] = \alpha_{\ell_1} \) and \([\ell_1, \ell_2; \alpha] = (\alpha_{\ell_2} - \alpha_{\ell_1})/(\ell_2 - \ell_1) \).

It is easy to verify that the divided differences satisfy the recursive relation

\[
[\ell_1, \ldots, \ell_k; \alpha] = \frac{[\ell_2, \ldots, \ell_k; \alpha] - [\ell_1, \ldots, \ell_{k-1}; \alpha]}{\ell_k - \ell_1}.
\]

We shall use the following two facts about divided differences for the proof of Lemma D.11.

The first fact is given in Lemma D.12 below which is a simple consequence of Kuczma [26, Theorem 15.3.1].
Lemma D.12. Fix \( r \geq 1 \) and \( n \geq r \). Suppose \( \theta \in \mathbb{R}^n \) satisfies \((D^{(r-1)}\theta)_{1} \leq \cdots \leq (D^{(r-1)}\theta)_{n-r+1}\). Then for every choice of indices \( 1 \leq i_1 < \cdots < i_{r+1} \leq n \), we have
\[
[i_2, \ldots, i_{r+1}; \theta] \geq [i_1, \ldots, i_r; \theta].
\]

Remark D.1. When \( r = 2 \), it is easy to see that Lemma 2.2 reduces to the well-known increasing slopes property of convex sequences.

The second fact about divided differences is given in Lemma D.13 below which is a consequence of Kuczma [26, Lemma 15.2.5 and Theorem 15.2.6].

Lemma D.13. Fix \( r \geq 1 \) and \( n \geq r \). For every choice of indices \( 1 \leq i_1 < i_2 < \cdots < i_r \leq n \), there exist non-negative real numbers \( \{c_i, 1 \leq i \leq n-r+1\} \) with \( \sum_{i=1}^{n-r+1} c_i = 1 \) such that
\[
[i_1, \ldots, i_r; \theta] = \frac{1}{(r-1)!} \sum_{i=1}^{n-r+1} c_i (D^{(r-1)}\theta)_i \quad \text{for every } \theta \in \mathbb{R}^n.
\]

Proof of Lemma D.11. Fix \( t > 0 \) and suppose that \( S := \{i_1, \ldots, i_m\} \) (with \( 1 \leq i_1 < \cdots < i_m \leq n \)) is a subset of \( \{1, \ldots, n\} \) that is \( t \)-shattered by \( \mathcal{C}(V) \). Let \( h_{i_1}, \ldots, h_{i_m} \) denote the associated levels and denote by \( h \) the vector in \( \mathbb{R}^m \) given by \((h_{i_1}, \ldots, h_{i_m})\). We shall then prove that \( m \) is bounded from above by the right hand side of (174). Note that we can assume that \( m \geq r \) (otherwise there is nothing to prove).

We first claim that
\[
[i_j, i_{j+1}, \ldots, i_{j+r-1}; h] \geq [i_{j-1}, \ldots, i_{j+r-2}; h] + t \sum_{k=j-1}^{j+r-1} (-\tau_{k,j})\{\tau_{k,j} < 0\} \quad (176)
\]
for every \( j = 2, \ldots, m - r + 1 \) where
\[
\tau_{k,j} := \prod_{j \leq \ell \leq j+r-1; \ell \neq k} \frac{1}{i_k - i_\ell} - \prod_{j-1 \leq \ell \leq j+r-2; \ell \neq k} \frac{1}{i_k - i_\ell}
\]
for \( k = j, \ldots, j+r-2 \) and
\[
\tau_{j-1,j} := (-1)^{r-1} \prod_{\ell = j}^{j+r-2} \frac{1}{i_\ell - i_{j-1}} \quad \text{and} \quad \tau_{j+r-1,j} := \prod_{\ell = j}^{j+r-2} \frac{1}{i_{j+r-1} - i_\ell}
\]
In the above, for \( r = 1 \), we take \( \tau_{j-1,j} = -1 \) and \( \tau_{j,j} = 1 \).

To see (176), note first that because \( S \) is \( t \)-shattered by \( \mathcal{C}_r(a, V) \), there exists \( \theta \in \mathcal{C}_r(a, V) \) such that \( \theta_{i_k} \leq h_{i_k} \) whenever \( \tau_{k,j} \geq 0 \) and \( \theta_{i_k} \geq h_{i_k} + t \) whenever \( \tau_{k,j} < 0 \). Because \( \theta \in \mathcal{C}_r(a, V) \), Lemma D.12 gives
\[
[i_j, i_{j+1}, \ldots, i_{j+r-1}; \theta] \geq [i_{j-1}, \ldots, i_{j+r-2}; \theta].
\]
It can be checked that the above inequality is equivalent to
\[ \sum_{k=j-1}^{j+r-1} \tau_{k,j} \theta_{ik} \geq 0 \]
which is further equivalent to
\[ \sum_{k=j-1}^{j+r-1} \tau_{k,j} \theta_{ik} \geq \sum_{k=j-1}^{j+r-1} (\tau_{k,j}) \theta_{ik} \{ \tau_{k,j} < 0 \}. \]

The above inequality, together with the fact that \( \theta_{ik} \leq h_{ik} \) when \( \tau_{k,j} \geq 0 \) and \( \theta_{ik} \geq h_{ik} + t \) when \( \tau_{k,j} < 0 \), gives (176).

From (176), it is easy that by recursive application, one obtains
\[ [i_j, i_{j+1}, \ldots, i_{j+r-1}; h] \geq [i_u, i_{u+1}, \ldots, i_{u+r-1}; h] + t \sum_{a=u}^{j-1} \sum_{k=a}^{j+r-1} (\tau_{k,a+1}) \{ \tau_{k,a+1} < 0 \} \]
for every \( 1 \leq u < j \leq m - r + 1 \). Taking \( u = 1 \) and \( j = m - r + 1 \), we obtain
\[ [i_{m-r+1}, \ldots, i_m; h] - [i_1, \ldots, i_r; h] \geq t T_r. \]

where
\[ T_r := \sum_{a=1}^{m-r} \sum_{k=a}^{m} (\tau_{k,a+1}) \{ \tau_{k,a+1} < 0 \}. \]

We now claim that
\[ [i_1, \ldots, i_r; h] \geq \frac{a}{(r-1)!} \quad \text{and} \quad [i_{m-r+1}, \ldots, i_m; h] \leq \frac{a + V}{(r-1)!}. \]

We shall prove the first inequality in (178) below. The proof of the second inequality will be similar. One can write \([i_1, \ldots, i_r; h]\) as \( \sum_{j=1}^{r} \beta_j h_{ij} \) for some real coefficients \( \beta_j \). Because \( S \) is \( t \)-shattered by \( \mathcal{C}_r(a, V) \), there exists \( \theta \in \mathcal{C}_r(a, V) \) such that \( h_{ij} \geq \theta_{ij} \) for \( \beta_j \geq 0 \) and \( h_{ij} < \theta_{ij} \) for \( \beta_j < 0 \). This implies that
\[ [i_1, \ldots, i_r; h] = \sum_{j=1}^{r} \beta_j h_{ij} \geq \sum_{j=1}^{r} \beta_j \theta_{ij} = [i_1, \ldots, i_r; \theta]. \]

Lemma D.13 now implies that, for some \( c_i \geq 0, 1 \leq i \leq n - r + 1 \) with \( \sum_{i=1}^{n-r+1} c_i = 1 \), we have
\[ [i_1, \ldots, i_r; \theta] = \frac{1}{(r-1)!} \sum_{i=1}^{n-r+1} c_i (D^{(r-1)} \theta)_i \geq \frac{a}{(r-1)!} \]
where the last inequality follows because \( \theta \in \mathcal{C}_r(a, V) \). This proves (178).

Combining (178) and (177), we obtain
\[ T_r \leq \frac{V t}{t(r-1)!}. \]
We now claim the following lower bound for $T_r$:

$$T_1 = m - 1 \quad \text{and} \quad T_r \geq \frac{(m - r)^r}{n^{r-1}(r-1)^{r-1}} \quad \text{for every } r \geq 2. \quad (179)$$

Before we prove (179), note first that as a consequence of the above pair of inequalities, inequality (174) holds with $C_1 = 1$ and

$$C_r = \left( \frac{(r-1)^{r-1}}{(r-1)!} \right)^{1/r} \quad \text{for } r \geq 2.$$

Therefore, to complete the proof of Lemma D.11, we only need to prove inequality (179).

To prove (179), we assume that $r \geq 2$ (the fact that $T_1 = m - 1$ is obvious) and note first that $\tau_{a+r-1,a+1} < 0$ for every $a = 1, \ldots, m - r$. As a result,

$$T_r \geq \sum_{a=1}^{m-r} (-\tau_{a+r-1,a+1}) \geq \sum_{a=1}^{m-r} \frac{1}{(i_{a+r-1} - i_a) \cdots (i_{a+1} - i_{a+r-2})}.$$

By the AM-GM inequality, we have

$$(i_{a+r-1} - i_a) \cdots (i_{a+1} - i_{a+r-2}) \leq \left( \frac{(i_{a+r-1} - i_a) + \cdots + (i_{a+1} - i_{a+r-2})}{r-1} \right)^{r-1}.$$

If we define $s_j := i_{j+1} - i_j$ for $j = 1, \ldots, m - 1$, then it is easy to see that

$$\frac{(i_{a+r-1} - i_a) + \cdots + (i_{a+1} - i_{a+r-2})}{r-1} = \sum_{j=0}^{r-2} \frac{j+1}{r-1} s_{a+j} \leq \sum_{j=0}^{r-2} s_{a+j}.$$

We have deduced therefore that

$$T_r \geq \sum_{a=1}^{m-r} \left( \frac{1}{\sum_{j=0}^{r-2} s_{a+j}} \right)^{r-1}.$$

We now use the convexity of the map $x \mapsto (1/x)^{r-1}$ for $x > 0$ to obtain

$$T_r \geq \frac{(m - r)^r}{\left( \sum_{a=1}^{m-r} \sum_{j=0}^{r-2} s_{a+j} \right)^{r-1}}.$$

Inequality (179) follows from here because

$$\sum_{a=1}^{m-r} \sum_{j=0}^{r-2} s_{a+j} = \sum_{j=0}^{r-2} \sum_{a=1}^{m-r} s_{a+j} = \sum_{j=0}^{r-2} (i_{m-r+j+1} - i_{j+1}) \leq n(r - 1).$$

This completes the proof of Lemma D.11.
Appendix E: Additional Simulation Results

The purpose of this section is to provide additional details for the main simulation section as well as to provide results for the function $f^*_3(x) := 14.77I\{0.1 < x \leq 0.13\} - 3.69I\{0.13 < x \leq 0.15\} + 7.39I\{0.15 < x \leq 0.23\} - 7.39I\{0.23 < x \leq 0.25\} + 11.08I\{0.25 < x \leq 0.4\} - 4.43I\{0.4 < x \leq 0.44\} + 3.32I\{0.44 < x \leq 0.65\} + 19.21I\{0.65 < x \leq 0.76\} + 7.76I\{0.76 < x \leq 0.78\} + 15.51I\{0.78 < x \leq 0.81\}$. This function (plotted in Figure 7) is similar to the blocks function of Donoho and Johnstone [9].

Note that in our simulation results for $f^*_1$, we computed the ideal penalized estimator with $\lambda$ taken to be $\lambda^*$ defined as in (27). We mentioned that $\lambda^*$ was computed by Monte-Carlo averaging based on a convex optimization scheme for computing $\lambda_{\theta^*}(z)$ for each $z \in \mathbb{R}^n$. Let us provide more details behind this convex optimization here. For general $r \geq 1$, it is easy to see (using the definition of $\lambda_{\theta^*}(z)$ and the subdifferential characterization in Proposition 2.5) that $\lambda_{\theta^*}(z)$ can be read-off as the optimizing value for $\lambda$ in the following
convex optimization problem:

\[
\begin{align*}
\text{minimize} \quad & \|z - v\| \\
\text{subject to} \quad & \sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r - 1} \right) v_i = 0 \quad \text{for} \quad j = 1, \ldots, r \\
& \sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r - 1} \right) v_i - \lambda \leq 0 \quad \text{for} \quad r < j \leq n \\
& \sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r - 1} \right) v_i + \lambda \geq 0 \quad \text{for} \quad r < j \leq n \\
& \sum_{i=j}^{n} \left( \frac{r + i - j - 1}{r - 1} \right) v_i - \lambda \text{sgn}((D^{(r)} \theta)_{j-r}) = 0 \quad \text{for} \quad r < j \leq n \\
& \text{with} \quad (D^{(r)} \theta)_{j-r} \neq 0.
\end{align*}
\]

This optimization problem can be solved efficiently by the convex optimization software MOSEK for \( r = 1 \). In fact, for computational reasons, it is easier to solve the dual of this problem. For \( r \geq 2 \) however, this problem becomes quite ill-conditioned and MOSEK seems to have trouble finding the global minimizer. This is why we could not compute the \( \lambda^* \) values for the function \( f_3^* \).

Fig 8: Plots when the true function is \( f_3^* \).

The simulation results for the function \( f_3^* \) (here \( r = 1 \) as \( f_3^* \) is a piecewise constant function) are given in Figure 8. It is clear from here that the behavior of the non-CV
estimators is in accordance with our theoretical results. The CV estimators seem to behave in a complicated manner in the bottom-left plot. Again, understanding the risk behavior of CV estimates in this setting is beyond the scope of the present paper.

References

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. Information and Inference iau005.
[2] ARNOLD, T. B. and TIBSHIRANI, R. J. (2016). Efficient implementations of the generalized lasso dual path algorithm. Journal of Computational and Graphical Statistics 25 1–27.
[3] BELLEC, P. C. (2018). Sharp oracle inequalities for Least Squares estimators in shape restricted regression. Ann. Statist. 46 745–780. MR3782383
[4] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). Classification and regression trees. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. MR726392
[5] BROCKMANN, M., GASSER, T. and HERRMANN, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. J. Amer. Statist. Assoc. 88 1302–1309. MR1245363
[6] BÜHLMANN, P. and VAN DE GEER, S. (2011). Statistics for high-dimensional data. Springer Series in Statistics. Springer, Heidelberg Methods, theory and applications. MR2807761
[7] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. The Annals of Statistics 42 2340–2381.
[8] DALALYAN, A., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. Bernoulli 23 552–581.
[9] DONOHOO, D. L. and JOHNSTONE, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 425–455.
[10] DONOHOO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. Annals of Statistics 26 879–921.
[11] DUMBGEN, L. (2010). Bounding standard gaussian tail probabilities. arXiv preprint. arXiv preprint arXiv:1012.2063.
[12] FAN, Z. and GUAN, L. (2017). l_0-estimation of piecewise-constant signals on graphs. arXiv preprint arXiv:1703.01421.
[13] FOYGEL, R. and MACKYEY, L. (2014). Corrupted sensing: Novel guarantees for separating structured signals. IEEE Transactions on Information Theory 60 1223–1247.
[14] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. Ann. Statist. 19 1–141. With discussion and a rejoinder by the author. MR1091842
[15] GAO, C., HAN, F. and ZHANG, C.-H. (2017). Minimax risk bounds for piecewise constant models. arXiv preprint arXiv:1705.06386.
[16] GOHBERG, I. and SEMENCUL, A. (1972). On the inversion of finite Toeplitz matrices and their continuous analogs. Mat. issled 2 201–233.
[17] Goldenshluger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* 6 135–170. MR1466625

[18] Golub, G. H. and Loan, C. F. V. (2013). *Matrix Computations*, Fourth ed. JHU Press.

[19] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics* 38. Cambridge University Press.

[20] Guntuboyina, A. and Sen, B. (2017). Nonparametric Shape-restricted Regression. *arXiv preprint arXiv:1709.05707*.

[21] Harchaoui, Z. and Lévy-Leduc, C. (2012). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*.

[22] Hiriart-Urruty, J.-B. and Lemaréchal, C. (2013). *Convex analysis and minimization algorithms I: Fundamentals* 305. Springer science & business media.

[23] Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes Technical Report. available at arXiv preprint arXiv:1107.3806.

[24] Johnstone, I. M. (2015). *Gaussian estimation: Sequence and wavelet models*. Available at [http://statweb.stanford.edu/~imj/GE09-08-15.pdf](http://statweb.stanford.edu/~imj/GE09-08-15.pdf).

[25] Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D. (2009). l_1 trend filtering. *SIAM Rev.* 51 339–360. MR2505584

[26] Kuczma, M. (2009). *An introduction to the theory of functional equations and inequalities: Cauchy’s equation and Jensen’s inequality*. Springer Science & Business Media.

[27] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

[28] LePSKI, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* 25 929–947. MR1447734

[29] Lévy-Leduc, C. and Harchaoui, Z. (2008). Catching change-points with lasso. In *Advances in Neural Information Processing Systems* 617–624.

[30] Lin, K., Sharpnack, J., Rinaldo, A. and Tibshirani, R. J. (2016). Approximate Recovery in Changepoint Problems, from _ℓ_2 Estimation Error Rates. *arXiv preprint arXiv:1606.06746*.

[31] Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *The Annals of Statistics* 25 387–413.

[32] Mangasarian, O. L. and Schumaker, L. L. (1971). Discrete splines via mathematical programming. *SIAM Journal on Control* 9 174–183.

[33] Müller, H.-G. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* 15 182–201. MR885731

[34] Ortelli, F. and van de Geer, S. (2018). On the total variation regularized estimator over the branched path graph. *arXiv preprint arXiv:1806.01009*.

[35] Oymak, S. and Hassibi, B. (2016). Sharp mse bounds for proximal denoising. *Foundations of Computational Mathematics* 16 965–1029.

[36] Pintore, A., Speckman, P. and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* 93 113–125. MR2277744
[37] Pollard, D. (1990). Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics 2. Institute of Mathematical Statistics, Hayward, CA.

[38] Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. IEEE transactions on information theory 57 6976–6994.

[39] Rockafellar, R. T. (1970). Convex Analysis. Princeton Univ. Press, Princeton, New Jersey.

[40] Rockafellar, R. T. and Wets, R. J.-B. (2009). Variational analysis 317. Springer Science & Business Media.

[41] Rudelson, M. and Vershynin, R. (2006). Combinatorics of random processes and sections of convex bodies. Annals of Mathematics 603–648.

[42] Rudin, L. I., Osher, S. and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60 259–268.

[43] Schrijver, A. (1986). Theory of linear and integer programming. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Ltd., Chichester A Wiley-Interscience Publication. MR874114 (88m:90090)

[44] Steidl, G., Didas, S. and Neumann, J. (2006). Splines in higher order TV regularization. International journal of computer vision 70 241–255.

[45] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 267–288.

[46] Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. The Annals of Statistics 42 285–323.

[47] Van de Geer, S. (2000). Applications of Empirical Process Theory. Cambridge University Press.

[48] Van de Geer, S. (2018). On tight bounds for the Lasso. arXiv preprint arXiv:1804.00989.

[49] Van de Geer, S. and Wainwright, M. (2015). On concentration for (regularized) empirical risk minimization. arXiv preprint arXiv:1512.00677.

[50] Van der Vaart, A. and Wellner, J. A. (1996). Weak Convergence and Empirical Process: With Applications to Statistics. Springer-Verlag.

[51] Wang, Y.-X., Smola, A. J. and Tibshirani, R. J. (2014). The Falling Factorial Basis and Its Statistical Applications. In ICML 730–738.

[52] Wang, Y.-X., Sharpnack, J., Smola, A. J. and Tibshirani, R. J. (2016). Trend filtering on graphs. The Journal of Machine Learning Research 17 1–41.

[53] Winkler, G. and Liebscher, V. (2002). Smoothers for discontinuous signals. Journal of Nonparametric Statistics 14 203–222.

[54] Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. J. Amer. Statist. Assoc. 96 247–259. MR1952735