Bernoulli **27**(1), 2021, 66–92
https://doi.org/10.3150/20-BEJ1223

Adaptive confidence sets in shape restricted regression

PIERRE C. BELLEC

Department of Statistics, Hill Center, Busch Campus, Rutgers University, Piscataway, NJ 08854, USA.
E-mail: pierre.bellec@rutgers.edu

A simple construction of adaptive confidence sets is proposed in isotonic, convex and unimodal regression. In univariate isotonic regression, the proposed confidence set enjoys uniform coverage over all non-decreasing regression functions. Furthermore, the diameter of the proposed confidence set automatically adapts to the unknown number of pieces of the true parameter, in the sense that the diameter is bounded from above by the minimax risk over the class of k-piecewise constant functions. The diameter of the confidence set is a simple increasing function of the number of jumps of the isotonic least-squares estimate.

A similar construction is proposed in convex regression where the true regression function is convex and piecewise affine. Here, the confidence set enjoys uniform coverage and its diameter automatically adapts to the number of affine pieces of the true regression function. The diameter of the confidence set is an increasing function of the number of affine pieces of the convex least-squares estimate.

We explain how to extend this technique to a non-convex set by proposing a similar adaptive confidence set in unimodal regression. The confidence set automatically adapts to the number of jumps of the true unimodal regression function and its diameter is an increasing function of the number of jumps of the unimodal least-squares estimate.

Keywords: adaptive confidence set; convex regresion; isotonic regression; piecewise affine; piecewise constant; shape constraints; unimodal regression

1. Introduction

Let $K \subset \mathbb{R}^n$ be a closed convex set. Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \quad i = 1, \dots, n,$$

where the vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T \in K$ is unknown, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$ is a noise vector with n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ where $\sigma > 0$ and $I_{n\times n}$ is the $n \times n$ identity matrix. Denote by $\mathbb{E}_{\boldsymbol{\mu}}$ and $\mathbb{P}_{\boldsymbol{\mu}}$ the expectation and the probability measure corresponding to the distribution of the random variable

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}\big(\mathbf{0}, \sigma^2 I_{n\times n}\big). \tag{1.1}$$

The vector $\mathbf{y} = (Y_1, \dots, Y_n)^T$ is observed and the goal is to estimate $\boldsymbol{\mu}$. Consider the scaled norm $\|\cdot\|_n$ defined by

$$\|\boldsymbol{u}\|_n^2 = \frac{1}{n}\sum_{i=1}^n u_i^2, \quad \boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n.$$

The error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ is given by $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_n^2$. Let $|\cdot|_2^2$ be the squared Euclidean norm, so that $\frac{1}{n}|\cdot|_2^2 = \|\cdot\|_n^2$. For a finite set E, let $|E|$ denote its cardinality. We use bold face for vectors and the components of any vector $\boldsymbol{v} \in \mathbb{R}^n$ are denoted by $v_1, \dots, v_n$. If $\boldsymbol{v} \in \mathbb{R}^n$ and $T \subset \{1, \dots, n\}$, we

may view $\boldsymbol{v}$ as a function from $\{1,\dots,n\}$ to $\mathbb{R}$ and denote by $\boldsymbol{v}_T = (v_i, i \in T) \in \mathbb{R}^T$ the restriction of this function to T. If it is clear from context, we identify the linear space $\mathbb{R}^T$ with $\mathbb{R}^{|T|}$ with the canonical isomorphism that maps $(v_i, i \in T)$ to $(v_{k_1},\dots,v_{k_{|T|}}) \in \mathbb{R}^{|T|}$ where $T = \{k_1,\dots,k_{|T|}\}$ with $k_1 < k_2 < \cdots < k_{|T|}$.

In this paper, we consider the particular case where K is a polyhedron, that is, an intersection of a finite number of half-spaces. If the true parameter $\boldsymbol{\mu}$ lies in a low-dimensional face of the polyhedron K, it has been shown that for some polyhedra K, the rate of estimation is of order $\frac{d\sigma^2}{n}$ up to logarithmic factors, where d is the dimension of the smallest face that contains $\boldsymbol{\mu}$ [3,11,12,18,19,21]; see also the survey [20]. This phenomenon appears, for example, if the polyhedron K is the cone of nondecreasing sequences [3,11] or the cone of convex sequences [3,19]. For these examples, if $\boldsymbol{\mu}$ lies in a d-dimensional face of the polyhedron K, the Least Squares estimator over K satisfies risk bounds and oracle inequalities with the parametric rate $\frac{d\sigma^2}{n}$, up to logarithmic factors. We consider the problem of confidence sets in this context. In particular, the present paper addresses the following questions.

- Is it possible to estimate or bound from below by a data-driven quantity the dimension d of the smallest face of the polyhedron K that contains the true parameter $\boldsymbol{\mu}$?
- Is it possible to construct a confidence set $\hat{C}_n$ such that:
  1. It enjoys uniform coverage over all $\boldsymbol{\mu} \in K$ (i.e., $\boldsymbol{\mu} \in \hat{C}_n$ with high probability).
  2. It adapts to the smallest low-dimensional face that contains $\boldsymbol{\mu}$ (i.e., the diameter of $\hat{C}_n$ should be of the order $\frac{d\sigma^2}{n}$ up to logarithmic factors if the smallest face that contains $\boldsymbol{\mu}$ has dimension d).

In this paper, we answer these questions for two particular convex polyhedra, the cone of nondecreasing sequences and the cone of convex sequences, as well as for the non-convex set of unimodal sequences.

The construction of adaptive confidence sets in isotonic or convex regression has been studied in [6,7,15]. These papers show that if the true regression function is simultaneously smooth and monotone, then it is possible to construct confidence sets that adapt to the unknown smoothness of the true regression function. In the present paper, there is no smoothness assumption and the goal is to construct confidence sets that adapt to the dimension d of the smallest face of the polyhedron. A related literature studies the optimal separation rates for testing and construction of confidence sets in sparse mean models and sparse linear regression [2,8,9,26,30]. This literature is reviewed in Section 4.2 where we discuss piecewise constant signals.

The rest of the paper is organized as follows. Section 2 gives the definition of honest and adaptive confidence sets. Section 3 defines the cone of nondecreasing sequences and recalls some material from [1,24] on the statistical dimension and the intrinsic volumes of closed convex cones. In Section 4, honest and adaptive confidence sets are provided for the cone of nondecreasing sequences and Section 4.2 shows that, on the other hand, such honest and adaptive confidence sets do not exist for piecewise constant sequences. Section 4.3 discusses the case of unknown noise level. Honest and adaptive confidence sets are then given in Section 5 for the cone of convex sequences. In Section 6, we show that a similar construction is possible for the non-convex set of unimodal sequences, which shows that our construction of confidence sets does not require convexity of the model.

2. Honest and adaptive confidence sets

Let $(E_k)_{k\in J}$ be a collection of subsets K indexed by some possibly infinite set J. We will refer to the sets $(E_k)_{k\in J}$ as the *models*. Most collection of models considered in the paper will be finite and ordered by inclusion, that is,

$$E_1 \subset \cdots \subset E_{k_{\text{MAX}}} = K \tag{2.1}$$

where $J = \{1, \dots, k_{\mathrm{MAX}}\}$. For any model $E_k \subset K$, the minimax risk on E_k is the quantity

$$R^*_{\mathbb{E}}(E_k) = \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in E_k} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_n^2,$$

where the infimum is taken over all estimators, that is, all random variables of the form $\hat{\boldsymbol{\mu}} = g(\mathbf{y})$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ is a Borel function. If $J = \{1, \dots, k_{\mathrm{MAX}}\}$ and (2.1) holds, the minimax risks satisfy

$$R^*_{\mathbb{E}}(E_1) \le \cdots \le R^*_{\mathbb{E}}(E_{k_{\mathrm{MAX}}}).$$

In that case, the collection $(E_k)_{k=1,\dots,k_{\mathrm{MAX}}}$ represents models of increasing complexity.

Similarly, if a confidence value $\alpha \in (0, 1)$ is given, one may define the minimax quantity

$$R^*_{\alpha}(E_k) = \inf\Big\{ R > 0 : \sup_{\hat{\boldsymbol{\mu}}} \inf_{\boldsymbol{\mu} \in E_k} \mathbb{P}_{\boldsymbol{\mu}}\big(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_n^2 \le R\big) \ge 1 - \alpha \Big\} \tag{2.2}$$

for all $k \in J$, where the supremum is taken over all estimators. This quantity represents the smallest radius, in a minimax sense, of a confidence ball with confidence level $1 - \alpha$. Similarly, if $J = \{1, \dots, k_{\max}\}$ and the models are ordered by inclusion as in (2.1), this quantity is an increasing function of k and we have

$$R^*_{\alpha}(E_1) \le \cdots \le R^*_{\alpha}(E_{k_{\mathrm{MAX}}})$$

for all $\alpha \in (0, 1)$.

The goal of this paper is to study confidence sets in shape restricted regression. A confidence set is a region $\hat{C}_n$ such that with high probability, the unknown parameter $\boldsymbol{\mu}$ belongs to $\hat{C}_n$. Let $\alpha \in (0, 1)$. If $\boldsymbol{\mu} \in E_{k^*}$ for some $k^* \in J$, the quantity (2.2) may be used to define the oracle region

$$\hat{C}^*_n\big(k^*\big) := \big\{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u} - \hat{\boldsymbol{\mu}}\|_n^2 \le R^*_{\alpha}(E_{k^*})\big\},$$

where $\hat{\boldsymbol{\mu}}$ is an estimator that achieves the supremum in (2.2) (we assume here that all infima and suprema in (2.2) are attained). Then, by definition of $R^*_{\alpha}(\cdot)$, we have that $\boldsymbol{\mu} \in \hat{C}^*_n(k^*)$ with probability at least $1 - \alpha$. We call $\hat{C}^*_n(k^*)$ an *oracle* region since it is inaccessible for two reasons. First, the radius $R^*_{\alpha}(E_{k^*})$ and the integer k^* must be known in order to construct $\hat{C}^*_n(k^*)$, that is, the knowledge of the smallest model that contains $\boldsymbol{\mu}$ is needed. Second, the oracle region $\hat{C}^*_n(k^*)$ is an Euclidean ball centered around the estimator $\hat{\boldsymbol{\mu}}$ that achieves the infimum in (2.2), and this estimator is unknown.

This paper studies the construction of data-driven confidence sets $\hat{C}_n$. We consider only $1 - \alpha$ confidence sets, which means that the true parameter $\boldsymbol{\mu}$ belongs to $\hat{C}_n$ with probability at least $1 - \alpha$, uniformly over all $\boldsymbol{\mu} \in K$.

We also want the diameter of the confidence set $\hat{C}_n$ to be of the same order as the diameter of the oracle region $\hat{C}^*_n(k^*)$, that is, the value $R^*_{\alpha}(E_{k^*})$. Furthermore, the construction of $\hat{C}_n$ should not require the knowledge of the smallest model that contains the true parameter $\boldsymbol{\mu}$: The knowledge of k^* is not needed to construct the confidence region $\hat{C}_n$. In that case, we say that the confidence set $\hat{C}_n$ is adaptive.

We now give a formal definition of these properties. For any $A \subset \mathbb{R}^n$, define the diameter of A for the scaled norm $\|\cdot\|_n$ by

$$\operatorname{diam} A := \sup_{\boldsymbol{v}, \boldsymbol{u} \in A} \|\boldsymbol{v} - \boldsymbol{u}\|_n.$$

In the following definition, for any $\boldsymbol{\mu}$ we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P}_{\boldsymbol{\mu}})$ rich enough so that $\mathbf{y}$ in (1.1) is a random variable. (For instance, one may take $\Omega = \mathbb{R}^n$ and $\mathcal{A}$ the Lebesgue measure on $\mathbb{R}^n$.)

Definition 2.1. Let $\alpha \in (0,1)$. Let $K \subset \mathbb{R}^n$ be a closed convex set and let $(E_k)_{k\in J}$ be a collection of subsets of K indexed by an arbitrary set J. Denote by $\hat{C}_n : \mathbb{R}^n \to \{B \subset \mathbb{R}^n\}$ a function such that (i) the event $\{\boldsymbol{\mu} \in \hat{C}_n(\mathbf{y})\}$ is $\mathcal{A}$-measurable, and (ii) $\operatorname{diam}(\hat{C}_n(\mathbf{y}))$ is a random variable in $(\Omega, \mathcal{A}, \mathbb{P}_{\boldsymbol{\mu}})$. If clear from context, we will write $\hat{C}_n = \hat{C}_n(\mathbf{y})$ for brevity. We say that $\hat{C}_n$ is an honest confidence set if

$$\inf_{\boldsymbol{\mu}\in K} \mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \in \hat{C}_n) \geq 1-\alpha. \tag{2.3}$$

We say that an honest confidence set $\hat{C}_n$ is adaptive in probability if for all $\gamma \in (0,1)$,

$$\inf_{k\in J} \inf_{\boldsymbol{\mu}\in E_k} \mathbb{P}_{\boldsymbol{\mu}}\left(\operatorname{diam}(\hat{C}_n)^2 \leq c' R^*_\alpha(E_k) \log\left(\frac{en}{\gamma\alpha}\right)^c\right) \geq 1-\gamma, \tag{2.4}$$

where $c' > 0$ and $c \geq 0$ are numerical constants. Alternatively to (2.4), we say that the confidence set $\hat{C}_n$ is adaptive in expectation if for all $k \in J$,

$$\sup_{\boldsymbol{\mu}\in E_k} \mathbb{E}_{\boldsymbol{\mu}}\left[\operatorname{diam}(\hat{C}_n)^2\right] \leq c' R^*_{\mathbb{E}}(E_k) \log\left(\frac{en}{\alpha}\right)^c, \tag{2.5}$$

where $c' > 0$ and $c \geq 0$ are numerical constants.

The role of the constant $c \geq 0$ is to allow for logarithmic factors. Inequality (2.3) requires that the true sequence $\boldsymbol{\mu}$ lies in $\hat{C}_n$ with probability $1-\alpha$ for all $\boldsymbol{\mu} \in K$. Inequality (2.4) implies that if the true parameter satisfies $\hat{\boldsymbol{\mu}} \in E_{k^*}$ for some $k^* \in J$, then the diameter of $\hat{C}_n$ is of the same order as the minimax quantity (2.2) of the model E_{k^*}, up to logarithmic factors.

We now consider a special case: confidence balls centered at the Least Squares estimator. The Least Squares estimator over a closed convex set K is defined by

$$\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) = \operatorname*{argmin}_{\boldsymbol{u}\in K} \|\mathbf{y}-\boldsymbol{u}\|_n^2 = \Pi_K(\mathbf{y})$$

where Π_K denotes the convex projection onto K. By definition of the convex projection onto K, we have $(\boldsymbol{u} - \Pi_K(\mathbf{y}))^T(\mathbf{y} - \Pi_K(\mathbf{y})) \leq 0$ for all $\boldsymbol{u} \in K$, which can be rewritten as

$$\left\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \mathbf{y}\right\|_n^2 \leq \|\boldsymbol{u}-\mathbf{y}\|_n^2 - \left\|\boldsymbol{u} - \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)\right\|_n^2. \tag{2.6}$$

If the confidence set $\hat{C}_n$ is an Euclidean ball, it is characterized by its center and its radius. Let $\alpha \in (0,1)$ be a confidence value, typically $\alpha = 0.05$. Let $K \subset \mathbb{R}^n$ be a closed convex set and let $(E_k)_{k\in J}$ be a collection of subsets of K indexed by an arbitrary set J. Let $\hat{r}$ be a positive random variable measurable with respect to $\mathbf{y}$ and let $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ be the Least Squares estimator over K. The set

$$\hat{C}_n = \left\{\boldsymbol{v} \in \mathbb{R}^n : \left\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{v}\right\|_n^2 \leq \hat{r}\right\} \tag{2.7}$$

is an honest confidence ball if (2.3) holds. The confidence ball $\hat{C}_n$ is said to be adaptive in probability if (2.4) holds, that is, for all $\gamma \in (0,1)$,

$$\inf_{k\in J} \inf_{\boldsymbol{\mu}\in E_k} \mathbb{P}_{\boldsymbol{\mu}}\left(\hat{r} \leq c' R^*_\alpha(E_k) \log\left(\frac{en}{\gamma\alpha}\right)^c\right) \geq 1-\gamma, \tag{2.8}$$

for all $\gamma \in (0,1)$ where $c' > 0$ and $c \geq 0$ are numerical constants. The confidence ball $\hat{C}_n$ is said to be adaptive in expectation if (2.5), that is,

$$\sup_{\boldsymbol{\mu} \in E_k} \mathbb{E}_{\boldsymbol{\mu}}[\hat{r}] \leq c' R^*_{\mathbb{E}}(E_k) \log\left(\frac{en}{\alpha}\right)^c, \tag{2.9}$$

for all $k \in J$, where $c' > 0$ and $c \geq 0$ are numerical constants.

3. Preliminaries

Throughout the paper, $C_1, C_2, C_3, \ldots$ denote positive absolute constants.

3.1. The cone of nondecreasing sequences and the models $(\mathcal{S}^{\uparrow}_n(k))_{k=1,\ldots,n}$

Let $\mathcal{S}^{\uparrow}_n$ be the set of all nondecreasing sequences, defined by

$$\mathcal{S}^{\uparrow}_n := \big\{\boldsymbol{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n : u_i \leq u_{i+1}, i = 1, \ldots, n-1\big\}.$$

For $n = 1$, let $\mathcal{S}^{\uparrow}_n = \mathbb{R}$. For all $n \geq 1$, define the cone of non-increasing sequences by $\mathcal{S}^{\downarrow}_n := -\mathcal{S}^{\uparrow}_n$.

For any $\boldsymbol{u} \in \mathcal{S}^{\uparrow}_n \cup \mathcal{S}^{\downarrow}_n$, let $k(\boldsymbol{u}) := |\{u_i, i = 1, \ldots, n\}|$ where $|A|$ denotes the cardinality of set A. The integer $k(\boldsymbol{u})$ is the smallest positive integer such that $\boldsymbol{u}$ is piecewise constant with $k(\boldsymbol{u})$ pieces. The integer $k(\boldsymbol{u}) - 1$ is also the number of jumps of $\boldsymbol{u}$, that is, the number of inequalities $u_i \leq u_{i+1}$ that are strict. Define the sets

$$\mathcal{S}^{\uparrow}_n(k) = \big\{\boldsymbol{u} \in \mathcal{S}^{\uparrow}_n : k(\boldsymbol{u}) \leq k\big\}, \quad k = 1, \ldots, n. \tag{3.1}$$

The set $\mathcal{S}^{\uparrow}_n(1)$ is the subspace of all constant sequences while $\mathcal{S}^{\uparrow}_n(2), \ldots, \mathcal{S}^{\uparrow}_n(n-1)$ are closed non-convex sets. This defines a collection of models ordered by inclusion,

$$\mathcal{S}^{\uparrow}_n(1) \subset \mathcal{S}^{\uparrow}_n(2) \subset \cdots \subset \mathcal{S}^{\uparrow}_n(n) = \mathcal{S}^{\uparrow}_n.$$

The minimax risk over the sets $\mathcal{S}^{\uparrow}_n(k)$ satisfies

$$C_4 \sigma^2 k \log\log(16n/k)/n \leq R^*_{\mathbb{E}}\big(\mathcal{S}^{\uparrow}_n(k)\big) \leq C_5 \sigma^2 k \log\log(16n/k)/n, \tag{3.2}$$

for $k \geq 2$, cf. [17]. The lower bound in (3.2) is proved in [17] using Fano's inequality, which actually yields a lower bound in probability and thus $C_6 k \log\log(16n/k)/n \leq R^*_{\alpha}(\mathcal{S}^{\uparrow}_n(k))$ for $\alpha \in (0, C_7)$. For for the upper bound, Markov's inequality implies that $R^*_{\alpha}(\mathcal{S}^{\uparrow}_n(k)) \leq R^*_{\mathbb{E}}(\mathcal{S}^{\uparrow}_n(k))/\alpha$. In summary, for $R^*_{\alpha}(\mathcal{S}^{\uparrow}_n(k))$,

$$C_8(\alpha)\sigma^2 k \log\log(16n/k)/n \leq R^*_{\alpha}\big(\mathcal{S}^{\uparrow}_n(k)\big) \leq \alpha^{-1} C_9 \sigma^2 k \log\log(16n/k)/n \tag{3.3}$$

for constants that only depend on $\alpha \in (0,1)$. Existing results also show that the dependence in α in the upper bound is at most logarithmic by properties of the Least-Squares estimator: for any $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}_n(k)$, the Least-Squares estimator satisfies

$$R^*_{\alpha}\big(\mathcal{S}^{\uparrow}_n(k)\big) \leq 2\sigma^2 k \log(en/k)/n + 4\sigma^2 \log(1/\alpha)/n,$$

cf. [3], Theorem 3.2. Thus, the quantity $R^*_{\alpha}(\mathcal{S}^{\uparrow}_n(k))$ is of order $k\sigma^2/n$, up to logarithmic factors in n and $1/\alpha$.

3.2. Statistical dimension and intrinsic volumes of cones

We recall here some properties of closed convex cones. Most of the material of the present section comes from [1,24]. In the present paper, a cone is always pointed at 0. A polyhedral cone is a closed convex cone of the form

$$K = \big\{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{u}^T \boldsymbol{v}_j \le 0 \text{ for all } j = 1, \dots, p\big\}, \tag{3.4}$$

where $\boldsymbol{v}_1, \dots, \boldsymbol{v}_p$ are vectors in $\mathbb{R}^n$, that is, K is the intersection of a finite number of half-spaces. The polar cone of K is defined as

$$K^\circ := \big\{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{v}^T \boldsymbol{\theta} \le 0 \text{ for all } \boldsymbol{v} \in K\big\}.$$

If K is a polyhedral cone, the face of K with outward vector $\boldsymbol{\theta} \in \mathbb{R}^n$ is the set

$$F(\boldsymbol{\theta}) := \Big\{\boldsymbol{u} \in K : \boldsymbol{u}^T \boldsymbol{\theta} = \sup_{\boldsymbol{v} \in K} \boldsymbol{v}^T \boldsymbol{\theta}\Big\}. \tag{3.5}$$

The face $F(\boldsymbol{\theta})$ is nonempty if and only if $\boldsymbol{\theta} \in K^\circ$. If K is the polyhedral cone (3.4) defined by the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$, a face of a polyhedral cone K has to be of the form

$$\big\{\boldsymbol{u} \in K : \boldsymbol{u}^T \boldsymbol{v}_j = 0 \text{ for all } j \in T\big\}$$

for some $T \subset \{1, \dots, k\}$. The dimension of a face F is the dimension of the affine span of F.

Definition 3.1 (Statistical dimension, Amelunxen et al. [1]). For any closed convex cone $K \subset \mathbb{R}^n$, define

$$\delta(K) := \mathbb{E}\big[\big|\Pi_K(\boldsymbol{g})\big|_2^2\big] = \mathbb{E}\big[\boldsymbol{g}^T \Pi_K(\boldsymbol{g})\big] = \mathbb{E}\Big[\Big(\sup_{\boldsymbol{\theta} \in K : |\boldsymbol{\theta}|_2 \le 1} \boldsymbol{g}^T \boldsymbol{\theta}\Big)^2\Big],$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, I_{n \times n})$. The quantity $\delta(K)$ is called the statistical dimension of the cone K.

It is also known that the following holds almost surely

$$\big|\Pi_K(\boldsymbol{g})\big|_2^2 = \boldsymbol{g}^T \Pi_K(\boldsymbol{g}) = \Big(\sup_{\boldsymbol{\theta} \in K : |\boldsymbol{\theta}|_2 \le 1} \boldsymbol{g}^T \boldsymbol{\theta}\Big)^2, \tag{3.6}$$

cf. [1], Proposition 3.1. The random variable (3.6) concentrates around its expectation. Indeed, the function $\boldsymbol{g} \mapsto |\Pi_K(\boldsymbol{g})|_2$ is 1-Lipschitz so that the Gaussian concentration theorem [5], Theorem 5.5, states that $|\Pi_K(\boldsymbol{g})|_2 \le \mathbb{E}|\Pi_K(\boldsymbol{g})|_2 + \sqrt{2\log(1/\alpha)}$ holds with probability at least $1 - \alpha$. Since $\mathbb{E}|\Pi_K(\boldsymbol{g})|_2 \le \delta(K)^{1/2}$ by Jensen's inequality, this implies a concentration result for the random variable (3.6): on the same event of probability at least $1 - \alpha$, one has

$$\big|\Pi_K(\boldsymbol{g})\big|_2^2 \le \delta(K) + 2\sqrt{2\log(1/\alpha)\delta(K)} + 2\log(1/\alpha) \le 2\delta(K) + 4\log(1/\alpha). \tag{3.7}$$

Our proofs also rely on the following facts given in [1], Proposition 3.1 (9), or [24], Section 5.2. If $K_1, K_2 \subset \mathbb{R}^n$ are two convex cones such that $K_1 \subset E_1$, $K_2 \subset E_2$ for two subspaces E_1, E_2 with $E_1 \perp E_2$, then $K = K_1 + K_2$ satisfies almost surely $\Pi_{K_1}(\boldsymbol{g}) = \Pi_{K_1}(\Pi_{E_1}(\boldsymbol{g}))$,

$$\Pi_K(\boldsymbol{g}) = \Pi_{K_1}(\boldsymbol{g}) + \Pi_{K_2}(\boldsymbol{g}), \qquad \big|\Pi_K(\boldsymbol{g})\big|_2^2 = \big|\Pi_{K_1}(\boldsymbol{g})\big|_2^2 + \big|\Pi_{K_2}(\boldsymbol{g})\big|_2^2.$$

Consequently $\Pi_{K_1}(\boldsymbol{g})$ and $\Pi_{K_2}(\boldsymbol{g})$ are independent if $\boldsymbol{g} \sim N(\boldsymbol{0}, I_{n\times n})$ and

$$\delta(K) = \delta(K_1) + \delta(K_2). \tag{3.8}$$

If $K_1 \in \mathbb{R}^q$, $K_2 \in \mathbb{R}^{n-q}$ and $K = K_1 \times K_2$ then (3.8) still holds by applying the previous argument to the two cones $K_1 \times \{0,\dots,0\} \subset \mathbb{R}^n$ and $\{0,\dots,0\} \times K_2 \subset \mathbb{R}^n$.

We now define the intrinsic volumes of a polyhedral cone, which are closely related to the statistical dimension.

Definition 3.2 (Intrinsic volumes of a polyhedral cone [1]). Let $K \subset \mathbb{R}^n$ be a polyhedral cone and let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, I_{n\times n})$. The intrinsic volumes of K are the real numbers

$$v_k(K) = \mathbb{P}\big(\Pi_K(\boldsymbol{g}) \text{ lies in the relative interior of a } k\text{-dimensional face of } K\big),$$

for all $k = 0,\dots,n$.

The intrinsic volumes of a polyhedral cone K define a probability distribution on the discrete set $\{0,\dots,n\}$. More precisely, define the random variable

$$V_K = \sum_{k=0}^{n} k \mathbf{1}_{\{\Pi_K(\boldsymbol{g}) \text{ lies in the relative interior of a } k\text{-dimensional face of } K\}}, \tag{3.9}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. The random variable V_K is valued in $\{0,\dots,n\}$ and satisfies $\mathbb{P}(V_K = k) = v_k(K)$ for all $k = 0,\dots,n$. The following identity was derived in [1,24]:

$$\delta(K) = \sum_{k=0}^{n} k v_k(K), \tag{3.10}$$

that is, the statistical dimension $\delta(K)$ is the expectation of the random variable V_K. Furthermore, the random variable V_K concentrates around its expected value. The following concentration inequality is given in [24], Corollary 4.10,

$$\mathbb{P}\big(V_K - \delta(K) \geq \lambda\big) \leq \exp\left(-\frac{\delta(K)}{2} h\left(\frac{\lambda}{\delta(K)}\right)\right), \quad \text{for all } \lambda > 0,$$

where $h(t) = (1+t)\log(1+t) - t$. Using the estimate $h^{-1}(t) \leq \sqrt{2t} + 3t$ (cf. [5], Corollary 12.12), we obtain

$$\mathbb{P}\big(V_K - \delta(K) \geq 2\sqrt{x\delta(K)} + 6x\big) \leq \exp(-x), \quad \text{for all } x > 0. \tag{3.11}$$

Deriving upper and lower bounds on the statistical dimension of a cone K may be a challenging problem. Some recipes to derive such bounds are proposed in [1,10]. An exact formula is available for the statistical dimension of the cone $\mathcal{S}_n^{\uparrow}$ [1], (D.12). It is given by

$$\delta\big(\mathcal{S}_n^{\uparrow}\big) = \delta\big(\mathcal{S}_n^{\downarrow}\big) = \sum_{k=1}^{n} \frac{1}{k}, \quad \text{so that} \quad \log n \leq \delta\big(\mathcal{S}_n^{\uparrow}\big) \leq \log(en). \tag{3.12}$$

Finally, we will need the following characterization of the faces of the cone $\mathcal{S}_n^{\uparrow}$.

Proposition 3.1. *Let $k \in \{1, \dots, n\}$.*

(i) *The faces of dimension k of the cone $\mathcal{S}_n^{\uparrow}$ are the sets*

$$F(S) := \big\{ \boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathcal{S}_n^{\uparrow} : u_{i-1} = u_i \text{ if } i \in S \big\} \tag{3.13}$$

where $S \subseteq \{2, \dots, n\}$ with $|S| = n - k$. The cone $\mathcal{S}_n^{\uparrow}$ has no face of dimension 0.

(ii) *For $k = 1, \dots, n$, the set $\mathcal{S}_n^{\uparrow}(k)$ is the union of all faces of dimension k.*

(iii) *For $K = \mathcal{S}_n^{\uparrow}$, the random variable V_K in (3.9) is equal to $k(\Pi_K(\boldsymbol{g}))$.*

Proof. (i) Let $K = \mathcal{S}_n^{\uparrow}$. Our definition of a face is given by an outward vector $\boldsymbol{\theta} \in \mathbb{R}^n$, and the corresponding face is given in (3.5). Every outward vector $\boldsymbol{\theta}$ gives a (possibly empty face) of K, and two distinct vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ may produce the same face in (3.5). The face with outward vector $\boldsymbol{\theta}$ is non-empty if and only if $\sup_{\boldsymbol{v} \in K} \boldsymbol{v}^T \boldsymbol{\theta} = 0$.

Here, $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$ if and only if $\boldsymbol{u} = X\boldsymbol{v}$ for some

$$\boldsymbol{v} \in \mathbb{R}^n \quad \text{with } v_2 \geq 0, \dots, v_n \geq 0 \tag{3.14}$$

and $X = (X_{ij}) \in \mathbb{R}^{n \times n}$ is the matrix with $X_{ij} = 1$ if $j \leq i$ and $X_{ij} = 0$ if $j > i$. Hence the face with outward vector $\boldsymbol{\theta}$ is non-empty if and only if $X^T\boldsymbol{\theta} \leq 0$ and $\sum_{i=1}^n \theta_i = 0$. Let $S \subset \{2, \dots, n\}$ be the set $\{i \in \{2, \dots, n\} : \theta_i + \theta_{i+1} + \cdots + \theta_n < 0\}$, that is, the indices of strict inequalities in $X^T\boldsymbol{\theta} \leq 0$. Then it is clear that if $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$ is of the form $\boldsymbol{u} = X\boldsymbol{v}$ with $\boldsymbol{v}$ in (3.14), inequality $\boldsymbol{u}^T\boldsymbol{\theta} = 0$ implies $\boldsymbol{v}^T X^T \boldsymbol{\theta} = 0$ and $v_i = 0$ for $i \in S$. Since $v_i = u_i - u_{i-1}$, if the face (3.5) for outward vector $\boldsymbol{\theta}$ is non-empty, then this face is of the form (3.13). Finally, the dimension of the face (3.13) is the rank of the matrix X with rows indexed in S removed, that is, $n - |S|$.

(ii) and (iii) follow by definition of $\mathcal{S}_n^{\uparrow}(k)$ in (3.1) since $\Pi_K(\boldsymbol{g})$ belongs to the relative interior of a face $F(S)$ if and only if $F(S)$ is the smallest face of K containing $\Pi_K(\boldsymbol{g})$. □

4. Nondecreasing sequences

4.1. Adaptive confidence sets for nondecreasing sequences

The estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ is the projection of $\mathbf{y}$ onto $\mathcal{S}_n^{\uparrow}$, so the vector $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ is nondecreasing. Let $\hat{k} = k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}))$ be the number of constant pieces of the Least Squares estimator. Using this notation, we define the statistic

$$\hat{r}_{\uparrow} = \frac{\sigma^2}{n}\Big(\sqrt{\hat{k}\log(en/\hat{k})} + \sqrt{2\big(\log(n/\alpha) + \hat{k}\log(en/\hat{k})\big)}\Big)^2. \tag{4.1}$$

Theorem 4.1. *For all $\alpha \in (0, 1)$ and all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$, the statistic $\hat{r}_{\uparrow}$ defined in (4.1) satisfies*

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|_n^2 \leq \hat{r}_{\uparrow}\big). \tag{4.2}$$

The above proposition shows that the confidence set (2.7) with $\hat{r} = \hat{r}_{\uparrow}$ satisfies condition (2.3). Up to constants and logarithmic factors, the number of constant pieces $\hat{k}$ of the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ bounds the loss $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|_n^2$ from above with high probability. Since $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ can be

computed in linear time, the integer $\hat{k}$ and the statistic $\hat{r}_\uparrow$ can also be computed in linear time. It is easy to compute $\hat{k}$ visually by drawing the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ and counting the number of jumps.

The proof of Theorem 4.1 relies on concentration properties of the random variable (3.6).

Proof of Theorem 4.1. Throughout the proof, we will consider partitions $(T_1, \dots, T_k)$ of $\{1, \dots, n\}$ such that each T_j satisfies $\max T_j < \min T_{j+1}$ as well as

$$T_j = \{\min T_j, \min T_j + 1, \dots, \max T_j\}, \tag{4.3}$$

that is, T_j contains all consecutive integers from $\min T_j$ to $\max T_j$.

Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ for notational simplicity. Then (2.6) with $\boldsymbol{u}$ replaced by $\boldsymbol{\mu}$ can be rewritten as $|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \le 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2$, which implies that $|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 \le \boldsymbol{\xi}^T\hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})/|(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})|_2$ has Euclidean norm 1. By definition of $k(\cdot)$, there exists a partition $(\hat{T}_1, \dots, \hat{T}_{\hat{k}})$ of $\{1, \dots, n\}$ such that $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ is constant on each $\hat{T}_j$, $j = 1, \dots, \hat{k}$. Since $\boldsymbol{\mu} \in \mathcal{S}_n^\uparrow$, both $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ and $\hat{\boldsymbol{\theta}}$ belong to the product cone

$$\hat{\mathcal{C}} := \mathcal{S}^\downarrow_{|\hat{T}_1|} \times \mathcal{S}^\downarrow_{|\hat{T}_2|} \times \cdots \times \mathcal{S}^\downarrow_{|\hat{T}_{\hat{k}}|}. \tag{4.4}$$

By the Gaussian concentration theorem [5], Theorem 5.5, 5.6,

$$\sup_{\boldsymbol{\theta} \in \mathcal{C}: |\boldsymbol{\theta}|_2 = 1} \boldsymbol{\xi}^T\boldsymbol{\theta} \le \mathbb{E} \sup_{\boldsymbol{\theta} \in \mathcal{C}: |\boldsymbol{\theta}|_2 = 1} \boldsymbol{\xi}^T\boldsymbol{\theta} + \sigma\sqrt{2x} \tag{4.5}$$

with probability at least $1 - e^{-x}$. Furthermore, $\mathbb{E}\sup_{\boldsymbol{\theta} \in \mathcal{C}: |\boldsymbol{\theta}|_2 = 1} \boldsymbol{\xi}^T\boldsymbol{\theta} \le \sigma\delta(\mathcal{C})^{1/2}$ by Jensen's inequality, and if $\mathcal{C}$ is of the form $\mathcal{C} = \mathcal{S}^\downarrow_{n_1} \times \cdots \times \mathcal{S}^\downarrow_{n_k}$ for positive integers $n_1, \dots, n_k$ such that $n_1 + \cdots + n_k = n$, then $\delta(\mathcal{C}) = \sum_{j=1}^k \delta(\mathcal{S}^\downarrow_{n_j}) \le \sum_{j=1}^k \log(en_j) \le k\log(en/k)$ thanks to (3.12) and (3.8).

Let $k = 1, \dots, n$ be fixed. There are $\binom{n-1}{k}$ partitions of $\{1, \dots, n\}$ of the form $(T_1, \dots, T_k)$ with $\max T_j < \min T_{j+1}$ for all $j = 1, \dots, k-1$ (each such partition defines a unique configuration of jumps). By the union bound and the inequality $\log\binom{n-1}{k} \le k\log(en/k)$, we have with probability at least $1 - e^{-x}$ the bound

$$\sup_{(T_1, \dots, T_k)} \Big(\sup_{\boldsymbol{\theta} \in \mathcal{S}^\downarrow_{|T_1|} \times \cdots \times \mathcal{S}^\downarrow_{|T_k|}: |\boldsymbol{\theta}|_2 = 1} \boldsymbol{\xi}^T\boldsymbol{\theta} \Big) \le \sigma\sqrt{k\log(en/k)} + \sigma\sqrt{2\big(x + k\log(en/k)\big)}.$$

Finally, we apply the union bound over all $k \in \{1, \dots, n\}$ and set $x = \log(n/\alpha)$. We have established that with probability at least $1 - \alpha$, for any random partition $(\hat{T}_1, \dots, \hat{T}_{\hat{k}})$ and $\hat{\mathcal{C}}$ in (4.4),

$$\sup_{\boldsymbol{\theta} \in \hat{\mathcal{C}}: |\boldsymbol{\theta}|_2 = 1} \boldsymbol{\xi}^T\boldsymbol{\theta} \le \sigma\sqrt{\hat{k}\log(en/\hat{k})} + \sigma\sqrt{2\big(\log(n/\alpha) + \hat{k}\log(en/\hat{k})\big)}. \tag{4.6}$$

In particular, this is true for the partition $(\hat{T}_1, \dots, \hat{T}_{\hat{k}})$ induced by $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ defined in the previous paragraph. Note also that the right-hand side of the previous display is equal to $(n\hat{r}_\uparrow)^{1/2}$. On this event of probability at least $1 - \alpha$ we have $|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 \le \boldsymbol{\xi}^T\hat{\boldsymbol{\theta}} \le (n\hat{r}_\uparrow)^{1/2}$ and the proof is complete. □

We have established the existence of an honest confidence interval of the form

$$\hat{C}_n := \big\{\boldsymbol{v} \in \mathcal{S}_n^\uparrow : \big\|\boldsymbol{v} - \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^\uparrow\big)\big\|_n^2 \le \hat{r}\big\}.$$

This confidence set has uniform coverage over all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$, that is, it satisfies (2.3). The next result implies that the diameter of this confidence set is minimax optimal up to logarithmic factors.

Theorem 4.2. *Let $(T_1, \dots, T_k)$ be a partition of $\{1, \dots, n\}$ with $\max T_j < \min T_{j+1}$ for every $j = 1, \dots, k-1$, and assume that $\boldsymbol{\mu}$ is constant on each T_j Let $\gamma \in (0,1)$. The random variable $\hat{k} = k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}))$ satisfies*

$$\begin{aligned} \hat{k} &\le k \log(en/k) + 2\sqrt{\log(1/\gamma) k \log(en/k)} + 6\log(1/\gamma) \\ &\le 2k \log(en/k) + 7\log(1/\gamma)) \end{aligned} \tag{4.7}$$

with probability greater than $1-\gamma$ as well as

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \le k \log(en/k). \tag{4.8}$$

Furthermore $\sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}(k)} \mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \ge k \log(n/k)$ when k divides n so that (4.8) is sharp.

Note that in Theorem 4.2 $\boldsymbol{\mu}$ is only assumed to be piecewise constant, but not necessarily monotone. This will be useful in Section 6.

Proof of Theorem 4.2. Define the closed convex cone

$$K = \mathcal{S}_{|T_1|}^{\uparrow} \times \mathcal{S}_{|T_2|}^{\uparrow} \times \cdots \times \mathcal{S}_{|T_k|}^{\uparrow} \subset \mathbb{R}^n \tag{4.9}$$

and let $\hat{\boldsymbol{\mu}}^* = \Pi_K(\mathbf{y})$. It is clear that

$$\min_{\boldsymbol{u} \in K} \sum_{j=1}^{k} |\mathbf{y}_{T_j} - \boldsymbol{u}_{T_j}|_2^2 = \min_{\boldsymbol{u}_1 \in \mathcal{S}_{|T_1|}^{\uparrow}, \dots, \boldsymbol{u}_k \in \mathcal{S}_{|T_k|}^{\uparrow}} \sum_{j=1}^{k} |\mathbf{y}_{T_j} - \boldsymbol{u}_{T_j}|_2^2 = \sum_{j=1}^{k} \min_{\boldsymbol{u}_j \in \mathcal{S}_{|T_j|}^{\uparrow}} |\mathbf{y}_{T_j} - \boldsymbol{u}_{T_j}|_2^2.$$

Thus, as $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi}$ and $\boldsymbol{\mu}$ is constant on each T_j, we have

$$\hat{\boldsymbol{\mu}}_{T_j}^* = \Pi_{\mathcal{S}_{|T_j|}^{\uparrow}}(\mathbf{y}_{T_j}) = \Pi_{\mathcal{S}_{|T_j|}^{\uparrow}}(\boldsymbol{\mu}_{T_j} + \boldsymbol{\xi}_{T_j}) = \boldsymbol{\mu}_{T_j} + \Pi_{\mathcal{S}_{|T_j|}^{\uparrow}}(\boldsymbol{\xi}_{T_j}).$$

As adding the constant sequence $\boldsymbol{\mu}_{T_j}$ does not modify the number of constant pieces (or the number of jumps), we have

$$k\big(\hat{\boldsymbol{\mu}}_{T_j}^*\big) = k\big(\boldsymbol{\mu}_{T_j} + \Pi_{\mathcal{S}_{|T_j|}^{\uparrow}}(\boldsymbol{\xi}_{T_j})\big) = k\big(\Pi_{\mathcal{S}_{|T_j|}^{\uparrow}}(\boldsymbol{\xi}_{T_j})\big) = k\big(\big(\Pi_K(\boldsymbol{\xi})\big)_{T_j}\big).$$

Let V_K be the random variable defined in (3.9). By the properties of product cones given in [24], Section 5.2, V_K has the same distribution as

$$V_{\mathcal{S}_{|T_1|}^{\uparrow}} + \cdots + V_{\mathcal{S}_{|T_k|}^{\uparrow}}$$

and the random variables $V_{\mathcal{S}_{|T_1|}^{\uparrow}}, \dots, V_{\mathcal{S}_{|T_k|}^{\uparrow}}$ are independent (cf. [24], (5.2)). By Proposition 3.1, for all $j = 1, \dots, k$ we have $k(\hat{\boldsymbol{\mu}}_{T_j}^*) = V_{\mathcal{S}_{|T_j|}^{\uparrow}}$ so that V_K and $\sum_{j=1}^{k} k(\hat{\boldsymbol{\mu}}_{T_j}^*)$ have the same distribution. By (3.10), $\mathbb{E}V_K = \delta(K)$ and by (3.11), with probability greater than $1-\gamma$ we have

$$V_K \le \delta(K) + 2\sqrt{\log(1/\gamma)\delta(K)} + 6\log(1/\gamma).$$

To bound $\delta(K)$ from above, we use that the statistical dimension of a direct product of cones is the sum of the statistical dimensions (cf. [1], Proposition 3.1, or the discussion preceding (3.8)),

$$\delta(K) = \sum_{j=1}^{k} \delta\big(\mathcal{S}^{\uparrow}_{|T_j|}\big) \le \sum_{j=1}^{k} \log\big(e|T_j|\big) \le k\log(en/k),$$

where we have used (3.12) and Jensen's inequality for the two inequalities.

The random variable V_K is distributed as $\sum_{j=1}^{k} k(\hat{\boldsymbol{\mu}}^*_{T_j})$. Thus, to complete the proof, it is enough to prove that almost surely, $\hat{k} := k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\uparrow}_n)) \le \sum_{j=1}^{k} k(\hat{\boldsymbol{\mu}}^*_{T_j})$. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\uparrow}_n)$ for notational simplicity. It is clear that

$$k(\hat{\boldsymbol{\mu}}) = \big|\{\hat{\mu}_i, i = 1,\dots,n\}\big| \le \sum_{j=1}^{k} k(\hat{\boldsymbol{\mu}}_{T_j}) = \sum_{j=1}^{k} \big|\{\hat{\mu}_i, i \in T_j\}\big|,$$

since a piece counted on the left-hand side must be counted at least once (and possibly multiple times) on the right-hand side. For all $j = 1,\dots,k$, the vectors $\hat{\boldsymbol{\mu}}_{T_j}$ and $\hat{\boldsymbol{\mu}}^*_{T_j}$ are solutions of the minimization problems

$$\hat{\boldsymbol{\mu}}^*_{T_j} = \operatorname*{argmin}_{\boldsymbol{v}\in\mathcal{S}^{\uparrow}_{|T_j|}} |\boldsymbol{v} - \mathbf{y}_{T_j}|_2^2, \hat{\boldsymbol{\mu}}_{T_j} = \operatorname*{argmin}_{\substack{\boldsymbol{v}\in\mathcal{S}^{\uparrow}_{|T_j|}:\\ \hat{\boldsymbol{\mu}}_{\min(T_j)}\le \boldsymbol{v}_1,\\ \boldsymbol{v}_{|T_j|}\le \hat{\boldsymbol{\mu}}_{\max(T_j)}}} |\boldsymbol{v} - \mathbf{y}_{T_j}|_2^2.$$

This means that $\hat{\boldsymbol{\mu}}_{T_j}$ solves a minimization problem with additional constraints at the boundary. By Lemma B.1, we have

$$k(\hat{\boldsymbol{\mu}}_{T_j}) \le k\big(\hat{\boldsymbol{\mu}}^*_{T_j}\big)$$

for all $j = 1,\dots,k$, which completes the proof of Equation (4.8).

To show that $\sup_{\boldsymbol{\mu}\in\mathcal{S}^{\uparrow}_n(k)} \mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \ge k\log(n/k)$ when k divides n, consider a partition $(T_1,\dots,T_k)$ with $|T_j| = n/k$ and $\max T_j \le \min T_{j+1}$ and set $\mu_i = xj$ if $i \in T_j$ for some parameter $x > 0$. The jumps of $\boldsymbol{\mu}$ have amplitude x. In the event $x > 2\max_{i=1}^{n} |\xi_i|$ we have $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\uparrow}_n) = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ where K is the cone (4.9). By dominated convergence and using the fact that adding a constant vector to $\boldsymbol{u} \in \mathbb{R}^{|T_j|}$ does not change the number of jumps of $\Pi_{\mathcal{S}^{\uparrow}_{|T_j|}}(\boldsymbol{u})$,

$$\lim_{x\to+\infty} \mathbb{E}_{\boldsymbol{\mu}}\big[k\big(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}^{\uparrow}_n\big)\big)\big] = \mathbb{E}_{\boldsymbol{\mu}}\big[k\big(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)\big)\big] = \mathbb{E}_{\mathbf{0}}\big[k\big(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)\big)\big] = \delta(K) = \sum_{j=1}^{k} \delta\big(\mathcal{S}^{\uparrow}_n\big(|T_j|\big)\big).$$

This quantity is greater than $k\log(n/k)$ due to the lower bound in (3.12). □

Corollary 4.3. *Let $J = \{1,\dots,n\}$ and define the collection of models $(E_k)_{k\in J} = (\mathcal{S}^{\uparrow}_n(k))_{k\in J}$. The random variable $\hat{r}_{\uparrow}$ defined in (4.1) satisfies (4.2), (2.8) and (2.9) with $\hat{r}$ replaced by $\hat{r}_{\uparrow}$. Thus, the ball centered at $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\uparrow}_n)$ of radius $\sqrt{\hat{r}_{\uparrow}}$ is an honest confidence set, which is adaptive in probability and in expectation with respect to the models $(\mathcal{S}^{\uparrow}_n(k))_{k=1,\dots,n}$.*

Proof. To show that (2.9) holds, for any $\boldsymbol{\mu} \in \mathcal{S}^{\uparrow}_n(k)$ we have

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{r}_{\uparrow}]^{1/2} \le \sigma n^{-1/2}\big(\sqrt{2\log(n/\alpha)} + (1+\sqrt{2})\sqrt{\mathbb{E}_{\boldsymbol{\mu}}\big[\hat{k}\log(en/\hat{k})\big]}\big)$$

thanks to $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and the triangle inequality for $\mathbb{E}_{\boldsymbol{\mu}}[(\cdot)^2]^{1/2}$. Since $x \mapsto x \log(en/x)$ is concave on $[1, n]$, $\mathbb{E}_{\boldsymbol{\mu}}[\hat{k} \log(en/\hat{k})] \le \mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \log(en/\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}])$. The bound (4.8) then yields $\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \times \log(en/\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}]) \le k \log(en/k)^2$ which completes the proof of (2.9) for the minimax risk given in (3.2), since (2.9) allows logarithmic factors.

To show that (2.8) holds for the minimax risk in (3.3), we instead use (4.7). This yields that with probability at least $1 - \gamma$,

$$\begin{aligned}\hat{r}_{\uparrow}^{1/2} &\le \sigma n^{1/2}\Big(\sqrt{2\log(n/\alpha)} + (1+\sqrt{2})\sqrt{\hat{k}\log(en/\hat{k})}\Big) \\ &\le \sigma n^{1/2}\Big(\sqrt{2\log(n/\alpha)} + (1+\sqrt{2})\sqrt{2k\log(en/k)\log(en) + 7\log(1/\gamma)\log(en)}\Big).\end{aligned}$$

This completes the proof since (2.8) allows logarithmic factors. □

4.2. Adaptive confidence sets do not exist for piecewise constant signal

The existence of honest and adaptive confidence sets with respect to the models $(\mathcal{S}_n^{\uparrow}(k))_{k=1,\dots,n}$ is surprising given the negative results for seemingly similar models. For the problem of constructing adaptive confidence sets in sparse linear regression with p covariates and $n \le p$ observations, where one observes

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j^* + \xi_i, \quad i = 1, \dots, n \text{ where } X_{ij}, \xi_i \sim^{iid} \mathcal{N}(0,1),$$

Nickl and van de Geer [26], Theorem 2, shows the impossibility to construct confidence sets for $\boldsymbol{\beta}^*$ that are (a) honest over all sparsity levels $k_1 \in \{k = 1, 2, 3, \dots : k \le n/\log p\}$ and (b) adaptive over sparsity levels $k_0 \in \{k = 1, 2, 3, \dots, : k_1 \le \sqrt{n}/\log p\}$. This impossibility result holds as $n, p, k \to +\infty$ with $|\boldsymbol{\beta}^*|_0 = p^{1-\eta}$ for some constant $\eta \in (0,1)$. Due to this impossibility result, confidence sets over all sparsity levels $k_1 \in \{k = 1, 2, 3, \dots : k \le n/\log p\}$ cannot have squared radius smaller than the minimax estimation rate $|\boldsymbol{\beta}^*|_0 \log(p)/n$ for $|\boldsymbol{\beta}^*|_0 \le \sqrt{n}/\log p$. This impossibility result for honest and adaptive confidence sets in sparse linear regression builds upon impossibility results for signal detection [22]. Similar impossibility results have been obtained for testing sparsity levels in the Gaussian mean model [2,9] and for testing sparsity levels in linear regression [8]. We refer the reader to Carpentier and Verzelen [9] and Table 1 therein for a complete characterization of the different phase transitions for testing sparsity levels in the Gaussian mean model, and [8], Table 1, for the corresponding phase transitions in sparse linear regression.

The models $(\mathcal{S}_n^{\uparrow}(k))_{k=1,\dots,n}$ can be related to sparse linear regression in the following way:

$$\mathcal{S}_n^{\uparrow}(k) = \big\{X\boldsymbol{v}, \boldsymbol{v} \in \mathbb{R} \times \mathbb{R}_+^{n-1} \text{ with } \|\boldsymbol{v}\|_0 = k-1\big\} \tag{4.10}$$

and $X \in \mathbb{R}^{n \times n}$ is the design matrix with entries $X_{ij} = I_{i \le j}$. Here, $\mathbb{R}_+ = [0, +\infty)$ and $I_{i \le j}$ is the indicator function equal to 1 if and only if $i \le j$. Thus, the model $\mathcal{S}_n^{\uparrow}(k)$ can be seen as a sparse linear regression model with design X and at most k nonzero entries, with an additional non-negativity constraint.

The major difference in (4.10) compared to sparse linear regression is the non-negativity of $v_2, \dots, v_n$ in (4.10), which encodes the shape constraint. If one drops the non-negativity constraints, the above model becomes the set of piecewise constant signals with $k-1$ jumps:

$$\mathcal{B}_n(k) = \big\{\boldsymbol{v} \in \mathbb{R}^n : \big|\{i = 2, \dots, n : v_i \ne v_{i-1}\}\big| = k-1\big\}.$$

The minimax rate over $\mathcal{B}_n(k)$ satisfies

$$R^*_{\mathbb{E}}\big(\mathcal{B}_n(k)\big) \asymp \begin{cases} \sigma^2/n & \text{if } k=1, \\ k\log\log(16n/k)\sigma^2/n & \text{if } k=2, \\ k\log(en/k)\sigma^2/n & \text{if } k\geq 3, \end{cases} \tag{4.11}$$

cf. [17], Theorem 4.1. The following result shows that honest and adaptive confidence sets do not exist for the models $(\mathcal{B}_n(k))_{k=1,\dots,n}$. This a consequence of the lower bound in [2].

Proposition 4.4. *Let $\alpha \in (0,1/8)$ independent of n. In the model (1.1), let $\hat{C}_n$ be a confidence set such that, for all $k \leq n^{1/2}$ inequality $\inf_{\boldsymbol{\mu}\in\mathcal{B}_n(k)} \mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu}\in\hat{C}_n) \geq 1-\alpha$ holds. Then there exists an absolute constant C_{12} such that*

$$\mathbb{E}_{\mathbf{0}}\big[\mathrm{diam}(\hat{C}_n)^2\big] \geq \sigma^2 C_{12} n^{-1/2} \tag{4.12}$$

for all n larger than some absolute constant $C_{10}>0$.

Thus a confidence set $\hat{C}_n$ that is honest with respect to the models $(\mathcal{B}_n(k))_{k\leq\sqrt{n}}$ must incur a squared radius of order at least $\sigma^2 n^{-1/2}$ which is larger than the minimax rate in (4.11) for $k \lll \sqrt{n}$. Furthermore, the discrepancy between the best possible squared radius and the minimax rate (4.11) is polynomial in n when $k=\lfloor n^\kappa \rfloor$ for $\kappa < 1/2$:

$$\frac{\mathbb{E}_{\mathbf{0}}[\mathrm{diam}(\hat{C}_n)^2]}{R^*_{\mathbb{E}}(\mathcal{B}_n(\lfloor n^\kappa\rfloor))} \geq C_{11}(\kappa)\frac{n^{1/2-\kappa}}{\log n}$$

for all confidence sets that are honest over $(\mathcal{B}_n(k))_{k\leq\sqrt{n}}$. The constant in the previous display depends on κ only. The $\sigma^2 n^{-1/2}$ lower bound in (4.12) also appears for confidence sets in sparse regression [8,22,26] and testing sparsity levels in sparse mean models [2,9]. The non-decreasing shape constraint of the previous section leads to a fundamentally different behavior: Theorems 4.1 and 4.2 show that it possible to construct confidence sets that are both honest and adaptive (up to logarithmic factors) over all models in (3.1), i.e., with squared radius of the same order as the minimax rate.

Because any $k/2$-sparse vector is is also piecewise constant with at most k jumps, Proposition 4.4 is a straightforward consequence of the lower bound given in [2], Theorem 1.

Proof of Proposition 4.4. Let $r_n^2 = 8\mathbb{E}_{\mathbf{0}}[\mathrm{diam}(\hat{C}_n)^2]$. We follow the argument laid out in [26], Section 3.1, to relate confidence sets to signal detection problems. Let $k_1 = \lfloor\sqrt{n}\rfloor$ be the largest integer smaller or equal to $\sqrt{n}$ and consider testing

$$H_0 : \boldsymbol{\mu} = \mathbf{0}, \quad \text{against } H_1 : \boldsymbol{\mu}\in\mathcal{B}_n(k_1) \text{ with } \|\boldsymbol{\mu}\|_n^2 \geq r_n^2.$$

With a minor abuse of notation, we denote by H_1 the set $H_1 = \{\boldsymbol{u}\in\mathcal{B}_n(k_1) \text{ with } \|\boldsymbol{u}\|_n^2 \geq r_n^2\}$. Consider the test statistic T_n valued in $\{0,1\}$ defined by $T_n=1$ if $\hat{C}_n \cap H_1 \neq \varnothing$ and $T_n=0$ otherwise. The type I and type II errors of T_n are bounded as follows:

- Under the null, on the event $\Omega_0 = \{\mathbf{0}\in\hat{C}_n\}\cap\{\mathrm{diam}(\hat{C}_n)^2 < 8\mathbb{E}_{\mathbf{0}}[\mathrm{diam}(\hat{C}_n)^2]\}$ we have $\hat{C}_n\cap H_1 = \varnothing$ so that $\mathbb{P}_{\mathbf{0}}(T_n=1) \leq \mathbb{P}(\Omega_0^c) \leq \alpha + 1/8$ by the union bound and Markov's inequality.
- Under the alternative, if $\boldsymbol{\mu}\in H_1$ then $\mathbb{P}_{\boldsymbol{\mu}}(T_n=0) \leq \mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu}\notin\hat{C}_n) \leq \alpha$.

The sum of type I and type II errors of the above test is thus at most $2\alpha + 1/8$. Define now $H_1' = \{\boldsymbol{\mu} \in \mathbb{R}^n : |\boldsymbol{\mu}|_0 \le k_1/2\}$. Since any $(k_1/2)$-sparse vector is also piecewise constant with k_1 pieces, $H_1' \subset H_1$ holds and

$$\mathbb{P}_{\mathbf{0}}(T_n = 1) + \sup_{\boldsymbol{\mu} \in H_1'} \mathbb{P}_{\boldsymbol{\mu}}(T_n = 0) \le \mathbb{P}_{\mathbf{0}}(T_n = 1) + \sup_{\boldsymbol{\mu} \in H_1} \mathbb{P}_{\boldsymbol{\mu}}(T_n = 0)$$
$$\le 2\alpha + 1/8 \le 3/8.$$

By [2], Theorem 1, if $r_n^2 \le C_{12}\sigma^2 n^{-1/2}$ for some absolute constant $C_{12} > 0$, there exists no test T_n with sum of type I and type II errors less than $3/8$ for testing $H_0 : \boldsymbol{\mu} = 0$ against H_1'. This implies (4.12) and completes the proof. □

4.3. Adaptivity to unknown noise level in non-decreasing sequences

Although the radius $\hat{r}_\uparrow$ in (4.1) depends on σ^2, Theorem 4.1 is also helpful to construct estimators $\hat{\sigma}^2$ of σ^2. Indeed, Theorem 4.1 implies that with probability at least $1 - \alpha - \beta$,

$$\begin{aligned} \big|1 - \sigma^{-1}\big\|\mathbf{y} - \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^\uparrow\big)\big\|_n\big| &\le \big|1 - \sigma^{-1}\|\boldsymbol{\xi}\|_n\big| + \sigma^{-1}\big\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^\uparrow\big) - \boldsymbol{\mu}\big\|_n \\ &\le \sqrt{2/n}\big(1 + \sqrt{\log(1/\beta)}\big) + \sigma^{-1}\hat{r}_\uparrow^{1/2} \end{aligned} \tag{4.13}$$

where we used that $\boldsymbol{\xi} \to |\|\boldsymbol{\xi}\|_n - \sigma|$ is a $n^{-1/2}$-Lipschitz function of $\boldsymbol{\xi}$ with $\mathbb{E}[|\|\boldsymbol{\xi}\|_n - \sigma|]^2 \le 2\sigma^2(1 - \sqrt{1 - 1/n}) \le 2\sigma^2/n$ and the Gaussian concentration result [5], Theorem 5.5, which gives $|\|\boldsymbol{\xi}\|_n - \sigma| \le \sigma\sqrt{2/n} + \sigma\sqrt{2\log(1/\beta)/n}$. Now let

$$\hat{\epsilon} = n^{-1/2}\big(\sqrt{2} + \sqrt{2\log(1/\beta)} + \sqrt{\hat{k}\log(en/\hat{k})} + \sqrt{2\big(\log(n/\alpha) + \hat{k}\log(en/\hat{k})\big)}\big)$$

be the right-hand side in (4.13), which only depends on α, $\hat{k}$, n. Set also $\hat{\sigma}^2 = \|\mathbf{y} - \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)\|_n^2$. Then on the above event of probability at least $1 - 2\alpha$, we have

$$|1 - \hat{\sigma}/\sigma| \le \hat{\epsilon}, \qquad \sigma(1 - \hat{\epsilon})_+ \le \hat{\sigma} \le \sigma(1 + \hat{\epsilon}), \qquad (1 + \hat{\epsilon})^{-1}\hat{\sigma} \le \sigma \le (1 - \hat{\epsilon})_+^{-1}\hat{\sigma}$$

so that for instance $(1 - \hat{\epsilon})^{-1}\hat{\sigma}$ provides a fully data-driven, high-probability upper bound for σ as soon as $\hat{\epsilon}$ is bounded away from 1. We obtain, as a consequence, the following data-driven analog of the results of the previous section with fully data-driven radius.

Theorem 4.5 (Unknown σ). *Let $\alpha, \beta \in (0, 1)$ be fixed constants independent of n. Consider the above notation for $\hat{\epsilon}$ and define*

$$\tilde{r}_\uparrow = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)\|_n^2}{(1 - \hat{\epsilon})_+^2 n}\big(\sqrt{\hat{k}\log(en/\hat{k})} + \sqrt{2\big(\log(n/\alpha) + \hat{k}\log(en/\hat{k})\big)}\big)^2.$$

Then we have

(i) $\mathbb{P}(\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow) - \boldsymbol{\mu}\|^2 \le \tilde{r}_\uparrow) \ge 1 - \alpha - \beta$.
(ii) $k(\boldsymbol{\mu})/n \to 0$ *as* $n \to +\infty$ *implies* $\hat{\epsilon} \to 0$ *in probability.*
(iii) $k(\boldsymbol{\mu})/n \to 0$ *as* $n \to +\infty$ *implies* $\tilde{r}_\uparrow = O_{\mathbb{P}}(\sigma^2(k(\boldsymbol{\mu})/n)\log(en/k(\boldsymbol{\mu}))^2$.

Hence the fully data-driven radius $\tilde{r}_\uparrow$ is honest and shrinks adaptively with respect to the models $(\mathcal{S}_n^\uparrow(k))_{k\le na_n}$ for any sequence $a_n \to 0$. Alternatively, one may use the median absolute deviation to estimate σ^2 as explained, for instance, in Remark 3.1 of [17].

Proof of Theorem 4.5. (i) follows from Theorem 4.1 and (4.13). For (ii), if $a_n = k(\boldsymbol{\mu})/n$ converges to 0 then by Theorem 4.2 with $\gamma = 1/n$ we have $\hat{k}/n \le 2a_n \log(ea_n^{-1}) + C_{13}\log(n)/n =: b_n$ with probability at least $1/n$. Furthermore, $a_n \to 0$ implies $b_n \to 0$. On the same event we have $\hat{\epsilon} \le (1+\sqrt{2})\{b_n \log(eb_n^{-1})\}^{1/2} + C_{14}\sqrt{\log(en)/n}$ which also converges to 0. For (iii), $\hat{\epsilon} \to 0$ in probability implies that $\hat{\sigma}^2(1-\hat{\epsilon})_+^{-2} = O_{\mathbb{P}}(\sigma^2)$, so that and Theorem 4.2 provide the required upper bound on $\tilde{r}_\uparrow$. □

5. Adaptive confidence sets for convex sequences

Confidence sets can also be obtained in univariate convex regression. If $n \ge 3$, define the set of convex sequences $\mathcal{S}_n^\cup$ by

$$\mathcal{S}_n^\cup := \big\{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : 2u_i \le u_{i+1} + u_{i-1}, i = 2, \dots, n-1\big\},$$

and define $\mathcal{S}_n^\cup = \mathbb{R}$ if $n = 1$ and $\mathcal{S}_n^\cup = \mathbb{R}^2$ if $n = 2$. For all $n \ge 1$, define the cone of concave sequences by $\mathcal{S}_n^\cap := -\mathcal{S}_n^\cup$.

For any $\boldsymbol{u} \in \mathcal{S}_n^\cup$, let $q(\boldsymbol{u}) - 1 \ge 0$ be the number of inequalities $2u_i \le u_{i+1} + u_{i-1}$, $i = 2, \dots, n-1$ that are strict. The integer $q(\boldsymbol{u})$ is also the smallest positive integer such that $\boldsymbol{u}$ is piecewise affine with $q(\boldsymbol{u})$ pieces. Define the sets

$$\mathcal{S}_n^\cup(q) = \big\{\boldsymbol{u} \in \mathcal{S}_n^\cup : q(\boldsymbol{u}) \le q\big\}, \quad q = 1, \dots, n-1.$$

The set $\mathcal{S}_n^\cup(1)$ is the subspace of all affine sequences while $\mathcal{S}_n^\cup(2), \dots, \mathcal{S}_n^\cup(n-2)$ are closed non-convex sets. We have

$$\mathcal{S}_n^\cup(1) \subset \mathcal{S}_n^\cup(2) \subset \cdots \subset \mathcal{S}_n^\cup(n-1) = \mathcal{S}_n^\cup.$$

These sets represent models of increasing complexity.

There exist numerical constants $c, c' > 0$ such that for all $\alpha \le (0, \min(c, 1))$ and any $q = 1, \dots, n-1$, we have

$$\frac{c'\sigma^2 q}{n} \le R_\alpha^*\big(\mathcal{S}_n^\cup(q)\big) \le \frac{16\sigma^2 q \log(en/q)}{n} + \frac{4\log(1/\alpha)}{n}, \tag{5.1}$$

cf. [3], Theorem 4.3, for the upper bound and [4], Proposition 7, for the lower bound. Thus, for $\alpha > 0$ small enough, the quantity $R_\alpha^*(\mathcal{S}_n^\cup(q))$ is of order $q\sigma^2/n$, up to logarithmic factors.

The statistical dimension of the cone $\mathcal{S}_n^\cup$ satisfies [3], Theorem 4.1,

$$\delta\big(\mathcal{S}_n^\cup\big) = \delta\big(\mathcal{S}_n^\cap\big) \le 8\log(en). \tag{5.2}$$

It is not known whether this upper bound is sharp. However, the fact that the statistical dimension of $\mathcal{S}_n^\cup$ grows slower than a logarithmic function of n is enough for the purpose of the present paper.

The following bound on the risk of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\cup)$ will be useful.

Proposition 5.1. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}} \big| \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\cup}\big) - \boldsymbol{\mu} \big|_2^2 \le \mathbb{E}_{\boldsymbol{\mu}} \Big[\Big(\sup_{\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{\mu}} : |\boldsymbol{v}|_2 \le 1} \boldsymbol{\xi}^T \boldsymbol{v} \Big)^2 \Big] = \sigma^2 \delta(\mathcal{T}_{\boldsymbol{\mu}}) \le 8\sigma^2 q(\boldsymbol{\mu}) \log \frac{en}{q(\boldsymbol{\mu})},$$

where $\mathcal{T}_{\boldsymbol{\mu}}$ is the tangent cone at $\boldsymbol{\mu}$ defined by $\mathcal{T}_{\boldsymbol{\mu}} := \{\boldsymbol{u} - t\boldsymbol{\mu}, \boldsymbol{u} \in \mathcal{S}_n^{\cup}, t \in \mathbb{R}\}$.

An outline of the proof of this result is as follows. More details may be found in [3], Proposition 4.3.

Outline of the proof of Proposition 5.1. The inequality $\mathbb{E}_{\boldsymbol{\mu}} |\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}|_2^2 \le \sigma^2 \delta(\mathcal{T}_{\boldsymbol{\mu}})$ was proved by [27] and it is a direct consequence of (2.6) with $\boldsymbol{u} = \boldsymbol{\mu}$. To bound from above the statistical dimension of $\mathcal{T}_{\boldsymbol{\mu}}$, we have the inclusion

$$\mathcal{T}_{\boldsymbol{\mu}} \subset \mathcal{S}_{|T_1|}^{\cup} \times \cdots \times \mathcal{S}_{|T_{q(\boldsymbol{\mu})}|}^{\cup},$$

where $(T_1, \ldots, T_{q(\boldsymbol{\mu})})$ is a partition of $\{1, \ldots, n\}$ such that $\boldsymbol{\mu}$ is affine on each T_j, $j = 1, \ldots, q(\boldsymbol{\mu})$. The formula for the statistical dimension of a direct product of cones (cf. [1], Proposition 3.1, or the discussion preceding (3.8)) yields

$$\delta\big(\mathcal{S}_{|T_1|}^{\cup} \times \cdots \times \mathcal{S}_{|T_{q(\boldsymbol{\mu})}|}^{\cup}\big) = \sum_{j=1}^{q(\boldsymbol{\mu})} \delta\big(\mathcal{S}_{|T_j|}^{\cup}\big) \le 8 \sum_{j=1}^{q(\boldsymbol{\mu})} \log\big(e|T_j|\big) \le 8 \log\big(en/q(\boldsymbol{\mu})\big),$$

where we used (5.2) and Jensen's inequality. □

We now turn to the construction of confidence sets. Recall that if $\boldsymbol{u} \in \mathcal{S}_n^{\cup}$ is a convex sequence, $q(\boldsymbol{u})$ is the number of pieces in the piecewise affine decomposition of $\boldsymbol{u}$. Let $\hat{q} := q(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}))$ be the number of affine pieces of the Least Squares estimator. Then, define the statistic

$$\hat{r}_{\cup} = \frac{\sigma^2 \hat{q}(16 + 24\log(n) + 4\log(1/\alpha))}{n}. \tag{5.3}$$

Similarly to the case of the statistic $\hat{r}_{\uparrow}$ in isotonic regression, the following result shows that the confidence ball (2.7) with $\hat{r} = \hat{r}_{\cup}$ enjoys uniform coverage over all $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$.

Theorem 5.2. *For all $\alpha \in (0, 1)$ and all $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$, the statistic $\hat{r}_{\cup}$ defined in (4.1) satisfies*

$$\big\| \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\cup}\big) - \boldsymbol{\mu} \big\|_n^2 \le \hat{r}_{\cup}, \tag{5.4}$$

with probability at least $1 - \alpha$.

The above result is analogous to Theorem 4.1. The numerical constants are slightly worse in the case of the present section because the upper bound (5.2) on the statistical dimension of the cone $\mathcal{S}_n^{\cup}$ is slightly worse than (3.12). The proof of Theorem 5.2 is similar to the proof of Theorem 4.1 and can be found in Appendix C.

Now, the goal is to show that the statistic $\hat{r}_{\cup}$ is of the same order as the minimax quantity (5.1). We employ a different strategy than in the previous section. For any function $g : \mathbb{R}^n \to \mathbb{R}^n$ which is weakly differentiable, the divergence of g is the random variable

$$D_g(\mathbf{y}) = \sigma^2 \sum_{i=1}^{n} \frac{\partial}{\partial y_i} g(\mathbf{y})_i.$$

It is well known that by Stein's identity, under suitable conditions on g (cf. [25], Section 2, or [29], Lemma 3.6), we have

$$\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} D_g(\mathbf{y}) = \mathbb{E}_{\boldsymbol{\mu}}\big[\boldsymbol{\xi}^T g(\mathbf{y})\big]. \tag{5.5}$$

The divergence of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) = \Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y})$ is given in [14], Proposition 2.7, (see also [25]). Namely, we have the following result.

Proposition 5.3 ([14,25]). *If $g(\cdot) = \Pi_{\mathcal{S}_n^{\cup}}(\cdot)$ is the projection onto the cone of convex sequences, then (5.5) holds and we have*

$$D_g(\mathbf{y}) = \hat{q} + 1$$

almost surely, where $\hat{q} = q(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}))$.

This result can be used to bound from above the expected radius of the statistic $\hat{r}_{\cup}$.

Theorem 5.4. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{q}] \le 8q(\boldsymbol{\mu}) \log \frac{en}{q(\boldsymbol{\mu})} - 1. \tag{5.6}$$

Consequently, for all $\alpha \in (0,1)$, the statistic (5.3) satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{r}_{\cup}] \le \frac{\sigma^2 q(\boldsymbol{\mu}) \operatorname{polylog}(n, 1/\alpha)}{n}. \tag{5.7}$$

where $\operatorname{polylog}(n, 1/\alpha) = 8\log(en)(16 + 24\log(n) + 4\log(1/\alpha))$.

Proof. By Proposition 5.3 and (5.5), we have

$$\sigma^2 \mathbb{E}_{\boldsymbol{\mu}}[1 + \hat{q}] = \mathbb{E}_{\boldsymbol{\mu}}\big[\boldsymbol{\xi}^T \Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y})\big] = \mathbb{E}_{\boldsymbol{\mu}}\big[\boldsymbol{\xi}^T \big(\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}\big)\big]. \tag{5.8}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\sigma^2 \mathbb{E}_{\boldsymbol{\mu}}[1 + \hat{q}] &\le \mathbb{E}_{\boldsymbol{\mu}}^{1/2}\left[\left(\frac{\boldsymbol{\xi}^T (\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu})}{|\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}|_2}\right)^2\right] \mathbb{E}_{\boldsymbol{\mu}}^{1/2} \big|\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}\big|_2^2, \\
&\le \sigma \sqrt{\delta(\mathcal{T}_{\boldsymbol{\mu}})} \mathbb{E}_{\boldsymbol{\mu}}^{1/2} \big|\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}\big|_2^2.
\end{aligned}$$

Using Proposition 5.1 completes the proof of (5.6). Inequality (5.7) is a direct consequence of (5.6) and of the definition of $\hat{r}_{\cup}$. □

The above result is different from Theorem 4.2 in isotonic regression. Theorem 4.2 controls both the expectation and the deviations of $\hat{k}$. In this section, Theorem 5.4 only controls the expectation of $\hat{q}$. This comes from the use of Stein's identity in the proof of Theorem 5.4, which yields a result only in expectation.

The arguments used to prove Theorem 4.2 are based on the concentration properties of the intrinsic volumes of cones, while the proof of Theorem 5.4 relies on Stein's identity and Proposition 5.3. Thus, we have presented two methods to bound from above the expected diameter of the confidence sets constructed in the present paper. The concentration inequalities used in Theorem 4.2 for non-decreasing sequences provides exponential deviation inequalities, which are much stronger than the

bounds in expectation obtained in Theorem 5.4 (Indeed, bounds in expectation only imply weak deviation bounds by Markov's inequality). On the other hand, the argument based on Stein's formula in Theorem 5.4 has wider applicability. For instance, it readily applies to $\hat{k}$ from Section 4: Since $\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}]\sigma^2 = \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu})]$ holds [25], we have

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \le \mathbb{E}\big[\big|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\uparrow}\big) - \boldsymbol{\mu}\big|_2^2\big]^{1/2} \sigma\delta(\mathcal{T}_{\boldsymbol{\mu}})^{1/2} \le \sigma^2 k(\boldsymbol{\mu}) \log\big(en/k(\boldsymbol{\mu})\big)$$

for $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ by the same argument as in (5.8) for the first inequality and [3], Theorem 3.2, for the second, $\mathcal{T}_{\boldsymbol{\mu}} = \{\boldsymbol{u} - t\boldsymbol{\mu}, (t, \boldsymbol{u}) \in \mathbb{R} \times \mathcal{S}_n^{\uparrow}\}$ is the tangent cone. This provides an alternative proof of (4.8).

We end this section with the following consequence of Theorems 5.2 and 5.4.

Corollary 5.5. *Let $J = \{1, \dots, n-1\}$ and define the collection of models $(E_k)_{k\in J} = (\mathcal{S}_n^{\cup}(k))_{k=1,\dots,n-1}$. The random variable $\hat{r}_{\cup}$ defined in (5.3) satisfies (5.4) and (2.9) with $\hat{r}$ replaced by $\hat{r}_{\cup}$. Thus, the ball centered at $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ of radius $\sqrt{\hat{r}_{\cup}}$ is an honest confidence set, which is adaptive in expectation with respect to the models $(\mathcal{S}_n^{\cup}(k))_{k=1,\dots,n-1}$.*

The proof follows by combining Theorems 5.2 and 5.4, similarly to the proof of Corollary 4.3 for non-decreasing sequences.

6. Non-convexity: Adaptive confidence sets for unimodal sequences

Let $m \in \{1, \dots, n\}$. A sequence $\boldsymbol{u} \in \mathbb{R}^n$ is unimodal with mode at position m if and only if $\boldsymbol{u}_{\{1,\dots,m\}}$ is non-increasing and $\boldsymbol{u}_{\{m,\dots,n\}}$ is nondecreasing–in other words, $\boldsymbol{u}$ belongs to the set

$$K_m := \big\{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : u_1 \ge \cdots \ge u_m \le u_{m-1} \le \cdots \le u_n\big\}.$$

Next, by taking the union over all possible locations for the mode, we define the set of all unimodal sequences as

$$\mathcal{U} := \bigcup_{m=1,\dots,n} K_m.$$

The set $\mathcal{U}$ is non-convex for $n \ge 3$.

Recall that if $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow} \cup \mathcal{S}_n^{\downarrow}$ is monotone, $k(\boldsymbol{u}) = |\{u_i, i = 1, \dots, n\}$ is the smallest positive integer k such that $\boldsymbol{u}$ is piecewise constant constant with k pieces. We extend the function k to the set of unimodal sequences by setting

$$\forall \boldsymbol{u} \in \mathcal{U}, \quad k(\boldsymbol{u}) = \min\big\{k \ge 1 : \exists \text{ partition } (T_1, \dots, T_k) : \boldsymbol{u}_{T_j} \text{ is constant for all } j = 1, \dots, k\big\},$$

where the partition $(T_1, \dots, T_k)$ is a partition of $\{1, \dots, n\}$ with $\max T_j < \min T_{j+1}$ for all $j = 1, \dots, k-1$. A unimodal sequence $\boldsymbol{u} \in \mathcal{U}$ has $k(\boldsymbol{u}) - 1$ jumps. Similarly to isotonic regression, define the models

$$\mathcal{U}(k) = \big\{\boldsymbol{u} \in \mathcal{U} : k(\boldsymbol{u}) \le k\big\}, \quad k = 1, \dots, n. \tag{6.1}$$

These sets define models of increasing complexity since

$$\mathcal{U}(1) \subset \mathcal{U}(2) \subset \cdots \subset \mathcal{U}(n) = \mathcal{U}$$

It is known that there exist numerical constants c, c', c'' such that for all $\alpha \le c$,

$$c'\sigma^2 k/n \le R^*_\alpha\big(\mathcal{U}(k)\big) \le c''\sigma^2\big(k\log(en/k) + \log(n/\alpha)\big)/n. \tag{6.2}$$

Indeed, the lower bound is a consequence of (3.3) and the inclusion $R^*_\alpha(\mathcal{S}^\uparrow_n(k)) \subset R^*_\alpha(\mathcal{U}(k))$, while the upper bound is proved in [3,13,16], see [3], Appendix C, for explicit constants. Thus, for $\alpha > 0$ small enough, the quantity $R^*_\alpha(\mathcal{S}^\uparrow_n(k))$ is of order $k\sigma^2/n$, up to logarithmic factors in n and $1/\alpha$. Similarly, the minimax risk over the sets $\mathcal{U}(k)$ satisfies

$$c'\sigma^2 k/n \le R^*_{\mathbb{E}}\big(\mathcal{U}(k)\big) \le \sup_{\boldsymbol{\mu}\in\mathcal{U}(k)} \mathbb{E}_{\boldsymbol{\mu}}\big\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) - \boldsymbol{\mu}\big\|_n^2 \le c''\sigma^2 k\log(en/k)/n, \tag{6.3}$$

for some numerical constants $c', c'' > 0$ for instance by integrating the bounds of [3], Appendix C, where $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) \in \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{U}} \|\boldsymbol{u} - \mathbf{y}\|$ is the non-convex unimodal Least-Squares estimator (if the non-convex optimization problem has several solutions, we break ties arbitrarily).

The results below show that, using the number of constant pieces $\hat{k} = p(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}))$ of the unimodal Least-Squares estimator, one can construct adaptive confidence sets with respect to the collection of models (6.1).

Theorem 6.1. *Let $\hat{p} = k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}))$ be the smallest integer such that $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U})$ is piecewise constant on $\hat{p}$ contiguous pieces. For all $\alpha \in (0,1)$ and all $\mu \in \mathcal{U}$, the statistic*

$$\hat{r}_{uni} := \frac{4\sigma^2}{n}\left(\sqrt{(\hat{p}+3)\log\left(\frac{en}{\hat{p}+3}\right)} + \sqrt{2\log\left(\frac{n^2}{\alpha}\right) + 2(\hat{p}+3)\log\left(\frac{en}{\hat{p}+3}\right)}\right)^2 \tag{6.4}$$

satisfies

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\big\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) - \boldsymbol{\mu}\big\|_n^2 \le \hat{r}_{uni}\big) \ge 1-\alpha. \tag{6.5}$$

Proof. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U})$ for brevity. By definition of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U})$, inequality $|\hat{\boldsymbol{\mu}} - \mathbf{y}|_2^2 \le |\boldsymbol{\mu} - \mathbf{y}|_2^2$ holds. Hence, $|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \le 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ and

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 \le 2\boldsymbol{\xi}^T\hat{\boldsymbol{\theta}} \quad \text{where } \hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})/|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2. \tag{6.6}$$

We split $\{1,\dots,n\}$ into a partition (L, C, R) as follows: L is the largest set of indices of the form $\{1,2,\dots,|L|\}$ such that both $\hat{\boldsymbol{\mu}}_L$ and $\boldsymbol{\mu}_L$ are non-increasing, R is the largest set of indices of the form $\{n-|R|+1, n-|R|+2,\dots,n\}$ such that $\hat{\boldsymbol{\mu}}_R$ and $\boldsymbol{\mu}_R$ are both non-decreasing, and $C = \{1,\dots,n\} \setminus (R \cup L)$ contains the remaining central indices, where by definition, $\hat{\boldsymbol{\mu}}_C - \boldsymbol{\mu}_C$ is either increasing or decreasing. Let $\hat{k}_L$ be the number of constant pieces of $\hat{\boldsymbol{\mu}}$ on L, and let $(\hat{S}_1,\dots\hat{S}_{\hat{k}_L})$ be a partition of L such that $\hat{\boldsymbol{\mu}}_L$ is constant on each $\hat{S}_j$. Similarly, let $\hat{k}_R$ be the number of constant pieces of $\hat{\boldsymbol{\mu}}$ on R and $(\hat{T}_1,\dots,\hat{T}_{\hat{k}_R})$ be a partition of R such that $\hat{\boldsymbol{\mu}}_R$ is constant on each $\hat{T}_j$. Then both $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}$ and $\hat{\boldsymbol{\theta}}$ belong to either one of the cones

$$\hat{\mathcal{D}}_+ := \big(\mathcal{S}^\uparrow_{|\hat{S}_1|} \times\cdots\times \mathcal{S}^\uparrow_{|\hat{S}_{\hat{k}_L}|}\big) \times \mathcal{S}^\uparrow_{|C|} \times \big(\mathcal{S}^\downarrow_{|\hat{T}_1|} \times\cdots\times \mathcal{S}^\downarrow_{|\hat{T}_{\hat{k}_R}|}\big).$$

or

$$\hat{\mathcal{D}}_- := \big(\mathcal{S}^\uparrow_{|\hat{S}_1|} \times\cdots\times \mathcal{S}^\uparrow_{|\hat{S}_{\hat{k}_L}|}\big) \times \mathcal{S}^\downarrow_{|C|} \times \big(\mathcal{S}^\downarrow_{|\hat{T}_1|} \times\cdots\times \mathcal{S}^\downarrow_{|\hat{T}_{\hat{k}_R}|}\big),$$

the only difference being the direction on the central indices in C. We now argue similarly to (4.6) as follows. If $\mathcal{C}$ is a nonrandom cone of the form

$$\mathcal{C} = \mathcal{S}^{\uparrow}_{n_1} \times \cdots \times \mathcal{S}^{\uparrow}_{n_k} \times \mathcal{S}^{\downarrow}_{m_1} \times \cdots \times \mathcal{S}^{\downarrow}_{m_l} \tag{6.7}$$

for positive integers $n_1, \ldots, n_k, m_1, \ldots, m_l$ such that $n_1 + \cdots + n_k + m_1 + \cdots + m_l = n$, then by the Gaussian concentration theorem, (4.5) holds with probability at least $1 - e^{-x}$. Furthermore, $\mathbb{E} \sup_{\boldsymbol{\theta} \in \mathcal{C}:|\boldsymbol{\theta}|_2=1} \boldsymbol{\xi}^T \boldsymbol{\theta} \le \sigma \delta(\mathcal{C})^{1/2}$ by Jensen's inequality, and $\delta(\mathcal{C}) = \sum_{j=1}^{k} \log(en_j) + \sum_{j=1}^{l} \log(em_j) \le (k+l)\log(en/(k+l))$ thanks to (3.12) and the fact that the statistical dimension of a product of cones is equal to the sum of the statistical dimensions (cf. [1], Proposition 3.1, or the discussion preceding (3.8)).

Let $s = 1, \ldots, n$ be fixed. There are fewer than $n\binom{n-1}{s}$ cones of the form (6.7) with $k + l = s$. By the union bound and inequality $\log\binom{n-1}{s} \le s \log(en/s)$, we have with probability at least $1 - e^{-x}$ the bound

$$\sup_{\mathcal{C}} \Big(\sup_{\boldsymbol{\theta} \in \mathcal{C}:|\boldsymbol{\theta}|_2=1} \boldsymbol{\xi}^T \boldsymbol{\theta} \Big) \le \sigma \sqrt{s \log(en/s)} + \sigma \sqrt{2\big(x + \log n + s \log(en/s)\big)}$$

where the supremum is taken over all cones $\mathcal{C}$ of the form (6.7) with $k + l = s$. Finally, we apply the union bound over all $s \in \{1, \ldots, n\}$ and set $x = \log(n/\alpha)$. We have established that with probability at least $1 - \alpha$,

$$\sup_{\boldsymbol{\theta} \in \hat{\mathcal{D}}_+ \cup \hat{\mathcal{D}}_- :|\boldsymbol{\theta}|_2=1} \boldsymbol{\xi}^T \boldsymbol{\theta} \le \sigma \sqrt{\hat{s} \log(en/\hat{s})} + \sigma \sqrt{2\big(\log\big(n^2/\alpha\big) + \hat{s} \log(en/\hat{s})\big)}.$$

where $\hat{s} = \hat{k}_L + \hat{k}_R + 1$. The proof is completed by combining this bound with (6.6), and observing that $\hat{s} = \hat{k}_R + 1 + \hat{k}_L \le \hat{p} + 3$. □

Finally, the following result shows that the confidence set of the previous theorem has optimal radius up to logarithm factors.

Theorem 6.2. *If $\boldsymbol{\mu} \in \mathcal{U}$, then*

$$\hat{p} := k\big(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U})\big) \le 4k(\boldsymbol{\mu}) \log\big(en/k(\boldsymbol{\mu})\big) + 14 \log(2n/\gamma)$$

with probability at least $1 - \gamma$. Furthermore, $\mathbb{E}[\hat{p}] \le 4k(\boldsymbol{\mu}) \log(en/k(\boldsymbol{\mu})) + 14 \log(2en)$.

Proof. If $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U})$ is a unimodal fit and $\hat{m}$ is the location of its mode, then the following facts hold true [28]: (a) the value of $\hat{\mu}_{\hat{m}}$ of $\hat{\boldsymbol{\mu}}$ at $\hat{m}$ is equal to $y_{\hat{m}}$, (b) $\hat{\boldsymbol{\mu}}_{\{1,\ldots,\hat{m}\}}$ is equal to the isotonic (decreasing) fit of $\mathbf{y}_{\{1,\ldots,\hat{m}\}}$, and (c) $\hat{\boldsymbol{\mu}}_{\{\hat{m},\ldots,n\}}$ is equal to the isotonic (increasing) fit of $\mathbf{y}_{\{\hat{m},\ldots,n\}}$. Hence, we can bound from above the number of constant pieces of $\hat{\boldsymbol{\mu}}$ by the number of constant pieces of two isotonic fits, one on $\{1, \ldots, \hat{m}\}$ and the other on $\{\hat{m}, \ldots, n\}$.

If m is a deterministic mode location, and $\boldsymbol{\mu}_{\{1,\ldots,m\}}$ has $k(\boldsymbol{\mu}_{\{1,\ldots,m\}})$ pieces, by Theorem 4.2 the isotonic (decreasing) fit of $\mathbf{y}_{\{1,\ldots,m}\}$ has at most

$$2k(\boldsymbol{\mu}_{\{1,\ldots,m\}}) \log\big(em/k(\boldsymbol{\mu}_{\{1,\ldots,m\}})\big) + 7 \log(1/\gamma)$$

constant pieces with probability $1 - \gamma$. Similarly, with at least probability $1 - \gamma$, the isotonic (increasing) fit of $\mathbf{y}_{\{m,\ldots,n\}}$ has at most

$$2k(\boldsymbol{\mu}_{\{m,\ldots,n\}}) \log\big(e(n-m+1)/k(\boldsymbol{\mu}_{\{m,\ldots,n\}})\big) + 7 \log(1/\gamma)$$

constant pieces with probability at least $1-\gamma$. By the union bound, the two previous sentences hold uniformly over all possible modes $m = 1, \dots, n$ with probability at least $1 - 2n\gamma$.

Hence, with probability at least $1-2n\gamma$, the number of constant pieces of the unimodal least-squares $\hat{\boldsymbol{\mu}}$ is bounded from above by

$$2k(\boldsymbol{\mu}_{\{1,\dots,\hat{m}\}}) \log\left(\frac{e\hat{m}}{k(\boldsymbol{\mu}_{\{1,\dots,\hat{m}\}})}\right) + 2k(\boldsymbol{\mu}_{\{\hat{m},\dots,n\}}) \log\left(\frac{e(n-\hat{m}+1)}{k(\boldsymbol{\mu}_{\{\hat{m},\dots,n\}})}\right) + 14 \log\frac{1}{\gamma}.$$

We first use that $\hat{m} \le n$ and $(n-\hat{m}+1) \le n$ to bound from above the numerator inside the logarithms. Next, we use $k(\boldsymbol{\mu}_{\{1,\dots,\hat{m}\}}) \le k(\boldsymbol{\mu})$ and the fact that $x \log(en/x)$ is increasing on $[1, n]$ to conclude that the previous display is bounded from above by $4k(\boldsymbol{\mu}) \log(en/k(\boldsymbol{\mu})) + 14 \log(1/\gamma)$. The result in expectation is obtained by integration, using the identity $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z > t)\, dt$ for every non-negative random variable Z. □

Corollary 6.3. *Let $J = \{1, \dots, n\}$ and define the collection of models $(E_k)_{k\in J} = (\mathcal{U}(k))_{k=1,\dots,n}$. The random variable $\hat{r}_{uni}$ defined in (6.4) satisfies (6.5), (2.8) and (2.9) with $\hat{r}$ replaced by $\hat{r}_{uni}$. Thus, the ball centered at $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ of radius $\sqrt{\hat{r}_{uni}}$ is an honest confidence set, which is adaptive in expectation with respect to the models $(\mathcal{U}(k))_{k=1,\dots,n}$.*

The proof follows by combining Theorems 6.1 and 6.2, similarly to the proof of Corollary 4.3 for non-decreasing sequences.

7. Concluding remarks

We have provided a simple construction of honest and adaptive confidence sets for isotonic, convex and unimodal regression. Our construction reveals that the complexity of the Least-Squares estimator in these problems, e.g. the number of jumps of the isotonic Least-Squares or the number of changes of slope of the convex Least-Squares, can be used to bound from above the error of the estimator (cf. Theorems 4.1, 5.2 and 6.1). Furthermore, the complexity of the Least-Squares estimator is these problems is not larger, up to logarithmic factors, than the complexity of the true mean vector (cf. Theorems 4.2, 5.4 and 6.2).

The construction of honest confidence sets in Theorems 4.1, 5.2 and 6.1 relies on a careful application of the Gaussian concentration theorem combined with union bounds and upper bounds on statistical dimensions of tangent cones. Such techniques can readily be extended to the setting of [12, 18,21], where bounds on the statistical dimensions of tangent cones are readily available. However, the techniques used in Theorems 4.2, 5.4 and 6.2 to control the size of such confidence sets do not directly extend to the settings considered in [12,18,21] and it is unclear at this point how to control adaptively the radius of the confidence sets in these settings.

For the piecewise constant signals without the monotonicity constraint studied in Section 4.2, construction of honest and adaptive confidence sets are impossible. This provides additional evidence that shape constraints make it possible to construct adaptive and honest confidence sets or confidence bands for models where such task is impossible in the absence of the shape constraint [6,7,15].

The present paper tackles shape constraint regression problems with fixed design: if $x_1 < \cdots < x_n$ are ordered design points then the model $\mathcal{S}_n^{\uparrow}$ of nondecreasing sequences is equal to $\{(f(x_1), \dots, f(x_n))^T, f : \mathbb{R} \to \mathbb{R} \text{ nondecreasing}\}$, Hence, the results of the paper readily extend to random design isotonic regression, provided that the loss is measured with the empirical loss $\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2$ where f^* is the true regression function and $x_1 < \cdots < x_n$ are the random design points, by setting $\mu_i = f^*(x_i)$. For convex regression, if $x_1 < \cdots < x_n$ are equispaced

(i.e., $|x_{i+1} - x_i|$ is the same for all i), the model $\mathcal{S}_n^{\cup}$ is the same as $\{(f(x_1), \dots, f(x_n))^T, f : \mathbb{R} \to \mathbb{R}$ convex$\}$. If the design points are not equispaced, then $\{(f(x_1), \dots, f(x_n))^T, f : \mathbb{R} \to \mathbb{R}$ convex$\}$ defines a model different than $\mathcal{S}_n^{\cup}$, but the same logarithmic bounds on the statistical dimension hold for non-equispaced design points [3], cf. Proposition 4.2, and consequently the results of Section 5 extend to non-equispaced designs.

Regarding random design, in order to extend the results of the present paper to i.i.d. real valued design points $x_1, \dots, x_n$, one would need to establish a relationship between the empirical loss $\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2$ and the population loss $\mathbb{E}_{x\sim F}[(\hat{f}(x) - f^*(x))^2 | \hat{f}]$ where F is the distribution of the design. The random design analysis in [21] is a good starting point for this purpose. Finally, let us mention that if the design is Gaussian and high-dimensional, upper bounds based on statistical dimensions can be obtained similarly as in the sequence model [23] and some techniques of the present paper may thus be applicable. We leave this research direction open for future work.

Appendix A: Nondecreasing sequences with bounded total variation

Let $V > 0$. If the unknown parameter $\boldsymbol{\mu}$ satisfies $\mu_n - \mu_1 \le V$, the risk of the Least Squares estimator satisfy [31], (28),

$$\mathbb{E}_{\boldsymbol{\mu}} \big\| \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\uparrow}\big) - \boldsymbol{\mu} \big\|_n^2 \le \sigma^2 \kappa^2 \bigg(\Big(\frac{V}{\sigma n} \Big)^{2/3} + \frac{\log(en)}{n} \bigg),$$

where $\kappa \le 3.6$. Thus, an explicit constant is readily available [31], (2.8).

In this section, we explain how to construct confidence sets with diameter of the same order as the right-hand side of the previous display. We proceed as follows.

The function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by $f(\boldsymbol{v}) = \| \Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu} + \sigma \boldsymbol{v}) - \boldsymbol{\mu} \|_n$ is Lipschitz with coefficient $\sigma/\sqrt{n}$ as for all $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^n$,

$$\begin{aligned} \big| f(\boldsymbol{v}) - f\big(\boldsymbol{v}'\big) \big| &\le \big\| \Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu} + \sigma \boldsymbol{v}) - \Pi_{\mathcal{S}_n^{\uparrow}}\big(\boldsymbol{\mu} + \sigma \boldsymbol{v}'\big) \big\|_n \\ &\le \sigma \big\| \boldsymbol{v} - \boldsymbol{v}' \big\|_n = (\sigma/\sqrt{n}) \big| \boldsymbol{v} - \boldsymbol{v}' \big|_2. \end{aligned} \tag{A.1}$$

By the Gaussian concentration inequality [5], Theorem 5.6, the following holds with probability greater than $1 - \alpha$

$$\big\| \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\uparrow}\big) - \boldsymbol{\mu} \big\|_n \le \mathbb{E}_{\boldsymbol{\mu}} \| \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\uparrow}\big) - \boldsymbol{\mu} \|_n + \sigma \sqrt{\frac{2\log(1/\alpha)}{n}}.$$

Using that $(a+b)^2 \le 2a^2 + 2b^2$, we obtain the following for all $\alpha \in (0, 1)$: If $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and $\mu_n - \mu_1 \le V$, then

$$\big\| \hat{\boldsymbol{\mu}}^{\mathrm{LS}}\big(\mathcal{S}_n^{\uparrow}\big) - \boldsymbol{\mu} \big\|_n^2 \le 2\kappa^2\sigma^2 \Big(\frac{V}{\sigma n} \Big)^{2/3} + \frac{2\kappa^2\sigma^2 \log(en) + 4\sigma^2 \log(1/\alpha)}{n}$$

with probability greater than $1 - \alpha$.

Let $V_{\boldsymbol{\mu}} = \mu_n - \mu_1$ and $\hat{V} = y_n - y_1$. The random variable $\hat{V} - V_{\boldsymbol{\mu}}$ is a centered Gaussian with variance $2\sigma^2$, so that

$$V_{\boldsymbol{\mu}} \le \hat{V} + 2\sigma \sqrt{\log(1/\alpha)}$$

with probability greater than $1 - \alpha$. Thus, we have established the following.

Proposition A.1. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$. Define the statistic $\hat{s}_{\uparrow}$ by*

$$\hat{s}_{\uparrow} = 2\kappa^2\sigma^2\left(\frac{\hat{V} + 2\sigma\sqrt{\log(1/\alpha)}}{\sigma n}\right)^{2/3} + \frac{2\kappa^2\sigma^2\log(en) + 4\sigma^2\log(1/\alpha)}{n}$$

where $\kappa \le 3.6$ is the constant from [31] *that appears in* (A.1)*. Then we have $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|_n^2 \le \hat{s}_{\uparrow}$ with probability greater than $1 - 2\alpha$.*

Furthermore, it is clear that $\hat{V} \le V_{\boldsymbol{\mu}} + 2\sigma\sqrt{\log(1/\gamma)}$ with probability greater than $1 - \gamma$ for all $\gamma \in (0, 1)$.

Proposition A.2. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and let $V = \mu_n - \mu_1$. Then the statistic $\hat{s}_{\uparrow}$ defined above satisfies*

$$\hat{s}_{\uparrow} \le 2\kappa^2\sigma^2\left(\frac{\hat{V} + 2\sigma\sqrt{\log(1/(\gamma\alpha))}}{\sigma n}\right)^{2/3} + \frac{2\kappa^2\sigma^2\log(en) + 4\sigma^2\log(1/\alpha)}{n} \tag{A.2}$$

with probability at least $1 - \gamma$ for all $\gamma \in (0, 1)$.

Theorem A.3. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$. The statistic $\min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$ satisfies*

$$\left\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}\left(\mathcal{S}_n^{\uparrow}\right) - \boldsymbol{\mu}\right\|_n^2 \le \min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$$

with probability at least $1 - 3\alpha$. Furthermore, for all $\gamma \in (0, 1)$, the statistic $\min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$ is bounded from above with probability at least $1 - 2\gamma$, by the minimum of the right-hand side of (4.7) *and the right-hand side of* (A.2)*.*

For all $V \ge \sigma$ and all $k = 1, \dots, n$, define the class

$$\mathcal{S}_n^{\uparrow}(k, V) := \left\{\boldsymbol{v} = (v_1, \dots, v_n)^T \in \mathcal{S}_n^{\uparrow} : k(\boldsymbol{v}) \le k \text{ and } v_n - v_1 \le V\right\}.$$

Since

$$c\sigma^2\min\left(\left(\frac{V}{\sigma n}\right)^{2/3}, \frac{k\log\log(16n/k)}{n}\right) \le R_{\mathbb{E}}^*\left(\mathcal{S}_n^{\uparrow}(k, V)\right)$$

for some absolute constant $c > 0$ by combining the lower bound in (3.2) with the lower bound in [31], (2.9). Thus, the statistic $\min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$ of Theorem A.3 provides an honest confidence ball centered at the Least Squares estimator, and this confidence ball is adaptive in expectation for the collection of models

$$\left(\mathcal{S}_n^{\uparrow}(k, V)\right)_{k\in\{1,\dots,n\}, V\in[\sigma,+\infty)}.$$

Appendix B: Technical lemma

Lemma B.1. *In the present lemma, all quantities are deterministic. Let $a \in [-\infty, +\infty)$ and $b \in (-\infty, +\infty]$ such that $a \le b$. Let $\mathbf{y} \in \mathbb{R}^n$. Define $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ as the unique solutions of the minimization problems*

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{v}\in\mathcal{S}_n^{\uparrow}} |\mathbf{y} - \boldsymbol{v}|_2^2, \tag{B.1}$$

$$\boldsymbol{\theta} \in \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{S}_n^{\uparrow}(a,b)} |\mathbf{y} - \boldsymbol{v}|_2^2 \tag{B.2}$$

where $\mathcal{S}_n^{\uparrow}(a,b) := \{\boldsymbol{v} = (v_1, \ldots, v_n)^T \in \mathcal{S}_n^{\uparrow} : a \le v_1, v_n \le b\}$. *Then* $k(\boldsymbol{\theta}) \le k(\boldsymbol{\theta}^*)$.

The intuition behind this lemma is that if a constraint is not saturated for $\boldsymbol{\theta}$, this constraint is not saturated for $\boldsymbol{\theta}^*$ either, so $\boldsymbol{\theta}^*$ has at least as many jumps as $\boldsymbol{\theta}$.

Proof of Lemma B.1. Let $T_a = \{i = 1, \ldots, n : \hat{\theta}_i^* \le a\}$, $T_b = \{i = 1, \ldots, n : \hat{\theta}_i^* \ge b\}$ and $T_c = \{i = 1, \ldots, n : a < \hat{\theta}_i^* < b\}$. We will prove that the unique minimizer $\boldsymbol{\theta}$ of the problem (B.2) is

$$\boldsymbol{\theta}_{T_a} = a\mathbf{1}_{T_a}, \qquad \boldsymbol{\theta}_{T_c} = \boldsymbol{\theta}^*_{T_c}, \qquad \boldsymbol{\theta}_{T_b} = b\mathbf{1}_{T_b}, \tag{B.3}$$

where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$. Then it is clear that $k(\boldsymbol{\theta}) = 1 + k(\boldsymbol{\theta}^*_{T_c}) + 1 \le k(\boldsymbol{\theta}^*_{T_a}) + k(\boldsymbol{\theta}^*_{T_c}) + k(\boldsymbol{\theta}^*_{T_b}) = k(\boldsymbol{\theta}^*)$.

First, by strong convexity there exists a unique solution to the minimization problem (B.2), and a unique solution to the minimization problem (B.1). Second, by the characterization of the projection onto the closed convex set $\mathcal{S}_n^{\uparrow}(a,b)$, if $\boldsymbol{\theta}$ satisfies

$$A_{\boldsymbol{u}} := (\boldsymbol{u} - \boldsymbol{\theta})^T(\mathbf{y} - \boldsymbol{\theta}) \le 0$$

for all $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}(a,b)$, then $\boldsymbol{\theta}$ is the unique solution to the minimization problem (B.2). Let $\boldsymbol{\theta}$ be defined by (B.3). By simple algebra, for all $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}(a,b)$,

$$\begin{aligned} A_{\boldsymbol{u}} = {} & \big(\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a} + \boldsymbol{\theta}^*_{T_a} - \boldsymbol{\theta}^*_{T_a}\big)^T\big(\mathbf{y}_{T_a} - \boldsymbol{\theta}^*_{T_a}\big) + (\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a})^T\big(\boldsymbol{\theta}^*_{T_a} - a\mathbf{1}_{T_a}\big) \\ & + \big(\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b} + \boldsymbol{\theta}^*_{T_b} - \boldsymbol{\theta}^*_{T_b}\big)^T\big(\mathbf{y}_{T_b} - \boldsymbol{\theta}^*_{T_b}\big) + (\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b})^T\big(\boldsymbol{\theta}^*_{T_b} - b\mathbf{1}_{T_b}\big) \\ & + \big(\boldsymbol{u}_{T_c} - \boldsymbol{\theta}^*_{T_c}\big)^T\big(\mathbf{y}_{T_c} - \boldsymbol{\theta}^*_{T_c}\big). \end{aligned}$$

If a vector $\boldsymbol{v}$ has nonnegative entries and a vector $\boldsymbol{x}$ has non-positive entries, then $\boldsymbol{v}^T\boldsymbol{x} \le 0$, so $(\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a})^T(\boldsymbol{\theta}^*_{T_a} - a\mathbf{1}_{T_a}) \le 0$ and $(\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b})^T(\boldsymbol{\theta}^*_{T_b} - b\mathbf{1}_{T_b}) \le 0$. Thus,

$$\begin{aligned} A_{\boldsymbol{u}} \le {} & \big(\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a} + \boldsymbol{\theta}^*_{T_a} - \boldsymbol{\theta}^*_{T_a}\big)^T\big(\mathbf{y}_{T_a} - \boldsymbol{\theta}^*_{T_a}\big) \\ & + \big(\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b} + \boldsymbol{\theta}^*_{T_b} - \boldsymbol{\theta}^*_{T_b}\big)^T\big(\mathbf{y}_{T_b} - \boldsymbol{\theta}^*_{T_b}\big) \\ & + \big(\boldsymbol{u}_{T_c} - \boldsymbol{\theta}^*_{T_c}\big)^T\big(\mathbf{y}_{T_c} - \boldsymbol{\theta}^*_{T_c}\big), \end{aligned}$$

and the right-hand side of the previous display is equal to

$$\big(\boldsymbol{v} - \boldsymbol{\theta}^*\big)^T\big(\mathbf{y} - \boldsymbol{\theta}^*\big), \tag{B.4}$$

where $\boldsymbol{v} \in \mathbb{R}^n$ is defined by

$$\boldsymbol{v}_{T_a} := \boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a} + \boldsymbol{\theta}^*_{T_a}, \qquad \boldsymbol{v}_{T_c} := \boldsymbol{u}_{T_c}, \qquad \boldsymbol{v}_{T_b} := \boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b} + \boldsymbol{\theta}^*_{T_b}.$$

We have $\boldsymbol{v} \in \mathcal{S}_n^{\uparrow}$ by definition of T_a, T_c and T_b. The quantity (B.4) is non-positive because $\boldsymbol{\theta}^*$ is the projection of $\mathbf{y}$ onto the convex set $\mathcal{S}_n^{\uparrow}$. Thus, we have established that $A_{\boldsymbol{u}} \le 0$ for all $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}(a,b)$, so that the unique solution of the minimization problem (B.2) is given by the expression (B.3). □

Appendix C: Proofs for convex sequences

Proof of Theorem 5.2. For any set T of the form (4.3), using the concentration property (3.7) of the random variable (3.6) with $K = \mathcal{S}^{\cap}_{|T|}$, we have with probability greater than $1-\alpha$,

$$\big|\Pi_{\mathcal{S}^{\cap}_{|T|}}(\boldsymbol{\xi}_T)\big|_2^2 \le 2\delta\big(\mathcal{S}^{\cap}_{|T|}\big) + 4\log(1/\alpha) \le 16\log(en) + 4\log(1/\alpha),$$

where we used (5.2) for the last inequality. There are less that n^2 sets $T \subset \{1,\dots,n\}$ of the form (4.3). Using the union bound for all sets T of the form (4.3), we have $\mathbb{P}(\Omega(\alpha)) \ge 1-\alpha$ where

$$\Omega(\alpha) := \left\{ \sup_{T\subset\{1,\dots,n\} \text{ of the form (4.3)}} \big|\Pi_{\mathcal{S}^{\cap}_{|T|}}(\boldsymbol{\xi}_T)\big|_2^2 \le \sigma^2\left(16\log(en) + 4\log\left(\frac{n^2}{\alpha}\right)\right)\right\}.$$

Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\cup}_n)$ for notational simplicity. Then (2.6) with $\boldsymbol{u}$ replaced by $\boldsymbol{\mu}$ can be rewritten as

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \le 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2.$$

By definition of $q(\cdot)$, there exists a partition $(\hat{T}_1,\dots,\hat{T}_{\hat{q}})$ of $\{1,\dots,n\}$ such that $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\cup}_n)$ is affine on each $\hat{T}_j$, $j=1,\dots,\hat{q}$. Furthermore, each $\hat{T}_j$ has the form (4.3) because $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^{\cup}_n) \in \mathcal{S}^{\cup}_n$. We have

$$\begin{aligned} 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 &= \sum_{j=1}^{\hat{q}} 2\boldsymbol{\xi}_{\hat{T}_j}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j} - \big|(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{T_j}\big|_2^2, \\ &\le \sum_{j=1}^{\hat{q}} \left(\frac{\boldsymbol{\xi}_{\hat{T}_j}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j}}{|(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j}|_2}\right)^2, \end{aligned}$$

where we have used $2ab - a^2 \le b^2$. By definition of $(\hat{T}_1,\dots,\hat{T}_{\hat{q}})$, $\hat{\boldsymbol{\mu}}$ is affine on $\hat{T}_j$ for each $j = 1,\dots,\hat{q}$, thus the vector $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j} \in \mathcal{S}^{\cap}_{|\hat{T}_j|}$ is a concave sequence. By taking the supremum, we obtain

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \le \sum_{j=1}^{\hat{q}} \sup_{\boldsymbol{v}\in\mathcal{S}^{\cap}_{|\hat{T}_j|}:|\boldsymbol{v}|_2^2\le 1} \big(\boldsymbol{\xi}_{\hat{T}_j}^T\boldsymbol{v}\big)^2 = \sum_{j=1}^{\hat{q}} \big|\Pi_{\mathcal{S}^{\cap}_{|\hat{T}_j|}}(\boldsymbol{\xi}_{|\hat{T}_j|})\big|_2^2,$$

where we used (3.6) for the last equality. On the event $\Omega(\alpha)$ and by definition of $\hat{r}_{\cup}$,

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \le \sum_{j=1}^{\hat{q}} \big|\Pi_{\mathcal{S}^{\cap}_{|\hat{T}_j|}}(\boldsymbol{\xi}_{|\hat{T}_j|})\big|_2^2 \le \sigma^2\hat{q}\big(16\log(en) + 4\log\big(n^2/\alpha\big)\big) = n\hat{r}_{\cup}.$$

□

Acknowledgements

This work was partially supported by GENES and by the ANR grant ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047, as well as the NSF grants DMS-1811976 and DMS-1945428.

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. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

[2] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

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

[4] Bellec, P.C. and Tsybakov, A.B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.* **16** 1879–1892. MR3417801

[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[6] Cai, T.T. and Low, M.G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240 https://doi.org/10.1214/009053606000000146

[7] Cai, T.T., Low, M.G. and Xia, Y. (2013). Adaptive confidence intervals for regression functions under shape constraints. *Ann. Statist.* **41** 722–750. MR3099119 https://doi.org/10.1214/12-AOS1068

[8] Carpentier, A. and Verzelen, N. (2019). Optimal sparsity testing in linear regression model. ArXiv preprint. Available at arXiv:1901.08802.

[9] Carpentier, A. and Verzelen, N. (2019). Adaptive estimation of the sparsity in the Gaussian vector model. *Ann. Statist.* **47** 93–126. MR3909928 https://doi.org/10.1214/17-AOS1680

[10] Chandrasekaran, V., Recht, B., Parrilo, P.A. and Willsky, A.S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474 https://doi.org/10.1007/s10208-012-9135-7

[11] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878 https://doi.org/10.1214/15-AOS1324

[12] Chatterjee, S., Guntuboyina, A. and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. MR3706788 https://doi.org/10.3150/16-BEJ865

[13] Chatterjee, S. and Lafferty, J. (2019). Adaptive risk bounds in unimodal regression. *Bernoulli* **25** 1–25. MR3892309 https://doi.org/10.3150/16-bej922

[14] Chen, X., Lin, Q. and Sen, B. (2020). On Degrees of Freedom of Projection Estimators With Applications to Multivariate Nonparametric Regression. *J. Amer. Statist. Assoc.* **115** 173–186. MR4078455 https://doi.org/10.1080/01621459.2018.1537917

[15] Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli* **9** 423–449. MR1997491 https://doi.org/10.3150/bj/1065444812

[16] Flammarion, N., Mao, C. and Rigollet, P. (2019). Optimal rates of statistical seriation. *Bernoulli* **25** 623–653. MR3892331 https://doi.org/10.3150/17-bej1000

[17] Gao, C., Han, F. and Zhang, C.-H. (2017). Minimax risk bounds for piecewise constant models. ArXiv preprint. Available at arXiv:1705.06386.

[18] Guntuboyina, A., Lieu, D., Chatterjee, S. and Sen, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** 205–229. MR4065159 https://doi.org/10.1214/18-AOS1799

[19] Guntuboyina, A. and Sen, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. MR3405621 https://doi.org/10.1007/s00440-014-0595-3

[20] Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statist. Sci.* **33** 568–594. MR3881209 https://doi.org/10.1214/18-STS665

[21] Han, Q., Wang, T., Chatterjee, S. and Samworth, R.J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[22] Ingster, Y.I., Tsybakov, A.B. and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131 https://doi.org/10.1214/10-EJS589

[23] Matey, N. (2019). Gaussian regression with convex constraints. In *Proceedings of Machine Learning Research* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 31–38. Available at http://proceedings.mlr.press/v89/neykov19b.html.

[24] McCoy, M.B. and Tropp, J.A. (2014). From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete Comput. Geom.* **51** 926–963. MR3216671 https://doi.org/10.1007/s00454-014-9595-4

[25] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. MR1810920 https://doi.org/10.1214/aos/1015956708

[26] Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170

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

[28] Stout, Q.F. (2008). Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.* **53** 289–297. MR2649085 https://doi.org/10.1016/j.csda.2008.08.005

[29] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794

[30] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672 https://doi.org/10.1214/12-EJS666

[31] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898 https://doi.org/10.1214/aos/1021379864

Received April 2019 and revised March 2020