Electronic Journal of Statistics
Vol. 14 (2020) 1508–1576
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1700

Estimating piecewise monotone signals

Kentaro Minami

The University of Tokyo / Preferred Networks, Inc.
e-mail: minami@preferred.jp

Abstract: We study the problem of estimating piecewise monotone vectors. This problem can be seen as a generalization of the isotonic regression that allows a small number of order-violating changepoints. We focus mainly on the performance of the nearly-isotonic regression proposed by Tibshirani et al. (2011). We derive risk bounds for the nearly-isotonic regression estimators that are adaptive to piecewise monotone signals. The estimator achieves a near minimax convergence rate over certain classes of piecewise monotone signals under a weak assumption. Furthermore, we present an algorithm that can be applied to the nearly-isotonic type estimators on general weighted graphs. The simulation results suggest that the nearly-isotonic regression performs as well as the ideal estimator that knows the true positions of changepoints.

Keywords and phrases: Piecewise monotone function, isotonic regression, nearly-isotonic regression, adaptive risk bounds.

Received May 2019.

Contents

1 Introduction 1509
1.1 Summary of theoretical results 1512
1.2 Organization 1513
1.3 Notation 1513
2 Related work 1514
3 Lower bounds 1516
3.1 Minimax lower bound 1516
3.2 Lower bound of isotonic regression with misspecified partitions 1517
4 Risk bounds for nearly-isotonic regression 1518
4.1 Risk bounds for constrained estimators 1518
4.2 Risk bounds for penalized estimators 1522
4.3 Application to piecewise monotone vectors 1524
5 Model selection based estimators 1525
6 Simulations 1527
6.1 Dealing with inconsistency at boundaries 1527
6.2 Simulation data 1528
6.3 Geological data 1530
7 Discussion 1531
7.1 Non-Gaussian noises 1532
7.2 Future directions 1532
A Algorithms for nearly-isotonic estimators 1533

A.1 Penalized estimators 1533
A.1.1 One-dimensional problem 1533
A.1.2 General graphs 1535
A.1.3 General convex loss functions 1535
A.2 Constrained estimators 1537
B Supplemental experiments 1538
C Proofs in Section 3 1539
C.1 Proof of Proposition 3.2 1539
C.2 Proof of Proposition 3.3 1540
D Proofs in Section 4 1541
D.1 Preliminaries 1541
D.2 Risk bounds for constrained estimators (Proof of Theorem 4.1) 1545
D.3 Proof of Corollary 4.4 1556
D.4 Risk bounds for penalized estimators (Proof of Theorem 4.7) . 1557
D.5 Proof of Corollary 4.12 1560
D.6 Subdifferential and weak decomposability 1562
D.6.1 Characterization of the subdifferential 1562
D.6.2 Weak decomposability 1564
E Proofs in Section 5 1568
E.1 Proof overview 1568
E.2 Controlling the normalized process 1569
E.3 Proof of Theorem 5.1 1572
F Auxiliary lemmas 1573
Acknowledgment 1573
References 1574

1. Introduction

Isotonic regression is a popular statistical method based on partial order structures, which has a long history in statistics (Ayer et al. 1955, Brunk 1955, van Eeden 1956). Suppose that $\theta^* \in \mathbb{R}^n$ is a monotone vector satisfying $\theta_1^* \le \theta_2^* \le \cdots \le \theta_n^*$, and y is a noisy observation of θ^*. The goal of the isotonic regression is to find a least-square fit under the monotone constraint:

$$\text{minimize } \|y-\theta\|_2 \quad \text{subject to } \theta_1 \le \theta_2 \le \cdots \le \theta_n. \tag{1}$$

In other words, the isotonic regression is the least squares estimator $\hat{\theta} = \hat{\theta}_{K_n^\uparrow}$ over a closed convex cone $K_n^\uparrow := \{\theta \in \mathbb{R}^n : \theta_1 \le \theta_2 \le \cdots \le \theta_n\}$. Broadly speaking, the isotonic regression is an example of *shape restricted regression*. For comprehensive reviews on this field, see Robertson et al. (1988), Groeneboom and Jongbloed (2014), Chatterjee et al. (2015), Guntuboyina and Sen (2018) and references therein.

In this paper, we study the problem of estimating *piecewise monotone* vectors, which can be regarded as a generalization of isotonic regression that allows order-violating changepoints. We formulate the problem precisely as follows. Let us

consider the Gaussian sequence model

$$y_i = \theta_i^* + \xi_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

where $y = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$ is the observed vector, $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_n^*)^\top \in \mathbb{R}^n$ is the unknown parameter of interest, and $\xi = (\xi_1, \xi_2, \ldots, \xi_n)^\top$ is the unobserved noise distributed according to the Gaussian distribution $N(0, \sigma^2 I_n)$. Given the noisy observation y, the problem is to find a good piecewise monotone approximation of θ^*. Here we define piecewise monotone vectors as follows.

Definition 1.1. Let $\Pi = (A_1, A_2, \ldots, A_m)$ be a connected partition of $[n] = \{1, 2, \ldots, n\}$, that is, there exists a sequence $1 = \tau_1 < \tau_2 < \cdots < \tau_m < \tau_{m+1} = n + 1$ such that $A_i = \{\tau_i, \tau_i + 1, \ldots, \tau_{i+1} - 1\}$ $(i = 1, 2, \ldots, m)$. We say that a vector $\theta \in \mathbb{R}^n$ is *piecewise monotone* on Π if the restriction on each A_i is monotone:

$$\theta_{\tau_i} \le \theta_{\tau_i + 1} \le \cdots \le \theta_{\tau_{i+1} - 1}, \quad \text{for } i = 1, 2, \ldots, m.$$

We also say that θ is m-piecewise monotone if θ is piecewise monotone on some partition Π with $|\Pi| = m$.

We are particularly interested in the case where the number of pieces m is larger than two but much smaller than n because it is reduced to simpler problems if otherwise. From Definition 1.1, a monotone vector in $K_n^{\uparrow}$ is m-piecewise monotone for any $m \ge 1$. In particular, the least squares estimators over 1-piecewise monotone vectors coincide with the isotonic regression. Besides, since any vector in $\mathbb{R}^n$ is n-piecewise monotone, the least squares estimator over n-piecewise monotone vectors is merely the identity function $\hat{\theta}_{\mathrm{id}} = y$.

In real-world applications, there are many signals that can be approximated by piecewise monotone vectors. Here, we provide a few examples. First, in seismology, geological observations such as tide gauge records (Nagao et al. 2013) and GPS records (Roggers and Dragert 2003) often consist of a long-term monotonic trend and discontinuous jumps caused by tectonic activities. In particular, Roggers and Dragert (2003) reported that GPS measurements that are nearby a subduction zone in North America can be approximated by a sawtooth function. The top panel of Figure 1 shows an example of GPS measurements. Second, the numbers of search queries for some words related to seasons (e.g., "Christmas" and "gift") can be seen as periodic piecewise monotone signals (see the bottom panel of Figure 1 for examples). Third, in the ranking systems in online shopping websites, sales ranks of rarely sold items behave like piecewise monotone signals because they suddenly rise every time the items are sold (Hattori and Hattori 2010).

In this paper, we focus on the performance of *nearly-isotonic regression* proposed by Tibshirani et al. (2011). Given $y \in \mathbb{R}^n$ and a tuning parameter $\lambda \ge 0$, the nearly-isotonic regression estimator $\hat{\theta}_\lambda$ is defined as

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})_+ \right\}, \tag{3}$$

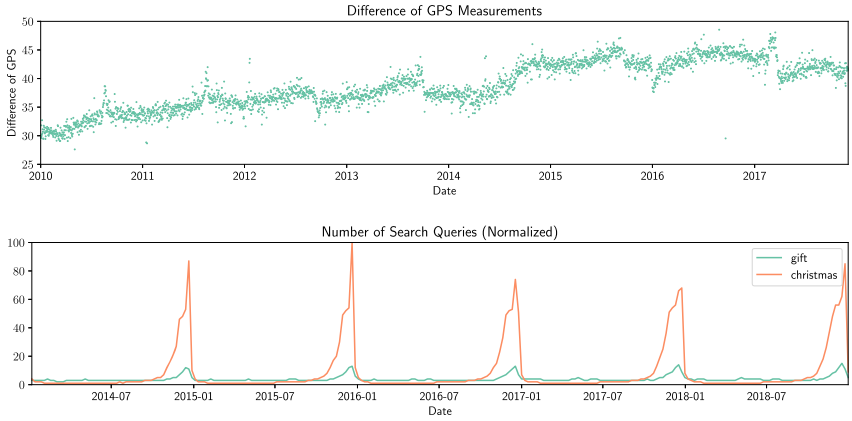

FIG 1. ***Examples of piecewise monotone signals in real-world data. Top**: The difference of the east-west component of GPS measurements between Victoria (British Columbia, Canada) and Seattle (United States). The trend factor seems to be approximated by a piecewise monotone signal. A possible reason for this behavior is the seismological phenomenon reported in Roggers and Dragert (2003). See Section 6.3 for a more detailed explanation of this data.* ***Bottom**: The numbers of search queries for two words "Christmas" and "gift" in Google Trends (https://www.google.com/trends).*

where $(z)_+ := \max\{z, 0\}$. Intuitively, the tuning parameter λ controls the degree of monotonicity. The term $(\theta_i - \theta_{i+1})_+$ poses a positive penalty if and only if the directed edge $(i, i+1)$ is *order violating*, i.e., $\theta_i > \theta_{i+1}$. Hence, a large value of $\lambda > 0$ makes the estimator $\hat{\theta}_\lambda$ close to a monotone vector. In particular, there is a sufficiently large λ such that the solution $\hat{\theta}_\lambda$ becomes exactly the same as the isotonic regression (1).

Our goal in this paper is to show that the nearly-isotonic regression can adapt to piecewise monotone vectors. As suggested in Tibshirani et al. (2011), the nearly-isotonic regression can fit to a "nearly monotone" vector that is close to $K_n^\uparrow$ in ℓ_2-sense. That is, the estimator performs well if θ^* has a small ℓ_2-misspecification error $\mathrm{dist}(\theta^*, K_n^\uparrow)$ defined as

$$\mathrm{dist}(\theta^*, K_n^\uparrow) := \inf_{\theta \in K_n^\uparrow} \|\theta^* - \theta\|_2.$$

Moreover, we can observe that the nearly-isotonic regression can fit to piecewise monotone vectors, even if θ^* is far from monotone in ℓ_2-sense. Figure 2 shows an example of the nearly-isotonic regression with $n = 100$. The true parameter θ^* (orange line) is 2-piecewise monotone. By varying the values of the tuning parameter $\lambda \geq 0$, the nearly-isotonic regression behaves as follows: If $\lambda = 0$, the nearly-isotonic regression is just the identity estimator $\hat{\theta}_{\mathrm{id}} = y$, which clearly overfits to the noisy observation. If λ is set to a sufficiently large value, $\hat{\theta}_\lambda$ coincides with the isotonic regression. In this example, however, the ℓ_2-misspecification error $\mathrm{dist}^2(\theta^*, K_n^\uparrow)$ is large compared with the normalized noise variance σ^2/n. We can see that the mean squared error (MSE) $\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta} - \theta^*\|_2^2$ of

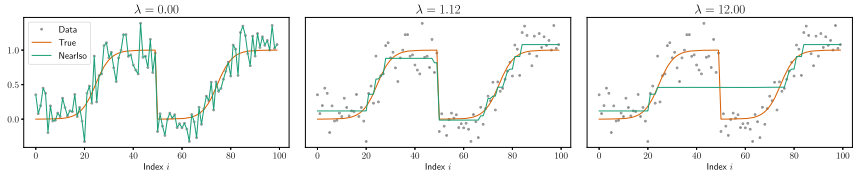

FIG 2. ***Examples of the nearly-isotonic regression estimators with different choices of tuning parameters.*** *The nearly-isotonic regression interpolates between the identity estimator* $\hat{\theta}_{\mathrm{id}} = y$ *and the isotonic regression* $\hat{\theta}_{K_n^{\uparrow}}$.

the isotonic regression can be much worse than that of the identity estimator, which coincides with σ^2/n (see Section 3.2). Indeed, we can choose a 2-piecewise monotone vector $\theta^* \in K^{\uparrow}_{n/2} \times K^{\uparrow}_{n/2}$ with arbitrarily large ℓ_2-misspecification error. If we choose an intermediate value of λ, the nearly-isotonic regression seems to fit to the true parameter. This suggests the adaptation property to piecewise monotone vectors.

1.1. Summary of theoretical results

In this paper, we investigate the adaptation property of the nearly-isotonic regression estimators defined in (3).

In the monotone regression setting (i.e., $m = 1$), it is known that the isotonic regression estimator $\hat{\theta}_{K_n^{\uparrow}}$ achieves the risk bound

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{K_n^{\uparrow}} - \theta^*\|_2^2 \le C\left(\frac{\sigma^2\mathcal{V}(\theta^*)}{n}\right)^{2/3} + \frac{C\sigma^2 \log en}{n},$$

where $\mathcal{V}(\theta) = \theta_n - \theta_1$ is the total variation of the monotone vector θ. It is also known that the rate $\mathrm{O}((\sigma^2\mathcal{V}/n)^{2/3})$ is minimax optimal under the assumption that θ^* is monotone and $\mathcal{V}(\theta^*) \le \mathcal{V}$ (Zhang 2002). Hence, a natural question is whether a similar rate can be achieved in piecewise monotone regression.

In Section 3.1, we provide the minimax lower bound over the class of piecewise monotone vectors. Let $\Theta_n(m, \mathcal{V})$ be the set of m-piecewise monotone vectors whose "upper" total variations are bounded by $\mathcal{V}$ (a precise definition is provided in Section 3.1). Then, the minimax risk over $\Theta_n(m, \mathcal{V})$ is bounded from below by a constant multiple of

$$\max\left\{\left(\frac{\sigma^2\mathcal{V}}{n}\right)^{2/3}, \frac{\sigma^2 m}{n}\log\frac{en}{m}\right\}.$$

In Section 5, we construct a concrete (but not computationally efficient) estimator that adaptively achieves this rate, and hence this lower bound is tight in the sense of the order in n, m, and $\mathcal{V}$. Intuitively, this suggest that the cost of not knowing the true partition is of order $\mathrm{O}(\frac{\sigma^2 m}{n}\log\frac{en}{m})$.

In Section 4, we provide the following risk bound for the nearly-isotonic regression estimator (3). A precise statement is given in Corollary 4.12.

Claim 1.2. Let θ^* be a piecewise monotone vector on a partition $\Pi = (A_1, A_2, \dots, A_m)$. Suppose that the following assumptions hold:

(a) The partition is equi-spaced: $|A_1| = |A_2| = \cdots = |A_m|$ $(= \frac{n}{m})$.
(b) For each segment A_j, $\theta^*_{A_j}$ is monotone and the total variation is bounded as $\mathcal{V}(\theta^*_{A_j}) \le \mathcal{V}/m$.
(c) $\theta^*_{A_j}$ satisfies an appropriate "growth condition" for each $j = 1, \dots, m$.

Then, the estimator (3) with optimally tuned parameter λ satisfies the following risk bound:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \le C\left\{\left(\frac{\sigma^2 \mathcal{V} \log en}{n}\right)^{2/3} + \frac{\sigma^2 m}{n}\log\frac{en}{m}\right\}. \tag{4}$$

The above claim is obtained as a corollary of a more general risk bound in Section 4. In the above statement, we make somewhat restrictive assumptions. Here, (a) and (b) are introduced just for the sake of notation simplicity, whereas (c) is an essential assumption. If we assume only (a) and (b), the rate that appeared in (4) is minimax optimal up to a logarithmic multiplication factor. However, we require an extra growth condition (c), which seems to be unavoidable for the estimator (3). We will provide a precise definition of the growth condition in Section 4.3.

1.2. Organization

The rest of this paper is organized as follows. In Section 2, we give a brief literature review on the shape restricted regression and regularization based estimators and relate our theoretical results to previous work. We provide lower bounds on the risks in the piecewise monotone regression problem in Section 3. In Section 4, we describe our main results on the risk upper bounds for the nearly-isotonic regression estimator and its constrained form variant. In particular, a precise statement of Claim 1.2 in the above is provided in Section 4.3. In Section 5, we discuss the attainability of the minimax lower bound; herein, we provide a concrete example of a model selection-based estimator that achieves the optimal rate. Furthermore, we present some numerical examples in Section 6. Finally, we present our conclusion in Section 7. We have also included appendices which contain additional numerical examples on two-dimensional signals, explanations of algorithms, and all proofs of the theoretical results.

1.3. Notation

Throughout this paper, we assume that $y = \theta^* + \xi$ is distributed according to an isotropic normal distribution $N(\theta^*, \sigma^2 I_n)$, where $\theta^* \in \mathbb{R}^n$ is the true mean

parameter of interest and $\xi \sim N(0, \sigma^2 I_n)$ is the noise vector. The symbol $\mathbb{E}_{\theta^*}$ denotes the expectation with respect to y.

We sometimes denote by C an absolute positive constant whose value may vary.

For any $\theta \in \mathbb{R}^n$, we define the total variation $\mathcal{V}(\theta)$ and the *lower total variation* $\mathcal{V}_-(\theta)$ by

$$\mathcal{V}(\theta) := \sum_{i=1}^{n-1} |\theta_i - \theta_{i+1}| \quad \text{and} \quad \mathcal{V}_-(\theta) := \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})_+,$$

where $(z)_+ := \max\{z, 0\}$ for any $z \in \mathbb{R}$. For example, if θ is monotone non-decreasing, then $\mathcal{V}(\theta) = \theta_n - \theta_1$ and $\mathcal{V}_-(\theta) = 0$. In this paper, the meaning of subscripts of θ depends on the context (e.g., θ_i, θ_A, $\hat{\theta}_\lambda$, and $\hat{\theta}_{K_n^\uparrow}$). If $A = \{\tau, \tau+1, \ldots, \tau+J-1\}$ is a connected subset of $[n]$, we denote by θ_A a sub-vector $(\theta_\tau, \theta_{\tau+1}, \ldots, \theta_{\tau+J-1})^\top \in \mathbb{R}^J$. We also denote by $\mathcal{V}^A(\theta_A)$ the total variation of θ_A.

2. Related work

There are two classes of estimators that are closely related to the nearly-isotonic regression (3): the isotonic regression and the fused lasso.

As we mentioned above, the isotonic regression is an instance of shape restricted regression. Many existing estimators in shape restricted regression can be formulated as least squares estimators (denoted by $\hat{\theta}_K$) onto closed convex sets (denoted by K). Examples include, but not limited to, the isotonic regression, the isotonic regression in two-dimensional grid or more general partial orders (see e.g., Robertson and Wright (1975) and Kyng et al. (2015)), and convex regression (Hildreth 1954).

Recently, researchers have developed two important techniques for analyzing risk behaviors of least squares estimators. First, Chatterjee (2014) proved that the Euclidean norm $\|\hat{\theta}_K - \theta^*\|_2$ is tightly concentrated around a certain quantity defined by the *localized Gaussian width*. As applications of Chatterjee's method, non-asymptotic upper bounds that have similar rates to the minimax risks have been proved for the isotonic regression (Chatterjee 2014, Bellec 2018), the multi-isotonic regression on two or more high dimension (Chatteejee et al. 2018, Han et al. 2019), the multi-dimensional convex regression (Han and Wellner 2016), and the constrained form trend filtering estimator (Guntuboyina et al. 2020). See also Section 2.2 in Bellec (2018) for a related result. Second, risk bounds based on the *statistical dimension* of the tangent cone of K has been developed by Oymak and Hassibi (2016) and Bellec (2018). This technique is useful because it takes into account the facial structure of K, which leads to risk bounds that are adaptive to low dimensional sub-structures. It has been shown that some least squares estimators are adaptive to piecewise constant vectors: for example, the isotonic regression (Bellec 2018) and the multi-isotonic regression (Chatteejee et al. 2018, Han et al. 2019). In particular, for the one-dimensional isotonic

regression, Chatterjee et al. (2015) and Bellec (2018) proved the following oracle inequality

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{K_n^\uparrow} - \theta^*\|_2^2 \le \inf_{\theta \in K_n^\uparrow}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + \frac{\sigma^2 k(\theta)}{n}\log\frac{en}{k(\theta)}\right\}, \tag{5}$$

where $k(\theta)$ is the number of constant pieces of θ. If θ^* is monotone and $k(\theta^*)$ is small, the right-hand side can be much smaller than the worst-case rate of $\mathrm{O}((\sigma^2\mathcal{V}/n)^{2/3})$. However, the first term in the right-hand side can become arbitrarily large if θ^* is not included in $K_n^\uparrow$.

The fused lasso (Tibshirani et al. 2005), also known as the total variation regularization (Rudin et al. 1992), is a penalized estimator defined as

$$\hat{\theta}_{\mathrm{FL},\lambda} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n}\left\{\frac{1}{2}\|y - \theta\|_2^2 + \lambda\sum_{i=1}^{n-1}|\theta_i - \theta_{i+1}|\right\}, \tag{6}$$

where $\lambda \ge 0$ is the tuning parameter. The fused lasso poses the penalty whenever $\theta_i \ne \theta_{i+1}$, whereas the penalty of the nearly-isotonic regression (3) activates only if $\theta_i > \theta_{i+1}$. Theoretical risk bounds for the fused lasso have been studied by Mammen and van de Geer (1997), Dalalyan et al. (2017), Lin et al. (2017), and Guntuboyina et al. (2020). In particular, Guntuboyina et al. (2020) showed an oracle inequality of the following form:

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathrm{FL},\lambda^*} - \theta^*\|_2^2 \\ &\le \inf_{\theta \in \mathbb{R}^n}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\frac{\sigma^2 k(\theta)}{n}\log\frac{en}{k(\theta)} + C\Delta_{\mathrm{FL}}(\theta)\right\},\end{aligned} \tag{7}$$

where λ^* is an optimally tuned parameter. One can control the quantity $\Delta_{\mathrm{FL}}(\theta)$ by assuming a mild regularity condition on θ^* so that the inequality (7) recovers the minimax rate for the piecewise constant vectors (see e.g., Gao et al. (2017)). However, even if θ^* is a monotone vector, (7) does not recover the rate of the isotonic regression (5) because $\Delta_{\mathrm{FL}}(\theta)$ becomes zero if and only if θ is just a constant vector.

Our risk bound for the nearly-isotonic regression in Section 4.2 fills the gap between the above risk bounds for the isotonic regression and the fused lasso. We will show an oracle inequality of the following form:

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathrm{NI},\lambda^*} - \theta^*\|_2^2 \\ &\le \inf_{\theta \in \mathbb{R}^n}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\frac{\sigma^2 k(\theta)}{n}\log\frac{en}{k(\theta)} + C\Delta_{\mathrm{NI}}(\theta)\right\}.\end{aligned}$$

Like in the case of the fused lasso (7), this inequality provides a meaningful risk bound even if we cannot approximate θ^* by a monotone vector. Furthermore, $\Delta_{\mathrm{NI}}(\theta)$ becomes zero for any monotone vector $\theta \in K_n^\uparrow$. Hence, our result can exactly recover the rate achieved by the isotonic regression (5).

3. Lower bounds

In this section, we provide lower bounds for the risk in one-dimensional piecewise monotone regression.

3.1. Minimax lower bound

We are interested in the lower bound for the minimax risk defined as

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta} \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|_2^2,$$

where $\Theta \subset \mathbb{R}^n$ is a set of piecewise monotone vectors, and the infimum is taken over all (measurable) estimators $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$. In particular, for $1 \le m \le n$, we consider the class of m-piecewise monotone vectors with a bounded total variation that is defined as follows.

Definition 3.1. Let $n \ge 2$ and $1 \le m \le n$. For any $\mathcal{V} > 0$, let $\tilde{\Theta}_n(m, \mathcal{V})$ denote the set of (at most) m-piecewise monotone vectors such that the upper total variation is bounded by $\mathcal{V}$. In other words, a vector $\theta \in \mathbb{R}^n$ is an element of $\tilde{\Theta}_n(m, \mathcal{V})$ if and only if the following conditions hold:

(i) θ is piecewise monotone on a connected partition $\Pi = \{A_1, \dots, A_{m^*}\}$ of $[n]$ whose cardinality $|\Pi| = m^*$ is not larger than m.
(ii) There exist numbers $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_{m^*}$ such that $\sum_{i=1}^{m^*} \mathcal{V}_i = \mathcal{V}$, $\mathcal{V}_i \ge 0$, and $\mathcal{V}(\theta_{A_i}) \le \mathcal{V}_i$ for all $i = 1, \dots, m^*$.

In addition, we also define $\Theta_n(m, \mathcal{V})$ as the set of m-piecewise monotone vectors such that the total variations for all pieces are uniformly bounded by $\mathcal{V}/m$. That is, $\Theta_n(m, \mathcal{V})$ is obtained by replacing (ii) by the following condition:

(ii)' $\mathcal{V}(\theta_{A_i}) \le \mathcal{V}/m$ for all $i = 1, \dots, m^*$.

First, we consider θ^* is piecewise monotone on a *known* partition $\Pi^* = \{A_1, A_2, \dots, A_{m^*}\}$ and that the total variation of the sub-vector $\theta^*_{A_i}$ is bounded as $\mathcal{V}(\theta_i^*) \le \mathcal{V}_i$ for each $i = 1, 2, \dots, m^*$. Then, the problem is decomposed into m^* independent subproblems of estimating monotone vectors θ_i^*. The minimax risk lower bound for monotone vectors has been proved by Zhang (2002) and Chatterjee et al. (2015). For simplicity in the notation, we assume here that $n_i = |A_i| \ge 2$ for all $i = 1, 2, \dots, m$. The minimax risk can be written as

$$\inf_{\hat{\theta}_i} \sup_{\substack{\theta^*_{A_i} \in K^{\uparrow}_{A_i}: \\ \mathcal{V}(\theta_i^*) \le \mathcal{V}_i}} \frac{1}{n_i} \mathbb{E}_{\theta^*_{A_i}} \|\hat{\theta}_i - \theta_i^*\|_2^2 \ge C_1 \left(\frac{\sigma^2 \mathcal{V}_i}{n_i}\right)^{2/3} \quad \text{for all } i = 1, \dots, m. \tag{8}$$

Hence, the minimax risk over $\tilde{\Theta}_n(m, \mathcal{V})$ is clearly bounded from below by

$$C_1 \sum_{i=1}^{m^*} \frac{n_i}{n} \left(\frac{\sigma^2 \mathcal{V}_i}{n_i}\right)^{2/3}. \tag{9}$$

If the partition Π^* is known, then this convergence rate can be obtained by concatenating the least squares estimators on all pieces. By Jensen's inequality, the quantity (9) is not larger than $(\sigma^2 \sum_i \mathcal{V}_i/n)^{2/3}$.

In the general setting, we have to deal with *unknown* partitions. The following proposition gives the lower bound over the class of piecewise monotone vectors in Definition 3.1.

Proposition 3.2. Let $n \geq 3$, $3 \leq m \leq n$, and $\mathcal{V} > 0$. Suppose that Θ is either $\tilde{\Theta}_n(m, \mathcal{V})$ or $\Theta_n(m, \mathcal{V})$ in Definition 3.1. Then, for any estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$, we have the following lower bound:

$$\sup_{\theta^* \in \Theta} \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|_2^2 \geq C \max \left\{ \left(\frac{\sigma^2 \mathcal{V}}{n} \right)^{2/3}, \quad \frac{\sigma^2 m}{n} \log \frac{en}{m} \right\}, \tag{10}$$

where $C > 0$ is a universal constant.

It remains to verify that the lower bound (10) is tight. Thus, in Section 5, we will construct an estimator that adaptively achieves a similar rate.

3.2. Lower bound of isotonic regression with misspecified partitions

Suppose that θ^* is an m-piecewise monotone vector. As we mentioned in the previous subsection, if we know the true partition on which θ^* is monotone, the least squares estimator can achieve the rate shown in (9). Here, we consider what happens if we underestimate the true number of the pieces.

We consider the risk behavior of the isotonic regression $\hat{\theta}_{K_n^\uparrow}$, which corresponds to the least squares estimator for the underestimated number of pieces as $m = 1$. If the true number of pieces is larger than or equal to two, θ^* may not be contained in $K_n^\uparrow$. Recall that $\mathrm{dist}(\theta^*, K_n^\uparrow)$ is the ℓ_2-misspecification error against the set of monotone vectors. Bellec (2018) showed that the isotonic regression is robust against a small ℓ_2-misspecification, that is, if $\mathrm{dist}(\theta^*, K_n^\uparrow) \leq \epsilon$, then

$$\frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta}_{K_n^\uparrow} - \theta^*\|_2^2 \leq \epsilon^2 + \frac{\sigma^2 k(\bar{\theta})}{n} \log \frac{en}{k(\bar{\theta})},$$

where $k(\bar{\theta})$ is the orthogonal projection of θ^* onto $K_n^\uparrow$. Conversely, if the ℓ_2-misspecification error is large, we see that the isotonic regression can have an arbitrarily large risk.

Proposition 3.3. There is a positive number $t = t_{n,\sigma^2}$ that depends on n and σ^2 such that if the true parameter θ^* satisfies $\mathrm{dist}(\theta^*, K_n^\uparrow) > t$, then the MSE of the isotonic regression is bounded from below as

$$\frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta}_{K_n^\uparrow} - \theta^*\|_2^2 > \sigma^2.$$

In this case, the isotonic regression has a strictly larger MSE than that of the identity estimator $\hat{\theta}_{\mathrm{id}} = y$.

We can easily check that there is a 2-piecewise monotone vector with an arbitrarily large ℓ_2-misspecification error. To see this, let $\theta^* \in \mathbb{R}^{2n}$ be a piecewise constant vector defined as $\theta^*_i = M > 0$ for $i = 1, \ldots, n$ and $\theta^*_i = 0$ for $i = n+1, \ldots, 2n$. Then, it is easy to see that $\mathrm{dist}(\theta^*, K^{\uparrow}_{2n}) = \sqrt{nM^2/2}$ diverges as $M \to \infty$. Figure 2 shows an example of a 2-piecewise monotone vector θ^* such that the isotonic regression has a larger squared loss value than the identity estimator.

4. Risk bounds for nearly-isotonic regression

In this section, we develop the risk bound for the nearly-isotonic regression estimator (3). Proofs of all the theorems and propositions in this section are presented in Appendix D.

4.1. Risk bounds for constrained estimators

Before considering the original version of the nearly-isotonic regression (3), we consider the performance of the *constrained form nearly-isotonic regression* $\hat{\theta}_{\mathcal{V}}$ defined by the following constrained optimization problem:

$$\text{minimize } \|y - \theta\|_2^2 \quad \text{subject to } \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})_+ \leq \mathcal{V}, \tag{11}$$

where $\mathcal{V} \geq 0$ is the tuning parameter. By the fundamental duality theorem in convex optimization, there exists a Lagrange multiplier $\lambda_{\mathcal{V}} \geq 0$ such that the regularization type formulation (3) admits the same solution $\hat{\theta}_{\lambda_{\mathcal{V}}} = \hat{\theta}_{\mathcal{V}}$. Hence, the solution path of penalized estimators $\{\hat{\theta}_\lambda : \lambda \geq 0\}$ and that of constrained estimators $\{\hat{\theta}_{\mathcal{V}} : \mathcal{V} \geq 0\}$ are equivalent. However, the properties of estimators with fixed values of $\lambda \geq 0$ and $\mathcal{V} \geq 0$ can be different in the following sense:

- From a computational perspective, calculating the constrained estimator (11) for a given $\mathcal{V} \geq 0$ is more difficult than the regularization estimator (3). For the regularization estimator (3), we can use the Modified Pool Adjacent Violators Algorithm (Modified PAVA) proposed by Tibshirani et al. (2011), which outputs the solution path for every $\lambda \geq 0$. In particular, given $\lambda \geq 0$, we can always obtain an *exact* solution $\hat{\theta}_\lambda$. However, to the best of our knowledge, there are no practical algorithms that obtain an exact solution for the constrained problem (11) that run as fast as the algorithms for the penalized problem (3). We present detailed explanations for the algorithms in Section A.
- From a statistical perspective, the correspondence between tuning parameters λ and $\mathcal{V}$ is not deterministic (i.e., it depends on the realization of the data y). For this reason, a risk bound that is obtained for one of (3) or (11) cannot be directly applied to the other.

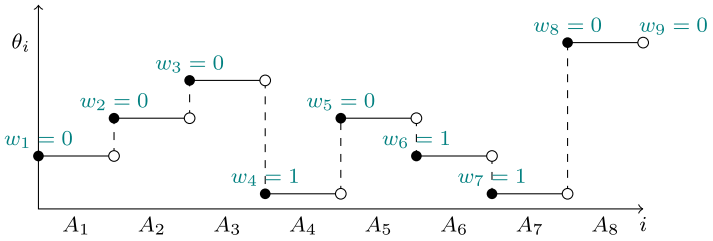

FIG 3. ***Illustration of the knot signs defined in*** (12)***.*** *In this example, θ is assumed to be k-piecewise constant with $k = 8$. The corresponding signs are given as $(w_1, w_2, \ldots, w_8, w_9) = (0, 0, 0, 1, 0, 1, 1, 0, 0)$. Moreover, if we assume $|A_1| = |A_2| = \cdots = |A_8|$, the quantity $M(\theta)$ defined in* (13) *is given as $M(\theta) = \frac{1}{|A_4|} + \frac{1}{|A_5|} + \frac{1}{|A_6|} + \frac{1}{|A_8|} = \frac{4k}{n}$.*

We show the main results on the adaptation property to piecewise monotone vectors in terms of sharp oracle inequality.

Before proceeding, we introduce some notations. Suppose that $\theta \in \mathbb{R}^n$ is piecewise constant on a connected partition $\Pi_{\mathrm{const}} = \{A_1, \ldots, A_k\}$ of $[n]$. We denote by $k(\theta) := |\Pi_{\mathrm{const}}|$ the number of pieces in which θ becomes constant. That is, there are integers $1 = \tau_1 < \cdots < \tau_{k+1} = n + 1$ such that (i) $A_i = \{\tau_i, \tau_i + 1, \ldots, \tau_{i+1} - 1\}$ for $i = 1, \ldots, k$ and (ii) for any $i \in [k]$, there exists $t_i \in \mathbb{R}$ such that $\theta_j = t_i$ for all $j \in A_i$. We define the sign $w_i \in \{0, 1\}$ associated with each knot τ_i $(i = 1, \ldots, k + 1)$ as

$$w_1 = w_{k+1} = 0 \quad \text{and}$$
$$w_i = \begin{cases} 1 & (t_{i-1} > t_i) \\ 0 & (t_{i-1} < t_i) \end{cases} \quad \text{for } i = 2, \ldots, k. \tag{12}$$

In other words, $w_i = 1$ if and only if the order violation $\theta_{j-1} > \theta_j$ occurs at $j = \tau_i$. See Figure 3 for the graphical illustration. Then, we define $M(\theta)$ as

$$M(\theta) := \sum_{j=2}^{k} \max\left\{\frac{1}{|A_j|}, \frac{k}{n}\right\} 1_{\{w_{j-1} \neq w_j\}}. \tag{13}$$

$M(\theta)$ determines the non-monotonicity of a piecewise constant vector θ. If θ is m-piecewise monotone, then it is clear that $M(\theta) \leq 2(m - 1)$. In particular, for any monotone vector θ, we have $M(\theta) = 0$. Based on these notations, we have the following sharp oracle inequality.

Theorem 4.1. For any $\theta^* \in \mathbb{R}^n$, the constrained nearly-isotonic regression (11) satisfies the following oracle inequality:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2$$
$$\leq \inf_{\substack{\theta \in \mathbb{R}^n: \\ \mathcal{V}_-(\theta) = \mathcal{V}}} \left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + C\sigma^2 \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)}\right\}. \tag{14}$$

Moreover, for any $\eta \in (0,1)$, we have

$$\begin{aligned}&\frac{1}{n}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \\ &\leq \inf_{\substack{\theta \in \mathbb{R}^n: \\ \mathcal{V}_-(\theta) = \mathcal{V}}} \left\{ \frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + C\sigma^2 \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)} \right\} \\ &+ \frac{4\sigma^2 \log \eta^{-1}}{n} \end{aligned} \tag{15}$$

with probability at least $1 - \eta$.

The following risk bound for the best choice of the tuning parameter $\mathcal{V} \geq 0$ is an immediate consequence of Theorem 4.1.

Corollary 4.2. Suppose $\theta^* \in \mathbb{R}^n$. Choose $\mathcal{V}^* \geq 0$ that minimizes the upper bound in (14) (thus, $\mathcal{V}^*$ depends on the true parameter θ^*). Then, we have

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}^*} - \theta^*\|_2^2 \\ &\leq \inf_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + C\sigma^2 \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)} \right\}. \end{aligned} \tag{16}$$

Also, choosing $\mathcal{V} := \mathcal{V}^*$ or $\mathcal{V} := \mathcal{V}_-(\theta^*)$, we have

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \leq C\sigma^2 \left\{ \frac{k(\theta^*)}{n} \log \frac{en}{k(\theta^*)} + \frac{M(\theta^*)}{k(\theta^*)} \log \frac{en}{k(\theta^*)} \right\}. \tag{17}$$

Remark 4.3. We briefly comment on the proof of Theorem 4.1 and Corollary 4.2. A key ingredient is to obtain a bound on the *statistical dimension* (Amelunxen et al. 2014) of the tangent cone of the constraint set $\{\theta \in \mathbb{R}^n : \mathcal{V}_-(\theta) \leq \mathcal{V}\}$. This methodology was first developed for the isotonic regression and the convex regression by Bellec (2018). In particular, our approach is inspired by the analysis of the constrained trend filtering estimators by Guntuboyina et al. (2020). See Appendix D for detailed proofs.

By restricting the region over which the infimum in (16) is taken, we have the oracle inequality for monotone vectors

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}^*} - \theta^*\|_2^2 \leq \inf_{\theta \in K_n^{\uparrow}} \left\{ \frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} \right\},$$

which recovers the existing results on the isotonic regression (Chatterjee et al. 2015, Bellec 2018) up to a constant multiplicative factor.

To understand the general upper bound in (16), we have to control the quantity $M(\theta)$ defined in (13). To this end, we consider the *minimal length condition*; we say that $\theta \in \mathbb{R}^n$ satisfies the minimal length condition for a constant $c > 0$ if it satisfies

$$\min\{|A_i| : 1 \leq i \leq k, w_i \neq w_{i+1}\} \geq \frac{cn}{k}, \tag{18}$$

where the partition $\Pi_{\mathrm{const}} = \{A_1, A_2, \ldots, A_k\}$ and the signs w_i $(i = 1, \ldots, k+1)$ are defined as in (13). Intuitively, a signal $\theta \in \mathbb{R}^n$ is well approximated by another signal that satisfies the minimal length condition if θ has "moderate slopes" around the order-violating jumps. For further discussion on such growth conditions, see Section 4.3.

Based on the minimal length condition, we have the following result from Theorem 4.1.

Corollary 4.4. Suppose that $\theta^* \in \mathbb{R}^n$ satisfies the minimal length condition (18) for a constant $c > 0$. Assume that θ^* is $k(\theta^*)$-piecewise constant and $m(\theta^*)$-piecewise monotone. Then, the constrained nearly-isotonic regression (11) satisfies

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \\ &\leq (\mathcal{V}_-(\theta^*) - \mathcal{V})^2 + C\sigma^2\left(\frac{k(\theta^*)}{n} + \frac{2c^{-1}(m(\theta^*)-1)}{n}\right)\log\frac{en}{k(\theta^*)}. \end{aligned} \tag{19}$$

In particular, if the tuning parameter $\mathcal{V}$ is chosen so that

$$(\mathcal{V}_-(\theta^*) - \mathcal{V})^2 \leq C'\frac{k(\theta^*)}{n}\log\frac{en}{k(\theta^*)}$$

for a positive constant C', we have

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \leq C''\sigma^2\left(\frac{k(\theta^*)}{n} + \frac{2c^{-1}(m(\theta^*)-1)}{n}\right)\log\frac{en}{k(\theta^*)},$$

where C'' is a positive constant.

Remark 4.5. If θ is k-piecewise constant and m-piecewise monotone, it is always true that $k \geq 2(m-1)$. Hence, the inequality (19) can be simplified as

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \leq (\mathcal{V}_-(\theta^*) - \mathcal{V})^2 + C(c)\sigma^2\frac{k(\theta^*)}{n}\log\frac{en}{k(\theta^*)},$$

where $C(c) > 0$ is a constant that depends on c alone.

Remark 4.6. We comment on the minimal length condition and the relation to estimation of piecewise constant vectors. We conjecture that the minimum length condition (18) is essentially unavoidable for the risk bound of the nearly-isotonic regression due to the following analogy to the fused lasso. The minimal length condition for the fused lasso is considered by Guntuboyina et al. (2020). For the fused lasso, Fan and Guan (2018) showed that the minimum length condition cannot be removed in the sense that there is a lower bound depending on the minimum length $\Delta = \min_i |A_i|$ (see also the experimental result by Guntuboyina et al. (2020), Remark 2.5).

4.2. Risk bounds for penalized estimators

In this section, we consider the risk bounds for the nearly-isotonic regression (3) in the original penalized form by Tibshirani et al. (2011).

Theorem 4.7. For any $\lambda \geq 0$, let $\hat{\theta}_\lambda$ denote the nearly-isotonic regression estimator defined in (3). Let θ^* and θ be any vectors in $\mathbb{R}^n$. Then, there exists a tuning parameter $\lambda^* = \lambda^*(\theta) \geq 0$ that depends only on θ such that, for any $\lambda \geq \lambda^*$, we have the following risk bound:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq \frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)} + 3(\lambda - \lambda^*)^2 M(\theta), \tag{20}$$

where $M(\theta)$ and $k(\theta)$ are defined similarly as in Theorem 4.1. Furthermore, for any $\eta \in (0,1)$, the inequality

$$\frac{1}{n}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq \frac{1}{n}\|\theta - \theta^*\|_2^2 + 2C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + 2C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)} + 6(\lambda - \lambda^*)^2 M(\theta) + \frac{16\sigma^2\log\eta^{-1}}{n} \tag{21}$$

holds with probability $1 - \eta$.

We comment on some direct consequences of Theorem 4.7. In this theorem, $\lambda^*(\theta)$ is defined as a function of θ. To understand the risk bound (20), we consider the choice of the tuning parameter $\lambda \geq 0$ that depends on the true parameter θ^*. Let $\bar{\theta}$ be a vector that minimizes the quantity

$$\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)}$$

among all $\theta \in \mathbb{R}^n$. Then, taking $\lambda^{**} := \lambda^*(\bar{\theta})$, we have the following oracle inequality which has the same form as (16):

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\lambda^{**}} - \theta^*\|_2^2 \leq \inf_{\theta\in\mathbb{R}^n}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)}\right\}.$$

Moreover, if $\lambda := \lambda^{**}$ or $\lambda := \lambda^*(\theta^*)$, we have

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq C\sigma^2\left\{\frac{k(\theta^*)}{n}\log\frac{en}{k(\theta^*)} + \frac{M(\theta^*)}{k(\theta^*)}\log\frac{en}{k(\theta^*)}\right\}.$$

Again, if we assume the minimal length condition (18) on θ^*, we obtain a simplified bound of the form (17).

We move on to discuss a precise expression of $\lambda^*(\theta)$ in Theorem 4.7. The next proposition provides an upper bound for $\lambda^*(\theta)$.

Proposition 4.8. Suppose $\theta \in \mathbb{R}^n$. Let $\Pi_{\mathrm{const}}(\theta) := \{A_1, A_2, \dots, A_k\}$ be the constant partition of θ, and $w_1, w_2, \dots, w_{k+1}$ be the associated signs defined in (12). Then, there is a universal constant $C > 0$ such that $\lambda^*(\theta)$ in Theorem 4.7 is bounded from above by

$$C\sigma \min\left\{ \frac{\|\theta\|_2}{\mathcal{V}_-(\theta)}, \left(\sum_{i=1}^{k} \frac{1_{\{w_i \neq w_{i+1}\}}}{|A_i|} \right)^{-1/2} \right\} \sqrt{\left(k(\theta) + \frac{nM(\theta)}{k(\theta)} \right) \log \frac{en}{k(\theta)}}.$$

The purpose of the choice of λ^* in Proposition 4.8 is to derive the theoretical convergence rate in terms of $k(\theta)$ and $M(\theta)$. However, different choices are possible if we are interested in other theoretical aspects (e.g., estimation consistency for changepoints). For the fused lasso estimator (6), several authors have studied theoretical choices of tuning parameters that result in risk upper bounds (Dalalyan et al. 2017, Lin et al. 2017, Guntuboyina et al. 2020).

Remark 4.9 (Example of parameter choice)**.** Here, we provide an example choice of the tuning parameter λ under a simple length condition. Let us assume that (i) θ^* is not globally monotone (i.e., $M(\theta^*) > 0$) and (ii) $|A_i|$ is of order n/k, that is,

$$c_1 \frac{n}{k} \leq |A_i| \leq c_2 \frac{n}{k}, \quad i = 1, \dots, k$$

holds for some $0 < c_1 < c_2$. Then, we can see that $\lambda^*(\theta^*)$ is bounded from above by

$$C'\sigma\sqrt{n \log en},$$

where C' is a constant that depends on C, c_1, c_2. For the fused lasso, the theoretical choice $\lambda = O(\sigma\sqrt{n \log en})$ has been suggested by Dalalyan et al. (2017) and Guntuboyina et al. (2020). For a detailed discussion, see Remark 2.7 by Guntuboyina et al. (2020) and references therein.

Remark 4.10. In general, the choice of the tuning parameter that minimizes the risk can be different from the theoretical suggestion. More importantly, we cannot obtain the value of λ suggested in Proposition 4.8 because it depends on the unknown true parameter θ^* and the noise standard deviation σ. In practice, there are two typical data-dependent choices of λ:

- **Stein's unbiased risk estimate:** If we know σ or its estimate value $\hat{\sigma}$, we can reasonably choose a parameter λ by minimizing Stein's unbiased risk estimate (SURE)

$$\mathrm{SURE}(\lambda) = \frac{1}{n}\|y - \hat{\theta}_\lambda\|_2^2 + \frac{2\hat{\sigma}^2}{n}\hat{\mathrm{df}}(\hat{\theta}_\lambda) + (\text{constant}). \tag{22}$$

  Here, $\hat{\mathrm{df}}(\hat{\theta}_\lambda) := k(\hat{\theta}_\lambda)$ is an unbiased estimate of the *degrees of freedom*. See Tibshirani et al. (2011) for the derivation.
- **Cross-validation:** We can also apply the cross-validation when the model (2) is interpreted as a discrete observation of a continuous signal. Specif-

ically, suppose that the data is generated according to the following nonparametric regression model:

$$y_i = f^*(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{23}$$

where $x_1 < x_2 < \ldots < x_n$ are given design points in $[0,1]$ and $f^* : [0,1] \to \mathbb{R}$ is an unknown piecewise monotone function. We define the nearly-isotonic regression estimator $\hat{f}_\lambda$ over the interval $[0,1]$ as follows: First, we determine the values $\hat{\theta}_{\lambda,i}$ $(i = 1, 2, \dots, n)$ by solving

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \sum_{i=1}^{n-1} \frac{(\theta_i - \theta_{i+1})_+}{x_{i+1} - x_i} \right\}. \tag{24}$$

Then, we define $\hat{f}_\lambda : [0,1] \to \mathbb{R}$ by interpolation. For instance, one can output a piecewise constant function so that $\hat{f}_\lambda(x_i) = \hat{\theta}_{\lambda,i}$. In this sense, given a new design point x^{new}, we can predict the value of $f^*(x^{\mathrm{new}})$ by $\hat{f}_\lambda(x^{\mathrm{new}})$. Hence, we can naturally apply the cross-validation in this situation.

4.3. Application to piecewise monotone vectors

To gain a deeper understanding of the adaptation property of the nearly-isotonic regression, we study the risk bound under a more specific assumption. We define the following *moderate growth condition* for piecewise monotone vectors.

Definition 4.11. Let $n \geq 2$. We say that a monotone vector $\theta \in K_n^{\uparrow}$ satisfies the moderate growth condition if

$$\theta_i \leq \theta_1 + \frac{i-1}{n-1}\mathcal{V}(\theta) \quad \text{for } i = 1, 2, \dots, \lceil n/2 \rceil$$

and

$$\theta_i \geq \theta_1 + \frac{i-1}{n-1}\mathcal{V}(\theta) \quad \text{for } i = \lceil n/2 \rceil, \lceil n/2 \rceil + 1, \dots, n.$$

Figure 4 gives an illustration of the moderate growth condition. In words, the signal $\theta \in \mathbb{R}^n$ satisfying the moderate growth condition is not larger than the linear signal in the left half of the domain, and not less than that in the right half of the domain. Intuitively, the role of the moderate growth condition is to guarantee the minimal length condition (18) for a piecewise constant approximation.

Suppose that the true signal θ^* is piecewise monotone and every segment satisfies the moderate growth condition. Then, the nearly-isotonic regression achieves a nearly minimax convergence rate as follows.

Corollary 4.12. Suppose that the following assumptions hold:

(a) The partition is equi-spaced: $|A_1| = |A_2| = \cdots = |A_m|$ $(= \frac{n}{m})$.
(b) $\theta^*_{A_j}$ is monotone and $\mathcal{V}(\theta^*_{A_j}) \leq \mathcal{V}/m$ for each $j = 1, \dots, m$.

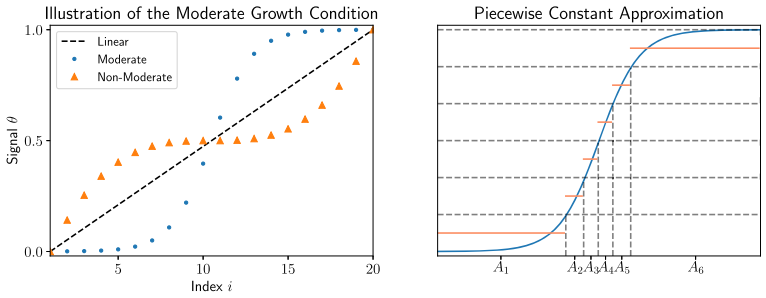

FIG 4. ***Illustration of the moderate growth condition. Left:*** *The plotted three signals are monotone vectors in* $K_n^{\uparrow}$ *with* $n = 20$ *and* $\mathcal{V}(\theta) = 1$. *The dotted line represents the linear signal* $\theta_i^{\mathrm{linear}} = i/n$ $(i = 1, 2, \ldots, n)$. *The blue circles depict an example of a signal that satisfies the moderate growth condition. That is, it is not larger than the linear signal* $\theta_i^{\mathrm{linear}}$ *for* $1 \le i \le 10$, *and not less than* $\theta_i^{\mathrm{linear}}$ *for* $10 \le i \le 20$. *On the other hand, the orange triangles depict a counterexample for this condition.* ***Right:*** *If* θ *satisfies the moderate growth condition, there is a* k*-piecewise monotone vector such that the lengths of segments at both ends are not less than* k/n. *See Appendix D.5 for a detailed explanation.*

(c) $\theta^*_{A_j}$ satisfies the moderate growth condition for each $j = 1, 2, \ldots, m$.

Then, the estimator (3) with optimally tuned parameter λ satisfies the following risk bound:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \le C \max\left\{ \left(\frac{\sigma^2 \mathcal{V} \log \frac{en}{m}}{n} \right)^{2/3}, \frac{\sigma^2 m}{n} \log \frac{en}{m} \right\}. \tag{25}$$

The risk bound (25) achieves the minimax rate over $\Theta_n(m, \mathcal{V})$ in Proposition 3.2 up to a multiplicative factor of $\log^{2/3} \frac{en}{m}$. We should note that the restrictive assumption (a) in Corollary 4.12 is employed merely for the sake of simplicity of the proof. We may relax this assumption as

$$\min_{1 \le i \le m} |A_i| \ge \frac{c' n}{m}$$

for some $c' > 0$.

5. Model selection based estimators

Here, we consider estimators obtained by model selection among all partitions Π. The main purpose of this section is to discuss whether the minimax lower bound in Proposition 3.2 can be achieved without any additional assumption such as the moderate growth condition.

Given a connected partition $\Pi = (A_1, A_2, \ldots, A_m)$ of $[n]$, we write $K_\Pi^{\uparrow}$ for the set of piecewise monotone vectors on Π, i.e.,

$$K_\Pi^{\uparrow} := K_{|A_1|}^{\uparrow} \times K_{|A_2|}^{\uparrow} \times \cdots \times K_{|A_m|}^{\uparrow}.$$

Let $\hat{\theta}_\Pi$ denote the projection estimator onto $K^\uparrow_\Pi$. By definition, $\hat{\theta}_\Pi$ is obtained by concatenating isotonic regression estimators defined in every segment.

If we know the true partition Π^* on which θ^* is piecewise monotone, then the risk of the projection estimator $\hat{\theta}_{\Pi^*}$ is bounded from above by

$$C \sum_{i=1}^{m} \frac{|A_i|}{n} \left(\frac{\sigma^2 \mathcal{V}^{A_i}(\theta^*_{A_i})}{|A_i|} \right)^{2/3}.$$

If the true partition is unknown, a natural idea is to select a data-dependent partition $\hat{\Pi}$ by a penalized selection rule:

$$\hat{\Pi} \in \operatorname*{argmin}_{\Pi} \left\{ \|y - \hat{\theta}_\Pi\|_2^2 + \mathrm{pen}(\Pi) \right\}. \tag{26}$$

Here, $\mathrm{pen}(\Pi)$ is a positive penalty for the partition Π.

The penalized selection rules have been well studied in statistics. In particular, Birgé and Massart (2001) and Massart (2007) developed non-asymptotic risk bounds for generic model selection settings in Gaussian sequence models. Hereafter, we construct a penalized selection estimator in the spirit of Theorem 4.18 in Massart (2007).

Instead of selecting $\hat{\theta}_\Pi$ according to (26), we introduce the *total variation sieves*. Namely, in addition to selecting partitions, we also select budgets of piecewise total variations as follows. Let $\Pi = (A_1, A_2, \ldots, A_m)$ be a connected partition. For any vector $\mathbf{V} = (\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m)$ with $\mathcal{V}_i \geq 0$ $(i = 1, 2, \ldots m)$, we define the set of piecewise monotone vectors with bounded total variations as

$$K^\uparrow_\Pi(\mathbf{V}) = K^\uparrow_\Pi(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m) := \{\theta \in K^\uparrow_\Pi : \mathcal{V}^{A_i}(\theta_{A_i}) \leq \mathcal{V}_i \text{ for } i = 1, 2, \ldots, m\}.$$

Then, we define $\hat{\theta}_{\Pi,\mathbf{V}}$ as the projection estimator onto $K^\uparrow_\Pi(\mathbf{V})$. Next, we define a countable set of vectors $\mathbf{V}$ as

$$\mathscr{V}(m) := \{(v(j_1), v(j_2), \ldots, v(j_m)) : \ (j_1, j_2, \ldots, j_m) \in \mathbb{N}^m\},$$

where $v(j) := j^{3/2}$. Finally, we select a pair $(\hat{\Pi}, \hat{\mathbf{V}})$ as the solution of the following minimization problem:

$$\min_{\Pi} \min_{\mathbf{V} \in \mathscr{V}(|\Pi|)} \left\{ \|y - \hat{\theta}_{\Pi,\mathbf{V}}\|_2^2 + \mathrm{pen}(\Pi, \mathbf{V}) \right\}. \tag{27}$$

With a careful choice of the penalty term $\mathrm{pen}(\Pi, \mathbf{V})$, we have the following result:

Theorem 5.1. There exists an absolute constant $C_{\mathrm{pen}} > 0$ such that the following statement holds. For any pair $(\Pi, \mathbf{V})$, define the penalty $\mathrm{pen}(\Pi, \mathbf{V})$ so that

$$\mathrm{pen}(\Pi, \mathbf{V}) = C_{\mathrm{pen}} \left(\sum_{i=1}^{m} \sigma^{4/3} |A_i|^{1/3} \mathcal{V}_i^{2/3} + \sigma^2 m \log \frac{\mathrm{e}n}{m} \right).$$

Let $(\hat{\Pi}, \hat{\mathbf{V}})$ be the minimizer in (27).

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}} - \theta^*\|_2^2 \\ &\le \min_{\Pi} \min_{\mathbf{V}\in\mathscr{V}(|\Pi|)} \left\{ \frac{3}{n}\mathrm{dist}^2(\theta^*, K_{\Pi}^{\uparrow}(\mathbf{V})) + \frac{2}{n}\mathrm{pen}(\Pi, \mathbf{V}) \right\} + \frac{256\sigma^2}{n}.\end{aligned}$$

In particular, if θ^* is piecewise monotone on $\Pi = (A_1, A_2, \ldots, A_m)$, we have

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}} - \theta^*\|_2^2 \\ &\le 2C_{\mathrm{pen}} \left\{ \sum_{i=1}^{m} \frac{|A_i|}{n} \left(\frac{\sigma^2(\mathcal{V}^{A_i}(\theta^*_{A_i}) + 1)}{|A_i|} \right)^{2/3} + \frac{\sigma^2 m}{n} \log \frac{en}{m} \right\} + \frac{256\sigma^2}{n}. \quad (28)\end{aligned}$$

We emphasize that Theorem 5.1 does not require any additional assumptions on θ^*, e.g., the minimum length condition or the moderate growth condition introduced in the previous section. Therefore, it suggests the existence of a penalized model selection estimator that achieves the minimax rate in Proposition 3.2. However, the estimator (27) is not practical for a computational reason because it is obtained through the minimization over exponentially many possible partitions Π.

The dependence on the total variation of each segment in (28) is $(\mathcal{V}^{A_i}(\theta^*_{A_i}) + 1)^{2/3}$ instead of $(\mathcal{V}^{A_i}(\theta^*_{A_i}))^{2/3}$. The additional constant 1 is due to the minimal resolution of the sieve. To establish a non-asymptotic risk bound for the penalized model selection estimator without sieves (i.e., (26)) and remove the dependence on the sieve resolution remains an open problem.

6. Simulations

We provide some numerical examples for piecewise monotone regression problems.

6.1. Dealing with inconsistency at boundaries

Before presenting the simulation results, we here explain a well-known practical issue in the isotonic regression literature and a regularization method to cope with it.

In the study of statistical estimation under monotonicity constraints, it is known that the least squares estimator $\hat{\theta}_{K_n^{\uparrow}}$ is inconsistent at the boundary points (see e.g., Groeneboom and Jongbloed (2014) and Woodroofe and Sun (1993)). A similar issue arises for the nearly-isotonic regression estimators. Since the penalty term in (3) does not activate if the orders are not violated at the boundary points (i.e., $y_1 < y_2$ or $y_{n-1} < y_n$), the nearly-isotonic regression is not robust against a negative noise at the left boundary or a positive noise at

the right boundary. To overcome this issue, we consider the following boundary correction regularization for the nearly-isotonic regression:

$$\hat{\theta}_{\mathrm{boundary},\lambda,\mu} = \operatorname*{argmin}_{\theta\in\mathbb{R}^n} \left\{ \frac{1}{2}\|y-\theta\|_2^2 + \lambda \sum_{i=1}^{n} (\theta_i - \theta_{i+1})_+ + \mu(\theta_n - \theta_1) \right\}, \quad (29)$$

where $\mu > 0$ is an additional tuning parameter. It can easily be checked that the solution is equivalent to that of the ordinary nearly-isotonic regression (3) applied to $\tilde{y} = (y_1 + \mu, y_2 \ldots, y_{n-1}, y_n - \mu)$. Similar regularization methods for isotonic regression have been studied by Chen et al. (2015), Wu et al. (2015) and Luss and Rosset (2017).

6.2. Simulation data

Here, we evaluate the performance of the nearly-isotonic regression and related estimators on simulated data. According to the one-dimensional regression model (23), we generated data with equi-spaced design points $x_i = (i-1)/n$ $(i = 1, 2, \ldots, n)$. For the true function f^*, we consider m-piecewise monotone functions defined as

$$f^{(m)}(x) := \sum_{j=1}^{m} f(mx - (j-1)) 1_{I_j}(x)$$

where $f : [0, 1) \to \mathbb{R}$ is a given monotone function and $I_j := [(j-1)/m, j/m)$ for $j = 1, 2, \ldots, m$. Following Meyer and Woodroofe (2000), we choose f from the following two monotone functions:

$$f_{\mathrm{sigmoid}}(x) = \mathrm{e}^{16x-8}/(1+\mathrm{e}^{16x-8}),$$
$$f_{\mathrm{cubic}}(x) = (2x-1)^3 + 1.$$

Figure 2 shows an example of $f = f_{\mathrm{sigmoid}}$ and $m = 2$. It is worth noting that the former sigmoidal function f_{sigmoid} satisfies the moderate growth condition (see Definition 4.11), whereas the latter cubic function f_{cube} does not. Hence, for the case of piecewise sigmoidal functions $f^{(m)}_{\mathrm{sigmoid}}$, the minimax rate of $\mathrm{O}(n^{-2/3})$ is achieved by both the nearly-isotonic regression and the fused lasso (see Corollary 4.12 above and Corollary 2.8 by Guntuboyina et al. (2020)).

In our experiments, the size n of the signal is chosen from $\{2^6, 2^7, \ldots, 2^{10}\}$. The noise standard deviation σ is assumed to be known and fixed to 0.25. We evaluated the MSE for the following four estimators:

- `Neariso`: The nearly-isotonic regression (3).
- `NearisoBC`: The nearly-isotonic regression with boundary correction (29)
- `Fused`: The fused lasso (6).
- `PO`: The projection estimator with the partition oracle, i.e., the projection estimator onto $K^{\uparrow}_{\Pi}$ provided with the true partition Π.

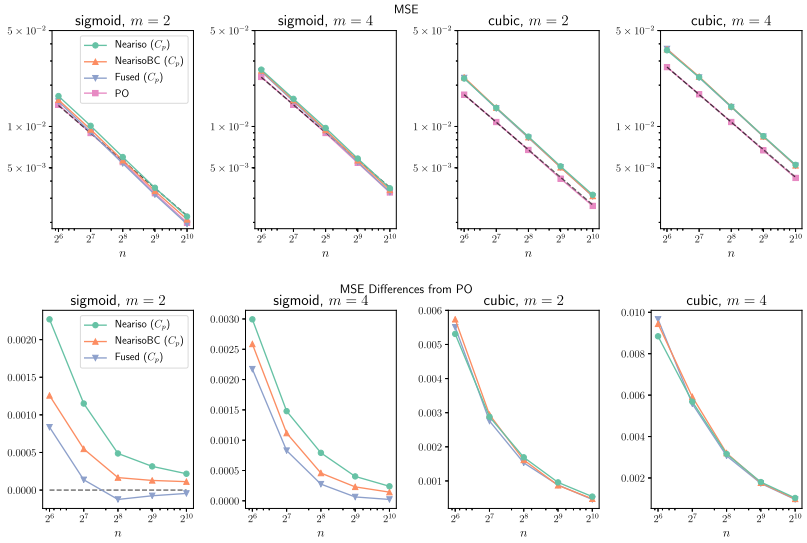

FIG 5. ***The risks of nearly-isotonic type estimators on simulated data.*** *The upper line shows log-log plots of the MSEs versus n. The lower line shows the difference of the MSEs between regularization type estimators (i.e.,* `Neariso` `NearisoBC` *and* `Fused`*) and the projection estimator with the oracle partition choice (*`PO`*).*

For `Neariso` and `Fused`, the tuning parameter λ is selected by generalized C_p criteria (i.e., minimizing SURE (22)). For `NearisoBC`, the tuning parameters (λ, μ) are selected by a similar criterion. To estimate the MSE, we generated 500 replications of the data and calculated the average value of the squared loss $\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2$.

Figure 5 presents the results for $m = 2, 4$ and $f = f_{\text{sigmoid}}, f_{\text{cubic}}$. The upper line shows log-log plots of the MSE versus n. In each setting, the three regularization based estimators (i.e., `Neariso` `NearisoBC` and `Fused`) performed as well as the ideal estimator `PO`, whereas the former three estimators do not use the information about the true partition. The risks of `PO` are well fitted by lines of slopes of $-2/3$, which means that the speed of the convergence is about the minimax optimal rate of $\mathrm{O}(n^{-2/3})$.

Next, we provide more detailed comparisons of regularization based estimators. The lower line in Figure 5 shows the difference of MSEs from that of `PO`. For piecewise sigmoidal functions, `NearisoBC` and `Fused` performed better than `Neariso`. Notably, in the case of $m = 2$, the risks of `Fused` were even better than `PO` for large values of n. A possible reason for the better performance of the fused lasso is that the sigmoidal function can be well approximated by a piecewise constant function near the boundaries. On the other hand, for piecewise cubic functions, `Neariso` performed slightly better than the other two estimators for small values of n.

6.3. Geological data

We conducted experiments on GPS data related to a seismological phenomenon reported by Roggers and Dragert (2003). The aim here is to investigate the performance of the nearly-isotonic type estimators on real-world data in which piecewise monotone approximations have already been justified in the previous work. For the signal y, we used the difference of the east-west components of GPS measurements between two observatories, which are located in Victoria (British Columbia, Canada) and Seattle (United States). The GPS data is provided by Melbourne et al. (2018). The top panel in Figure 6 shows the plot. The data period starts on January 1, 2010, and ends on December 2, 2017. After removing missing records, the size of the signal is $n = 2885$. The increasing trend of the signal is considered to be caused by the subduction process at the plate boundary. We can also see periodic reversals in the signal, and the entire signal may be approximated by a piecewise monotone signal. Such reversals may be related to the seismological phenomenon so-called the episodic tremor and slip. According to Roggers and Dragert (2003), such slip events were observed in every 13 to 16 months in their data taken from 1997 to 2003.

GPS data contains several anomalous values. For the signal y considered above, most of the values y_i are between 20 and 50, except for a single outlier $y_{2344} = 139.34$. The behaviors of the estimators are extremely affected by the existence of such outliers. In our situation, we can manually remove the anomalous value (denoted by $\tilde{y}$). However, it is often difficult to distinguish outliers in practical situations. From this perspective, we also considered the robust M-estimation version of the nearly-isotonic regression defined as (34) with $\mathcal{L}(\theta; y) = \sum_{i=1}^{n} \ell_\delta(\theta_i - y_i)$. Here, ℓ_δ is the Huber loss:

$$\ell_\delta(u) := \begin{cases} \dfrac{1}{2}u^2 & (|u| \le \delta) \\ \delta|u| - \dfrac{1}{2}\delta^2 & (|u| > \delta) \end{cases},$$

which is commonly used in the robust regression literature.

We applied the nearly-isotonic regression (3) and its robust variant to the signals y and $\tilde{y}$ in the above. The tuning parameters λ were determined by the 5-fold cross-validation, and δ in the Huber loss was fixed as $\delta = 0.01$.

First, we consider the case where the outlier is removed manually. The second panel in Figure 6 shows the result for the cross-validated nearly-isotonic regression. The vertical lines denote the locations of downward jumps in the estimators. We can see that the period of jump clusters is about 12 to 14 months, which is close to that of the seismological slip events suggested by Roggers and Dragert (2003).

Next, we consider the case where the signal contains an outlier. In this case, the value of the squared loss largely depends on the error at the coordinate of the outlier. Then, the cross-validation may choose a large tuning parameter, and the resulting estimator becomes close to a monotone signal. The third panel

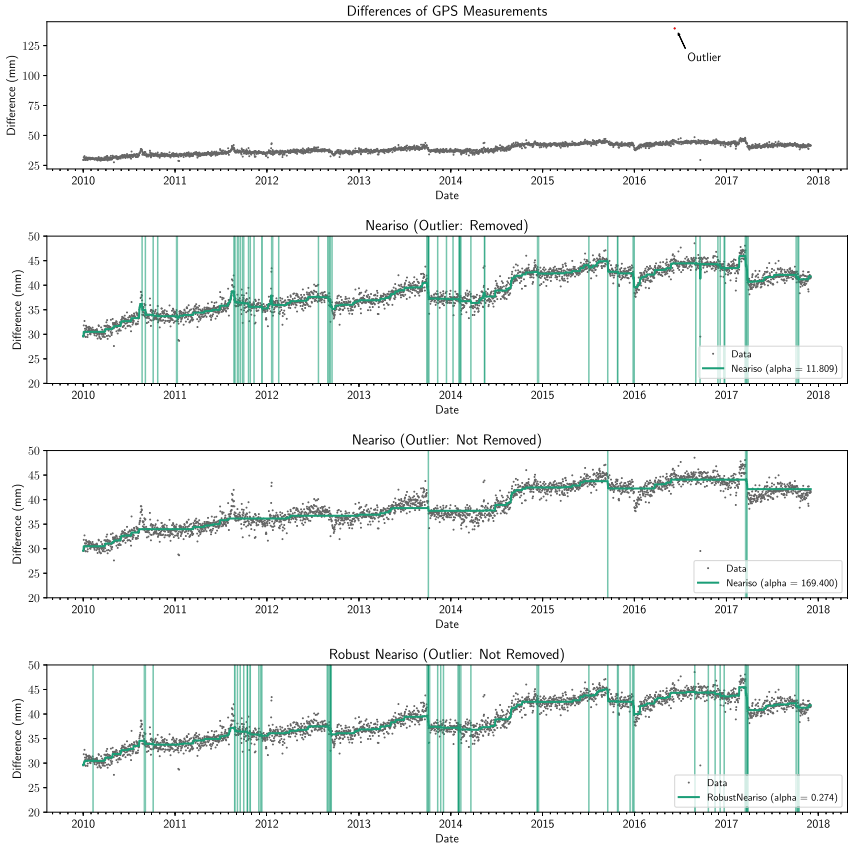

FIG 6. ***Nearly-isotonic type estimators applied to GPS data***. *See the text for details.*

in Figure 6 shows that the number of downward jumps is considerably less than the number that is expected from the known frequency of the slip events. Conversely, the fourth panel in Figure 6 shows that the robust version of the nearly-isotonic regression outputs similar clusters of change points as in the second panel.

7. Discussion

In this paper, we studied the problem of estimating piecewise monotone signals. The classical isotonic regression estimator cannot be applied in this setting because of the existence of arbitrarily large downward jumps. We derived the minimax risk lower bound over piecewise monotone signals with bounded upper total variations. The minimax rate is tight up to multiplicative constant because it can be achieved by a (computationally inefficient) model selection based estimator. Our main results show that the nearly-isotonic regression es-

timator achieves this rate under an additional growth condition. An advantage of the nearly-isotonic regression is that the estimator can be calculated efficiently on arbitrary directed graphs by parametric max-flow algorithms. The simulation results demonstrate that the nearly-isotonic regression has an almost similar convergence rate as the ideal estimator that knows the true partition.

7.1. Non-Gaussian noises

In this paper, we provided risk bound for the nearly-isotonic regression under the assumption that the noise distribution is Gaussian. However, in practice, this assumption is too restrictive. We here briefly discuss the risk bound with non-Gaussian error distributions.

Suppose that $\xi_1, \dots, \xi_n$ are i.i.d. random variables with $\mathbb{E}[\xi_1] = 0$ and $\mathrm{Var}(\xi_1) = \sigma^2$. Then, we can see that the "expectation bound" (20) holds with a different constant $C' > 0$. See Remark D.14 in the appendix for the key ingredients for the derivation. On the other hand, the "high-probability bound" (21) does not hold in general since it requires a more strong concentration property (i.e., the Gaussian concentration).

7.2. Future directions

An interesting direction for future work is to investigate the optimal rate of piecewise monotone regression on higher dimensional grids or general graphs. Recently, several researchers have analyzed the risk bounds for the isotonic regression estimators on two or more higher dimensional grid graphs (Chatteejee et al. 2018, Han et al. 2019). It is natural to ask whether one can construct a computationally efficient estimator that is adaptive to piecewise monotone vectors on a given graph. We believe that the nearly-isotonic type estimator (32) is a candidate. A major difficulty is to determine an appropriate graph topology. Given a partial order $\preceq$ on a set $V = [n]$, the corresponding isotonic regression estimator is uniquely determined. However, there are many directed acyclic graphs that correspond to partial order $\preceq$. Hence, the graph topology for the nearly-isotonic type estimators is not unique. To control the connectivity, it may be useful to introduce edge weightings proposed by Fan and Guan (2018).

Another direction is to develop a model selection method for least squares estimators over unbounded cones. We introduced sieves on the total variation in Section 5 to construct an estimator that is adaptive to piecewise monotone vectors. In practice, sieve-based methods can be computationally inefficient. Conversely, if the true vector θ^* is monotone, the isotonic regression automatically achieves the minimax rate with respect to the total variation. We conjecture that it is also possible to select the least squares estimator $\hat{\theta}_\Pi$ without using sieves. In particular, we leave it as an open question whether the adaptive risk bound is achieved by the penalized selection rule of the form (26).

Appendix A: Algorithms for nearly-isotonic estimators

In this section, we present algorithms for the nearly-isotonic regression and related estimators and discuss their computational complexities. Note that the main purpose of this section is to give a review of existing algorithms, and hence most results presented in this section are not new (except for Proposition A.1).

A.1. Penalized estimators

Here, we introduce two algorithms to solve the penalized form nearly-isotonic regression (3). In Section A.1.1, we introduce the solution path algorithm developed by Tibshirani et al. (2011). The advantage of the solution path algorithm is that it outputs the solutions $\hat{\theta}_\lambda$ for every $\lambda \geq 0$ simultaneously. However, the solution path algorithm cannot be applied to the estimators with general weights and graphs. In Section A.1.2, we provide another algorithm that outputs the exact solution for a single λ. The latter algorithm can be applied to the nearly-isotonic type estimators defined on any weighted directed graphs.

A.1.1. One-dimensional problem

The modified pool adjacent violators algorithm (modified PAVA, Tibshirani et al. (2011)) is the algorithm used to calculate the solution path for the problem (3). Here, we present a variant of the modified PAVA for the following weighted version of the estimator:

$$\hat{\theta}_\lambda = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \sum_{i-1}^{n} c_i(\theta_i - \theta_{i+1})_+ \right\}, \tag{30}$$

where $c_i > 0$ $(i = 1, 2, \ldots, n-1)$ are positive weight parameters. Letting $c_i = (x_{i+1} - x_i)^{-1}$, this formulation covers the nearly-isotonic regression for general increasing design points (24).

The derivation of Algorithm 1 is straightforward from the original paper of Tibshirani et al. (2011). We should note that the validity of this algorithm crucially depends on the property that the solution path is piecewise linear and "agglomerative". It is well known that the piecewise linearity of the solution path holds for many classes of regularization estimators (Rosset and Zhu 2007). We say that the solution path $\{\hat{\theta}_\lambda\}_{\lambda \geq 0}$ is *agglomerative* if it satisfies the following condition: if $\hat{\theta}_{\lambda,i} = \hat{\theta}_{\lambda,j}$ holds for some $\lambda = \lambda_0$, then the same equality holds for any $\lambda \geq \lambda_0$. For the constant weights ($c_i \equiv 1$), such agglomerative property was proved by Tibshirani et al. (2011). However, for general non-unitary edge weights ($c_i \neq 1$), this need not be true. Here, we provide the following proposition to ensure the agglomerative property for non-unitary edge weights.

Proposition A.1. The solution path of weighted nearly-isotonic regression (30) is piecewise linear and agglomerative if the edge weights satisfy the following

Algorithm 1: Modified Pool Adjacent Violators Algorithm (Tibshirani et al. 2011)

Input: $y \in \mathbb{R}^n$, $c_1, \ldots, c_{n-1} > 0$
Output: Set of finitely many breakpoints $\Lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_N\}$, solution path $\{\hat{\theta}_\lambda\}_{\lambda \in \Lambda}$

$\lambda_0 \leftarrow 0$, $\hat{\theta}_{\lambda_0} \leftarrow y$
Let Π_0 be the constant partition of $\hat{\theta}_{\lambda_0}$. Below, the solution $\hat{\theta}_{\lambda_i}$ is kept to be constant on Π_i.
for $i = 1, 2, \ldots$ **do**
  Let $k = |\Pi_{i-1}|$. Let $A_j = \{\tau_j, \tau_j + 1, \ldots, \tau_{j+1} - 1\}$ be the j-th element in the partition Π_{i-1}, and t_j be the value of $\hat{\theta}_{\lambda_{i-1}}$ on A_j $(j = 1, 2, \ldots, k)$.
  Set $s_0 = s_k = 0$ and $c_0 = 0$. Compute $s_j = 1_{\{t_j > t_{j+1}\}}$ for $j = 1, 2, \ldots, k-1$.
  Compute the slopes m_j $(j = 1, 2, \ldots, k)$ by

$$m_j = \frac{c_{\tau_j - 1} s_{j-1} - c_{\tau_{j+1}-1} s_j}{|A_j|}.$$

  Compute δ by

$$\delta = \min_{1 \le j \le k-1} \frac{t_{j+1} - t_j}{m_j - m_{j+1}}.$$

  If $\delta \le 0$, then terminate.
  $\lambda_i \leftarrow \lambda_{i-1} + \delta$.
  Set $\hat{\theta}_{\lambda_i}$ to be the piecewise constant vector whose values on A_j are $t_j + m_j \delta$ $(j = 1, 2, \ldots, k)$.
  Set Π_i to be the constant partition of $\hat{\theta}_{\lambda_i}$.
end

concavity condition.

$$c_{j-1} + c_{j+1} \le 2c_j \quad \text{for all } j = 0, 1, \ldots, n-2, \tag{31}$$

where we defined $c_0 := 0$. In particular, this condition implies that Algorithm 1 outputs the exact solution path.

The condition (31) demands that c_j can be written as $c_j = f(j)$ for some concave function $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $f(0) = 0$ and $f(x) > 0$ for all $x > 0$. In particular, for any $i \le j \le k$, we have

$$c_j \ge \frac{(k-j)c_i + (j-i)c_k}{k-i}$$

and

$$c_j \ge \frac{j}{k} c_k.$$

Proof sketch of Proposition A.1. We can prove the validity of Algorithm 1 by a similar argument as Tibshirani et al. (2011) if we assume the piecewise linearity and the agglomerative property. The piecewise linearity is already shown in Rosset and Zhu (2007). Hence, it remains to prove the agglomerative property

under the condition (31). To this end, we leverage the "agglomerative clustering condition" defined in Appendix D.6. In particular, we defer the details to Remark D.25 as well as Remark D.27. □

A.1.2. General graphs

Let $G = (V, E)$ be a directed graph with $V := [n]$. Suppose that each edge $(i, j) \in E$ is equipped with a positive weight $c_{(i,j)} > 0$. We define the *generalized nearly-isotonic regression* as

$$\hat{\theta}_{G,\lambda} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \mathcal{V}_G(\theta) \right\} \tag{32}$$

where $\mathcal{V}_G$ is a nearly-isotonic type penalty defined as

$$\mathcal{V}_G(\theta) := \sum_{(i,j) \in E} c_{(i,j)} (\theta_i - \theta_j)_+. \tag{33}$$

For any choices of G and c, $\mathcal{V}_G$ becomes a convex function. Clearly, the lower total variation $\mathcal{V}_-$ is a special case where $E = \{(i, i+1) : i = 1, 2, \ldots, n-1\}$ and $c_{(i,i+1)} \equiv 1$. Thus, (32) can be regarded as a generalization of the nearly-isotonic regression to general directed graphs.

The problem of the form (32) has been well studied in the optimization literature. In particular, we can see that solving (32) is equivalent to solving a certain parametrized family of minimum-cut problems. For detailed explanations of such an equivalence, see Obozinski and Bach (2016) and Chapter 8 in Bach (2013). Hence, (32) can be solved by the parametric max-flow algorithm (Gallo et al. 1989) that runs in $\mathrm{O}(n|E| \log \frac{n^2}{|E|})$. Conversely, it has been pointed out by Mairal et al. (2011) that, for many practical instances, some simplified variants of the parametric max-flow algorithm output the solution faster than the original algorithm by Gallo et al. (1989). We remark that Hochbaum and Queyranne (2003) also developed the relationship between the isotonic regression and the parametric max-flow algorithm.

Algorithm 2 shows the Divide-and-Conquer algorithm (Chapter 9 of Bach (2013)) that solves (32). In the inner loop, the algorithm recursively solves max-flow problems by defining smaller networks (Algorithm 3). See Figure 7 for examples of networks used in the first two recursions in the algorithm.

A.1.3. General convex loss functions

In practice, we are often interested in general convex loss functions other than the squared loss. Here, we consider a generalized problem of the following form:

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; y) + \lambda \mathcal{V}_G(\theta) \right\}, \tag{34}$$

Algorithm 2: Divide-and-Conquer algorithm for the generalized nearly-isotonic regression 32

Input: $y \in \mathbb{R}^V$, a directed graph $G = (V, E)$ with positive edge weights $\{c_{(i,j)}\}$, a tuning parameter $\lambda \geq 0$.
Output: The solution $\hat{\theta}_\lambda$ of (32)

1 Construct a flow network $\mathcal{N}$ by adding a source node s and a sink node t to the graph G.
2 Compute $\hat{\theta}_\lambda = \mathrm{Prox}_{\lambda F_\mathcal{N}}(y)$ according to Algorithm 3.

Algorithm 3: $\mathrm{Prox}_{\lambda F_\mathcal{N}}(y)$

Input: A flow network $\mathcal{N} = (V \cup \{s\} \cup \{t\}, E, c)$, $y \in \mathbb{R}^V$ and $\lambda > 0$.
Output: Proximal operator $\mathrm{Prox}_{\lambda F_\mathcal{N}}(y)$.

1 Let $\alpha \leftarrow \frac{1}{|V|}(\sum_{i \in V} y_i - \lambda F_\mathcal{N}(V))$, where $F_\mathcal{N}(V)$ is the capacity of the edge (s, t).
2 **if** $|V| = 1$ **then**
  return $\hat{\theta} = \alpha$
end
3 Find a subset $A \subseteq V$ minimizing the function $A \mapsto \lambda F_\mathcal{N}(A) - \sum_{i \in A} y_i + \alpha|A|$. Herein, $F_\mathcal{N}$ is the s-t cut function of the network $\mathcal{N}$. This step is equivalent to solving the max-flow problem defined by the flow network in Figure 7-(a).
4 **if** $\lambda F_\mathcal{N}(A) - \sum_{i \in A} y_i + \alpha|A| = 0$ **then**
  return $\hat{\theta} = \alpha 1_V$.
end
5 Let $\hat{\theta}_A \leftarrow \mathrm{Prox}_{\lambda F_{\mathcal{N}|A}}(y_A)$, where $\mathcal{N}|A$ is the reduction of $\mathcal{N}$ on A. The corresponding network is obtained by shrinking nodes $V \setminus A$ into the sink node t (Figure 7-(b)).
6 Let $\hat{\theta}_{V \setminus A} \leftarrow \mathrm{Prox}_{\lambda F_{\mathcal{N}^A}}(y_{V \setminus A})$, where $\mathcal{N}^A$ is the contraction of $\mathcal{N}$ by A. The corresponding network is obtained by shrinking nodes A into the source node s and adding $-F_\mathcal{N}(A)$ to the capacity of (s, t) (Figure 7-(c)).

where $\theta \mapsto \mathcal{L}(\theta; y)$ is a convex loss function for any $y \in \mathbb{R}^n$. As an example, this formulation contains the M-estimator in the regression setting $\mathcal{L}(\theta; y) = \frac{1}{2}\ell(y_i - \langle x_i, \theta \rangle)$, where $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$ $(i = 1, 2, \ldots, n)$ are the observed data and $\ell : \mathbb{R} \to \mathbb{R}$ is a convex function.

We can also obtain algorithms that output approximate minimizers of (34) as follows. First of all, note that Algorithm 2 outputs the *proximal operator* of the regularization term $\mathcal{V}_G(\theta)$. Once we have an oracle for the proximal operator, we can apply *proximal gradient methods* to solve (34). In particular, if $\mathcal{L}(\theta; y)$ is convex and smooth, the Fast Iterative Shrinkage Thresholding Algorithm (FISTA, Beck and Teboulle (2009)) outputs an $\mathrm{O}(\epsilon)$-optimal solution after $\mathrm{O}(\epsilon^{-2})$ evaluations of the proximal operator.

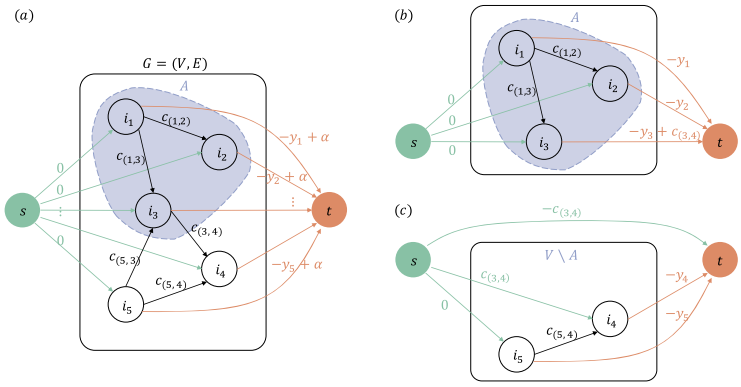

FIG 7. ***Flow networks in Algorithm 3.*** *Algorithm 3 requires to solve minimum s-t cut problems (or equivalently maximum flow problems) defined on certain flow networks. (a) A network that corresponds to the minimization problem in line 3. (b) A network that corresponds to the function* $B \mapsto \lambda F_{\mathcal{N}|A}(B) - y(B)$ *in line 5. (c) A network that corresponds to the function* $B \mapsto \lambda F_{\mathcal{N}^A}(B) - y(B)$ *in line 6. Note that we assumed* $\lambda = 1$ *in this example.*

A.2. Constrained estimators

Consider the following generalized version of the constrained form of nearly-isotonic regression (11):

$$\text{minimize } \|y - \theta\|_2^2 \quad \text{subject to } \sum_{(i,j)\in E} c_{(i,j)}(\theta_i - \theta_j)_+ \le \mathcal{V}. \tag{35}$$

Unlike the penalized estimators, it is difficult to find an exact solution of (35). However, since problem (35) is an instance of a quadratic programming problem, there are polynomial time algorithms to obtain approximate solutions. Here, we explain the *existence* of such algorithms. The following result is a direct application of Theorem 1 by Lee et al. (2018), which provides a convergence guarantee of a variant of cutting plane methods.

Proposition A.2. Suppose that $G = ([n], E)$ is a directed graph equipped with positive weights $c_{(i,j)}$ for every $(i, j) \in E$. Let $y \in \mathbb{R}^n$ be any vector and $\mathcal{V} > 0$. Then, for any $\epsilon > 0$, there exists a randomized algorithm that outputs $\tilde{\theta}$ satisfying

$$\mathcal{V}_G(\tilde{\theta}) := \sum_{(i,j)\in E} c_{(i,j)}(\tilde{\theta}_i - \tilde{\theta}_j)_+ \le \mathcal{V} + 2\epsilon \sum_{(i,j)\in E} c_{(i,j)}$$

and

$$\|y - \tilde{\theta}\|_2 \le \min_{\theta\in\mathbb{R}^n:\ \mathcal{V}_G(\theta)\le\mathcal{V}} \|y - \theta\|_2 + 2\epsilon\|y\|_2$$

with a probability of 0.99. The overall complexity of the algorithm is $\mathrm{O}((n + |E|)n^2 \log^{\mathrm{O}(1)} \frac{n}{\epsilon|E|})$.

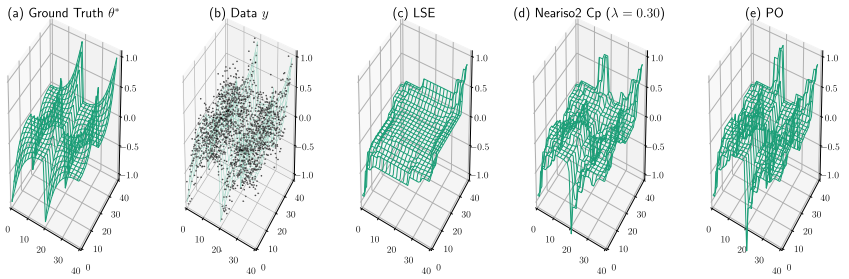

FIG 8. ***Examples of estimators for piecewise monotone matrices.*** *The true parameter θ^* is a 32×32 matrix that is monotone on each 16×16 segment. The bivariate isotonic regression (`LSE`) does not capture the piecewise monotone structure. The solution of the nearly-isotonic regression (`Neariso2`) seems to be close to the partition oracle (`PO`).*

Remark A.3. In practice, due to computational considerations, we recommend to use the penalized estimator (33) instead of the constrained estimator (35). For the penalized estimator, we empirically observed that Algorithm 2 runs sufficiently fast graphs with several hundreds of nodes. For the constrained estimator, Proposition A.2 theoretically guarantees polynomial time solvability of the constrained problem (35), whereas it does not provide a practical algorithm.

Appendix B: Supplemental experiments

To understand the behavior of the nearly-isotonic regression in more generic settings, we present additional simulation results for the nearly-isotonic regression on general graphs (32). Here, we consider the problem of estimating piecewise monotone signals on two-dimensional grids.

We say that an $n_1 \times n_2$ matrix θ is monotone if $\theta_{ij} \le \theta_{kl}$ whenever $i \le k$ and $j \le l$. In other words, θ is monotone if it has no order-violating edges in the two-dimensional grid graph $G_2 = (V_2, E_2)$, where $V_2 = [n_1] \times [n_2]$ is the set of all subscripts (i, j) and

$$\begin{aligned} E_2 :=& \{((i,j),(i,j+1)) \ : \ 1 \le i \le n_1, 1 \le j \le n_2 - 1\} \\ & \cup \{((i,j),(i+1,j)) \ : \ 1 \le i \le n_1 - 1, 1 \le j \le n_2\}. \end{aligned}$$

We say that θ is piecewise monotone if there is a partition Π of V such that, for each $A \in \Pi$, A is a weakly connected component of G_2 and θ_A has no order-violating edges in the induced subgraph. For simplicity of experimental settings, we here only consider "block" type partitions, i.e., we say that Π is of block type if it can be represented as a product of two partitions of the two coordinates. The left panel in Figure 8 is an example of two-dimensional piecewise monotone signals on a block type partition.

We compare the following three estimators:

- `LSE`: The bivariate isotonic regression (see e.g., Robertson et al. (1988)).

- `Neariso2`: The two-dimensional nearly-isotonic regression with C_p-tuned parameter.
- `PO`: The bivariate isotonic regression applied to the true partition.

For monotone matrices, Chatteejee et al. (2018) proved that `LSE` is minimax rate optimal with respect to $n = n_1 n_2$. Hence, the partition oracle estimator `PO` can be regarded as an ideal benchmark that is minimax optimal over piecewise monotone matrices. On the other hand, if the true matrix θ^* is piecewise monotone, the risk of `LSE` can be arbitrarily large for the same reason as Proposition 3.3. `Neariso2` is the special case of the generalized nearly-isotonic regression (32) applied to the graph G_2 defined above. `Neariso2` was originally discussed in Tibshirani et al. (2011), but no experimental results have been presented. Figure 8 shows examples of the solutions of the three estimators.

We construct an $n \times n$ matrix θ^* as follows: We define a $k \times k$ small monotone matrix U, and then we define θ^* as an $mk \times mk$ block matrix by repeating U for m times both in rows and columns (thus $n = mk$). We choose the small matrix $U = (U_{ij})$ from

$$U_{ij}^{\mathrm{cubic2d}} = (x_i + x_j - 1)^3$$

or

$$U_{ij}^{\mathrm{cubic1d}} = (2x_i - 1)^3,$$

where we write $x_i = \frac{i-1}{k-1}$ for $i = 1, 2, \ldots, k$. With the former choice, θ^* becomes an m^2-piecewise monotone matrix. With the latter choice, θ^* becomes an m-piecewise monotone matrix such that θ^*_{ij} does not depend on j.

We generated noisy observations y by adding independent Gaussian noises $\xi_{ij} \sim N(0, (0.25)^2)$ to every entries of θ^*. To estimate the MSE, we used 500 replications of the data. Figure 9 shows the results. Clearly, the risks of `LSE` (blue triangles) are much larger than those of the other two estimators. `Neariso2` (green circles) has slightly larger risks compared to `PO` (magenta squares), while their slopes seem to be close.

To visualize convergence rates, we fit the risks of `PO` by monomials $\propto n^{-a}$ ($a > 0$), and plotted as dashed lines in Figure 9. The values of the exponent a are respectively as follows: 0.58 (`cubic2d`, $m = 2$); 0.56 (`cubic2d`, $m = 4$); 0.50 (`cubic1d`, $m = 2$); 0.45 (`cubic2d`, $m = 4$). We should note that, in monotone matrix estimation, the theoretical convergence rate of `LSE` is known to be $\tilde{\mathrm{O}}(n^{-1/2})$ (Chatteejee et al. 2018).

Appendix C: Proofs in Section 3

C.1. Proof of Proposition 3.2

Let Θ be either $\tilde{\Theta}_n(m, \mathcal{V})$ or $\Theta_n(m, \mathcal{V})$, which are defined in Definition 3.1. The minimax lower bound (10) is proved by combining the following two lower bounds:

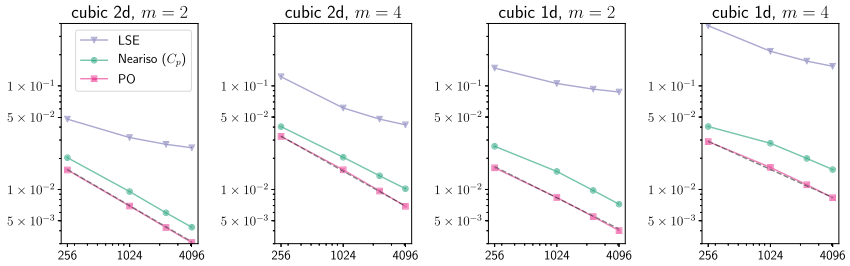

FIG 9. ***The risks in piecewise monotone matrix estimation***. *See the text for details.*

(i) **(Lower bound for monotone vectors (Zhang 2002, Chatterjee et al. 2015))** Let $\mathcal{K}(\mathcal{V}) = \{\theta \in K_n^{\uparrow} : \mathcal{V}(\theta) \le \mathcal{V}\}$ be the set of monotone vectors with bounded total variations. There is a universal constant $C_1 > 0$ such that for any estimator $\hat{\theta}$,

$$\sup_{\theta^* \in \mathcal{K}(\mathcal{V})} \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|_2^2 \ge C_1 \left(\frac{\sigma^2 \mathcal{V}}{n} \right)^{2/3}.$$

(ii) **(Lower bound for piecewise constant vectors)** Let $\mathcal{C}(m)$ be the set of m-piecewise constant vectors in $\mathbb{R}^n$, i.e., $\theta \in \mathcal{C}(m)$ if $|\{i : \theta_i \ne \theta_{i+1}\}| \le m - 1$. The minimax lower bound over $\mathcal{C}(m)$ can be related to sparse estimation as follows. Let X be an $n \times n$ matrix whose (i, j) entries are given as $1_{\{i \ge j\}}$. Then, $\mathcal{C}(m)$ contains the set $\{\theta = X\beta : \|\beta\|_0 \le m\}$, and the lower bound for the minimax risk over $\mathcal{C}(m)$ follows from the well-known results for ℓ_0 balls (e.g., Raskutti et al. (2011), Theorem 3-(b)). In particular, for any $m \ge 3$, the following lower bound is presented in Gao et al. (2017):

$$\sup_{\theta^* \in \mathcal{C}(m)} \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|_2^2 \ge C_2 \frac{\sigma^2 m}{n} \log \frac{en}{m},$$

where $C_2 > 0$ is a universal constant.

It remains to show that Θ contains $\mathcal{K}(\mathcal{V})$ and $\mathcal{C}(m)$. $\mathcal{C}(m) \subseteq \Theta$ is obvious because an m-piecewise constant vector is also an m-piecewise monotone vector such that the piecewise total variations are zero. From the definition, it is also clear that $\mathcal{K}(\mathcal{V}) \subseteq \tilde{\Theta}_n(m, \mathcal{V})$. If $\theta \in \mathcal{K}(\mathcal{V})$, the jumps $\theta_{i+1} - \theta_i$ that strictly exceeds $\mathcal{V}/m$ cannot occur more than $m - 1$ times. Hence, we can choose a partition Π with $|\Pi| \le m$ so that each $A \in \Pi$ does not contain such large jumps, which implies that $\theta \in \Theta_n(m, \mathcal{V})$.

C.2. Proof of Proposition 3.3

The following theorem in the seminal paper of Chatterjee (2014) provides useful upper and lower bounds for the risk of the least square estimator over any closed convex set K.

Theorem C.1 (Chatterjee (2014), Corollary 1.2)**.** Let $K \subseteq \mathbb{R}^n$ be any closed convex set, and let $\hat{\theta}_K$ denote the least squares estimator over K. For any $\theta^* \in \mathbb{R}^n$, define the function $g_{\theta^*} : \mathbb{R}_+ \to \mathbb{R} \cup \{-\infty\}$ as

$$g_{\theta^*}(t) := \mathbb{E}_{Z \sim N(0,\sigma^2 I_n)} \left[\sup_{\theta \in K : \|\theta - \theta^*\|_2 \le t} \langle Z, \theta - \theta^* \rangle \right] - \frac{t^2}{2}.$$

Here, if the set $\{\theta \in K : \|\theta - \theta^*\|_2 \le t\}$ is empty, we define $g_{\theta^*}(t) = -\infty$. Then, g_{θ^*} is strictly concave for $t \ge \mathrm{dist}(\theta^*, K)$ and has a unique maximizer t_{θ^*}. Moreover, there are universal constants $C_1, C_2 > 0$ such that

$$\frac{1}{n} \max\left\{ t_{\theta^*}^2 - C_1 t_{\theta^*}^{3/2}, 0 \right\} \le \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta}_K - \theta^*\|_2^2 \le \frac{C_2}{n} \max\left\{ t_{\theta^*}^2, \sigma^2 \right\}. \tag{36}$$

To prove Proposition 3.3, we use the lower bound in (36). Note that for a sufficiently large $t_0 > 0$, $t \mapsto t^2 - Ct^{3/2}$ is a strictly increasing in $t \in [t_0, \infty)$. For any n and σ^2, choose $t \ge t_0$ so that $t^2 - Ct^{3/2} \ge n\sigma^2$. Then, for any θ^* such that $\mathrm{dist}(\theta^*, K) \ge t$, we have

$$\frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta}_K - \theta^*\|_2^2 \ge \frac{1}{n}(t_{\theta^*}^2 - C_1 t_{\theta^*}^{3/2}) \ge \frac{1}{n}(t^2 - C_1 t^{3/2}) \ge \sigma^2.$$

Remark C.2. We should note that the above proof is valid for *any* closed convex set K. For the specific choice of $K = K_n^{\uparrow}$, the lower bound of t_{n,σ^2} used in the proof can be quite conservative. In practice, the risk of the isotonic regression estimator can be larger than σ^2 under a smaller value of ℓ_2-misspecification error.

Appendix D: Proofs in Section 4

D.1. Preliminaries

To state the results for risk upper bounds, we first introduce some quantities related to Gaussian processes.

Definition D.1. Let C be a closed convex set in $\mathbb{R}^n$. Let $\mathbb{E}$ denote the expectation with respect to an isotropic Gaussian random variable $Z \sim N(0, I_n)$.

(i) The *Gaussian width* of C is defined as

$$w(C) := \mathbb{E}\left[\sup_{\theta \in C} \langle Z, \theta \rangle \right].$$

(ii) The *Gaussian mean squared distance* is defined as

$$\mathbf{D}(C) := \mathbb{E}[\mathrm{dist}^2(Z, C)],$$

where $\mathrm{dist}(z, C) := \inf_{x \in C} \|x - z\|_2$.

(iii) Suppose that C is a convex cone. The *statistical dimension* of C is defined as

$$\delta(C) := \mathbb{E}\left[\left(\sup_{\theta \in C : \|\theta\|_2 \leq 1} \langle Z, \theta \rangle\right)^2\right].$$

We present some historical remarks on these definitions. The three quantities in Definition D.1 can be interpreted as complexity measures for the subset C in the Euclidean space. The Gaussian width has been well studied in convex geometry, signal processing, high-dimensional statistics, and empirical process theory; See e.g., Section 7.8 in Vershynin (2018) for a literature review. The definition of the Gaussian mean squared distance is due to Oymak and Hassibi (2016). As we will see in Lemma D.4 below, the Gaussian mean squared distance is useful to provide the risk bounds for proximal denoising estimators. The statistical dimension was defined in Amelunxen et al. (2014). Recently, Bellec (2018) pointed out that the statistical dimension characterizes the adaptive risk bounds for some shape restricted estimators including the isotonic regression and the convex regression.

As suggested by the definitions, these three quantities are closely related to each other. In particular, if C is a convex cone, these are comparable as follows.

Proposition D.2. Let C be a closed convex cone.

(i) (Amelunxen et al. (2014), Proposition 10.2) Let $S_{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^n$. Then, we have $w^2(C \cap S_{n-1}) \leq \delta(C) \leq w^2(C \cap S_{n-1}) + 1$.

(ii) (Amelunxen et al. (2014), Proposition 3.1) Let C° be the polar cone of C defined as

$$C^\circ := \{x \in \mathbb{R}^n : \langle x, z \rangle \leq 0 \text{ for all } z \in C\}.$$

Then, we have $\mathbf{D}(C) = \delta(C^\circ)$.

Now, we introduce two general results for risk bounds for general projection estimators and proximal denoising estimators.

Let K be a closed convex set in $\mathbb{R}^n$, and define the projection estimator onto K as $\hat{\theta}_K = \mathrm{argmin}_{\theta \in K} \|y - \theta\|_2$. Bellec (2018) proved the following oracle inequality that relates the risk of the projection estimator to the statistical dimension of the tangent cone of K. Here, the tangent cone $T_K(\theta)$ of K at $\theta \in K$ is defined as

$$T_K(\theta) := \mathrm{closure}(\{t(z - \theta) : t \geq 0, z \in K\}).$$

Lemma D.3 (Bellec (2018), Corollary 2.2)**.** Let $\theta^* \in \mathbb{R}^n$ be any vector, and suppose that the observation y is drawn according to $N(\theta^*, \sigma^2 I_n)$. Then, we have the following risk bound:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_K - \theta^*\|_2^2 \leq \inf_{\theta \in K}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + \frac{\sigma^2}{n}\delta(T_K(\theta))\right\}.$$

Moreover, for any $\eta \in (0,1)$, the inequality

$$\frac{1}{n}\|\hat{\theta}_K - \theta^*\|_2^2 \le \inf_{\theta \in K}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + \frac{2\sigma^2}{n}\delta(T_K(\theta))\right\} + \frac{4\sigma^2\log(\eta^{-1})}{n}$$

holds with probability at least $1-\eta$.

Next, we provide a general result for proximal denoising estimators. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and $\lambda \ge 0$. We define the proximal denoising estimator $\hat{\theta}_\lambda$ as

$$\hat{\theta}_\lambda := \operatorname*{argmin}_{\theta \in \mathbb{R}^n}\left\{\frac{1}{2}\|y-\theta\|_2^2 + \sigma\lambda f(\theta)\right\}. \tag{37}$$

The class of proximal denoising estimators contains the soft-thresholding estimator (Donoho et al. 1992), the total variation regularization (Rudin et al. 1992), the trend filtering (Kim et al. 2009) and the nearly-isotonic regression (Tibshirani et al. 2011). Oymak and Hassibi (2016) pointed out that the risk bound of proximal denoising estimators can be characterized by the Gaussian mean squared distance of the set $\lambda\partial f(\theta^*)$. Remarkably, based on this technique, Guntuboyina et al. (2020) proved sharp adaptation results for the trend filtering estimators. The following oracle inequality can be regarded as a generalization of Theorem 2.2 in Oymak and Hassibi (2016). For the sake of completeness, we also provide its proof below.

Lemma D.4. Let $\theta^* \in \mathbb{R}^n$ be any vector, and suppose that the observation y is drawn according to $N(\theta^*, \sigma^2 I_n)$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and let $\hat{\theta}_\lambda$ denote the proximal denoising estimator defined as (37). Then, we have

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \le \inf_{\theta \in \mathbb{R}^n}\left\{\frac{1}{n}\|\theta-\theta^*\|_2^2 + \frac{\sigma^2}{n}\mathbf{D}(\lambda\partial f(\theta))\right\}. \tag{38}$$

Moreover, for any $\eta \in (0,1)$, the inequality

$$\frac{1}{n}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \le \inf_{\theta \in \mathbb{R}^n}\left\{\frac{1}{n}\|\theta-\theta^*\|_2^2 + \frac{2\sigma^2}{n}\mathbf{D}(\lambda\partial f(\theta^*))\right\} + \frac{16\sigma^2\log(\eta^{-1})}{n} \tag{39}$$

holds with probability at least $1-\eta$.

Proof. Below, we write $\hat{\theta} := \hat{\theta}_\lambda$. To prove (38), it suffices to show that we have almost surely

$$\|\hat{\theta} - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2 \le \sigma^2\mathbf{D}(\lambda\partial f(\theta))$$

for any fixed vector $\theta \in \mathbb{R}^n$. We will assume $\theta \ne \hat{\theta}$ because otherwise the inequality is trivial.

From the first order optimality condition of the convex minimization problem (37), we have

$$\langle \theta - \hat{\theta}, y - \hat{\theta}\rangle \le \sigma\lambda(f(\theta) - f(\hat{\theta})) \quad \text{for any } \theta \in \mathbb{R}^n.$$

See Lemma 6.1 in van de Geer (2015) for a formal proof. Using the elementary fact that $2\langle u, v\rangle = \|u\|_2^2 + \|v\|_2^2 - \|u-v\|_2^2$ and substituting $y = \theta^* + \sigma z$, we have

$$\|\hat{\theta} - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2 \leq 2\sigma\lambda(f(\theta) - f(\hat{\theta})) - 2\sigma\langle z, \theta - \hat{\theta}\rangle - \|\theta - \hat{\theta}\|_2^2. \tag{40}$$

Now, take $v \in \partial f(\theta)$ arbitrarily. From the definition of the subgradient, we have

$$f(\theta) - f(\hat{\theta}) \leq \langle v, \theta - \hat{\theta}\rangle.$$

Hence, the right-hand side of (40) is bounded from above by

$$\begin{aligned}
&2\sigma\langle \lambda v - z, \theta - \hat{\theta}\rangle - \|\theta - \hat{\theta}\|_2^2 \\
&= 2\sigma \left\langle \lambda v - z, \frac{\theta - \hat{\theta}}{\|\theta - \hat{\theta}\|_2} \right\rangle \|\theta - \hat{\theta}\|_2 - \|\theta - \hat{\theta}\|_2^2 \\
&\leq \sigma^2 \left\langle \lambda v - z, \frac{\theta - \hat{\theta}}{\|\theta - \hat{\theta}\|_2} \right\rangle^2 \quad (\because 2ab - b^2 \leq a^2) \\
&\leq \sigma^2 \|\lambda v - z\|_2^2 \quad (\because \text{The Cauchy–Schwarz inequality}).
\end{aligned}$$

Since the choice of $v \in \partial f(\theta)$ is arbitrary, we have

$$\|\hat{\theta} - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2 \leq \sigma^2 \inf_{v \in \partial f(\theta)} \|\lambda v - z\|_2^2 = \sigma^2 \mathrm{dist}^2(z, \lambda \partial f(\theta)). \tag{41}$$

By taking the expectation of both sides, (38) is proved.

To prove the high-probability bound (39), we use the well-known Gaussian concentration inequality (see e.g., Theorem 5.6 in Boucheron et al. (2013)); for any L-Lipschitz function $h : \mathbb{R}^n \to \mathbb{R}$ and $\eta \in (0, 1)$, we have

$$\Pr_{Z \sim N(0, I_n)} \left\{ h(Z) - \mathbb{E}[h] \geq \sqrt{2L^2 \log \eta^{-1}} \right\} \leq \eta.$$

In fact, the map $z \mapsto \mathrm{dist}(z, \lambda \partial f(\theta))$ is a 2-Lipschitz function because, for any $z_1, z_2 \in \mathbb{R}^n$, we have

$$\begin{aligned}
|\mathrm{dist}(z_1, \lambda \partial f(\theta)) - \mathrm{dist}(z_2, \lambda \partial f(\theta))| &\leq \|(z_1 - P(z_1)) - (z_2 - P(z_2))\|_2 \\
&\leq 2\|z_1 - z_2\|_2,
\end{aligned}$$

where P is the orthogonal projection map onto the set $\lambda \partial f(\theta)$. Now, we take $\bar{\theta}$ as

$$\bar{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \|\theta - \theta^*\|_2^2 + \sigma^2 \left(\sqrt{\mathbf{D}(\lambda \partial f(\theta))} + \sqrt{8 \log \eta^{-1}} \right)^2 \right\}.$$

Combining (41) and the Gaussian concentration applied for $\theta = \bar{\theta}$, we have the desired result. □

D.2. Risk bounds for constrained estimators (Proof of Theorem 4.1)

In this subsection, we provide the proof of Theorem 4.1 as an application of Lemma D.3. To this end, we have to evaluate the statistical dimension of the tangent cone of a convex set

$$K_{-}(\mathcal{V}) := \{\theta \in \mathbb{R}^n : \mathcal{V}_{-}(\theta) \le \mathcal{V}\} = \left\{\theta \in \mathbb{R}^n : \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})_+ \le \mathcal{V}\right\}. \quad (42)$$

It is not surprising that the analysis of the tangent cone of $K_{-}(\mathcal{V})$ goes very similar to that of the set with bounded total variation $K(\mathcal{V}) = \{\theta \in \mathbb{R}^n : \mathcal{V}(\theta) \le \mathcal{V}\}$ in Guntuboyina et al. (2020). Our goal is to show the following upper bound for the statistical dimension:

Proposition D.5. Suppose that θ is a vector with $\mathcal{V}_{-}(\theta) = \mathcal{V}$. Then, there exists a universal constant $C > 0$ such that

$$\delta(T_{K_{-}(\mathcal{V})}(\theta)) \le Cn \left\{ \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)} \right\},$$

where $M(\theta)$ is defined in (13).

We briefly outline the proof for this result. We divide the proof into four steps: First, we provide some useful characterizations of the tangent cone. Second, we decompose the tangent cone into finitely many pieces so that the Gaussian widths become easy to evaluate. Third, we provide the concrete upper bounds the Gaussian widths of these pieces. Lastly, we combine the upper bounds and apply Lemma D.3 to complete the proof.

Step 1: Characterizing the tangent cone If $\mathcal{V}_{-}(\theta) < \mathcal{V}$, θ is contained in the interior of $K_{-}(\mathcal{V})$, and the tangent cone becomes the entire Euclidean space $\mathbb{R}^n$. Hereafter, we assume that θ lies on the boundary of $K_{-}(\mathcal{V})$, that is, $\mathcal{V}_{-}(\theta) = \mathcal{V}$. Let us recall the definition of the sign of jumps w_i in (12). Roughly speaking, the tangent cone of $K_{-}(\mathcal{V})$ is characterized by the sign of jumps.

Lemma D.6. Let θ be a vector in $\mathbb{R}^n$ such that $\mathcal{V}_{-}(\theta) = \mathcal{V}$. Let $\Pi = \{B_1, B_2, \dots, B_{k'}\}$ be any connected refinement [1] of the constant partition $\Pi_{\mathrm{const}}(\theta)$ of θ. Let $1 = \tau_1 < \tau_2 < \cdots < \tau_{k'} < \tau_{k'+1} = n + 1$ be a sequence such that $B_i = \{\tau_i, \tau_i + 1, \dots, \tau_{i+1} - 1\}$ for any $i \in \{1, 2, \dots, k'\}$. We define the signs $w_2, w_3, \dots, w_{k'} \in \{0, 1\}$ as

$$w_i = \begin{cases} 1 & \text{if } \theta_{\tau_i - 1} > \theta_{\tau_i} \\ 0 & \text{if } \theta_{\tau_i - 1} < \theta_{\tau_i} \\ \text{arbitrary value in } \{0, 1\} & \text{if } \theta_{\tau_i - 1} = \theta_{\tau_i} \end{cases}.$$

[1] Here, we say that Π is a connected refinement of another connected partition Π' if, for any $B \in \Pi$, there exists a unique element $A \in \Pi'$ such that $B \subseteq A$.

For any Π and $w_2, w_3, \ldots, w_{k'}$ taken as above, we define a convex cone $T(\Pi, w)$ as

$$T(\Pi, w) = \left\{ v \in \mathbb{R}^n : \sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(v_{B_i}) \le \sum_{i=2}^{k'} w_i (v_{\tau_i} - v_{\tau_i - 1}) \right\}, \tag{43}$$

where $\mathcal{V}_-^{B_i}(v_{B_i})$ is the lower total variation for the restricted vector v_{B_i}. Then, for the tangent cone $T_{K_-(\mathcal{V})}(\theta)$, we have the followings:

(i) If $\Pi = \Pi_{\mathrm{const}}(\theta)$, then $T_{K_-(\mathcal{V})}(\theta) = T(\Pi, w)$.
(ii) If Π is a connected refinement of $\Pi_{\mathrm{const}}(\theta)$ and w is taken arbitrarily as above, then $T_{K_-(\mathcal{V})}(\theta) \subseteq T(\Pi, w)$.

Proof. First, we show that $T_{K_-(\mathcal{V})}(\theta) \subseteq T(\Pi, w)$. By the definition of the tangent cone, it suffices to show that $v := z - \theta \in T(\Pi, w)$ holds for any $z \in K_-(\mathcal{V})$. Note that θ is constant on every $B_i \in \Pi$ since Π is finer than the constant partition of θ. Since the lower total variation is not changed by adding any constant value to each coordinates, we have $\mathcal{V}_-^{B_i}(z_{B_i} - \theta_{B_i}) = \mathcal{V}_-^{B_i}(z_{B_i})$. Then, we have

$$\begin{aligned}
&\sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(v_{B_i}) - \sum_{i=2}^{k'} w_i (v_{\tau_i} - v_{\tau_i - 1}) \\
&= \sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(z_{B_i}) + \sum_{i=2}^{k'} w_i (z_{\tau_i - 1} - z_{\tau_i}) - \sum_{i=2}^{k'} w_i (\theta_{\tau_i - 1} - \theta_{\tau_i}) \\
&\le \underbrace{\sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(z_{B_i}) + \sum_{i=2}^{k'} (z_{\tau_i - 1} - z_{\tau_i})_+}_{= \mathcal{V}_-(z) \le \mathcal{V}} - \underbrace{\sum_{i=2}^{k'} w_i (\theta_{\tau_i - 1} - \theta_{\tau_i})}_{= \mathcal{V}_-(\theta) = \mathcal{V}} \\
&\le 0,
\end{aligned}$$

which proves $v \in T(\Pi, w)$ and hence (ii).

Next, we prove that $T(\Pi, w) \subseteq T_{K_-(\mathcal{V})}(\theta)$ under the assumption

$$\Pi = \Pi_{\mathrm{const}}(\theta) = \{B_1, B_2, \ldots, B_k\}.$$

In this case, the definition of $w_2, \ldots, w_k$ coincides that in (12). Fix any $v \in T(\Pi, w)$. We want to show that z is obtained as $v = t(z - \theta)$ for some $t > 0$ and $z \in K_-(\mathcal{V})$. To this end, we check that there exists a (sufficiently small) $t^{-1} > 0$ such that $\theta + t^{-1} v \in K_-(\mathcal{V})$. Here, we have

$$\begin{aligned}
\mathcal{V}_-(\theta + t^{-1} v) &= \sum_{i=1}^{k} \mathcal{V}_-^{B_i}(\theta_{B_i} + t^{-1} v_{B_i}) \\
&\quad + \sum_{i=2}^{k} ((\theta_{\tau_i - 1} + t^{-1} v_{\tau_i - 1}) - (\theta_{\tau_i} + t^{-1} v_{\tau_i}))_+ \\
&= t^{-1} \sum_{i=1}^{k} \mathcal{V}_-^{B_i}(v_{B_i}) + \sum_{i=2}^{k} ((\theta_{\tau_i - 1} + t^{-1} v_{\tau_i - 1}) - (\theta_{\tau_i} + t^{-1} v_{\tau_i}))_+.
\end{aligned}$$

Recall that $w_2, \dots, w_k$ are chosen so that $(\theta_{\tau_i-1} - \theta_{\tau_i})_+ = w_i(\theta_{\tau_i-1} - \theta_{\tau_i})$. We can choose sufficiently small $t^{-1} > 0$ so that

$$((\theta_{\tau_i-1} + t^{-1}v_{\tau_i-1}) - (\theta_{\tau_i} + t^{-1}v_{\tau_i}))_+ = w_i((\theta_{\tau_i-1} + t^{-1}v_{\tau_i-1}) - (\theta_{\tau_i} + t^{-1}v_{\tau_i}))$$

for every $i = 2, 3, \dots, k$. Indeed, if we choose $t^{-1} > 0$ so that

$$t^{-1}|v_{\tau_i-1} - v_{\tau_i}| < \theta_{\tau_i-1} - \theta_{\tau_i} \quad \text{for every } i = 2, 3, \dots, k,$$

the signs of θ do not change by adding $t^{-1}v$. Consequently, we have

$$\begin{aligned}
\mathcal{V}_-(\theta + t^{-1}v) &= t^{-1}\sum_{i=1}^{k} \mathcal{V}_-^{B_i}(z_{B_i}) \\
&\quad + \sum_{i=2}^{k} w_i((\theta_{\tau_i-1} + t^{-1}v_{\tau_i-1}) - (\theta_{\tau_i} + t^{-1}v_{\tau_i})) \\
&= \mathcal{V}_-(\theta) + t^{-1}\left\{\sum_{i=1}^{k} \mathcal{V}_-^{B_i}(v_{B_i}) + \sum_{i=2}^{k} w_i(v_{\tau_i-1} - v_{\tau_i})\right\} \\
&\le \mathcal{V}_-(\theta) = \mathcal{V}.
\end{aligned}$$

This proves that $T(\Pi, w) \subseteq T_{K_-(\mathcal{V})}(\theta)$ and hence (i). □

From Proposition D.2-(i), we can bound the statistical dimension by the Gaussian width as follows:

$$\delta(T_{K_-(\mathcal{V})}(\theta)) \le w^2(T_{K_-(\mathcal{V})}(\theta) \cap S_{n-1}) + 1 \le w^2(T_{K_-(\mathcal{V})}(\theta) \cap B_n) + 1.$$

Here, $B_n := \{v \in \mathbb{R}^n : \|v\|_2 \le 1\}$ is the unit ball in $\mathbb{R}^n$. Hence, it suffices to consider the set $T_{K_-(\mathcal{V})}(\theta) \cap B_n$. In analogy to Lemma B.2 in Guntuboyina et al. (2020), we obtain the following characterization of this set.

Lemma D.7. Let θ be a vector in $\mathbb{R}^n$ such that $\mathcal{V}_-(\theta) = \mathcal{V}$. Let $\Pi = \{B_1, B_2, \dots, B_{k'}\}$ be any connected refinement of $\Pi_{\mathrm{const}}(\theta)$. Define the signs $w_2, w_3, \dots, w_{k'}$ as in Lemma D.6, and let $w_1 = w_{k'+1} = 0$. Then, for every $v \in T_{K_-(\mathcal{V})}(\theta)$ with $\|v\|_2 \le 1$, there exists indices $\ell_1 \in B_1, \ell_2 \in B_2, \dots, \ell_{k'} \in B_{k'}$ such that

$$\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) \le \left(\sum_{i=1}^{k'} \frac{1}{|B_i|} 1_{\{w_i \ne w_{i+1}\}}\right)^{\frac{1}{2}}, \tag{44}$$

where we define $\Gamma_i(v, \ell_i)$ as

$$\Gamma_i(v, \ell_i) := \mathcal{V}_-^{B_i}(v_{B_i}) - w_i(v_{\tau_i} - v_{\ell_i}) - w_{i+1}(v_{\ell_i} - v_{\tau_{i+1}-1}) \quad \text{for } i = 1, 2, \dots, k'. \tag{45}$$

Proof. Fix $v \in T_{K_-(\mathcal{V})}(\theta) \cap B_n$. By Lemma D.6, we have

$$\sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(v_{B_i}) \le \sum_{i=2}^{k'} w_i(v_{\tau_i} - v_{\tau_i-1}) = \sum_{i=1}^{k'+1} w_i(v_{\tau_i} - v_{\tau_i-1}). \tag{46}$$

Let $\ell_1 \in B_1, \ell_2 \in B_2, \dots, \ell_{k'} \in B_{k'}$ be indices which will be specified later. Defining $\Gamma_i(v, \ell_i)$ as in (45), we can rewrite (46) as

$$\begin{aligned}\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) &\le \sum_{i=1}^{k'} w_i(v_{\ell_i} - v_{\tau_i}) + \sum_{i=1}^{k'} w_{i+1}(v_{\tau_{i+1}-1} - v_{\ell_i}) + \sum_{i=1}^{k'+1} w_i(v_{\tau_i} - v_{\tau_i - 1}) \\ &= \sum_{i=1}^{k'} (w_i - w_{i+1}) v_{\ell_i} \\ &\le \sum_{i=1}^{k'} 1_{\{w_i \neq w_{i+1}\}} |v_{\ell_i}| \end{aligned} \tag{47}$$

Now, let t_i denote the ℓ_2 norm of v_{B_i} for $i = 1, 2, \dots, k'$. By the assumption, $\sum_{i=1}^{k'} t_i^2 = \|v\|_2^2 \le 1$. Then, for any $i \in \{1, 2, \dots, k'\}$, there exists $\ell_i \in B_i$ such that $|v_{\ell_i}| \le t_i / \sqrt{|B_i|}$. For these choices of ℓ_i, the right-hand side of (47) is bounded from above by

$$\begin{aligned}\sum_{i=1}^{k'} \frac{t_i}{\sqrt{|B_i|}} 1_{\{w_i \neq w_{i+1}\}} &\le \left(\sum_{i=1}^{k'} \frac{1}{|B_i|} 1_{\{w_i \neq w_{i+1}\}} \right)^{1/2} \left(\sum_{i=1}^{k'} t_i^2 \right)^{1/2} \\ &\le \left(\sum_{i=1}^{k'} \frac{1}{|B_i|} 1_{\{w_i \neq w_{i+1}\}} \right)^{1/2}, \end{aligned}$$

which proves the desired result. □

Remark D.8. Note that $\Gamma_i(v, \ell_i)$ is always non-negative. This is checked as follows: First, the lower total variation is always larger than the difference of boundary points, that is, for every $v \in \mathbb{R}^m$, we have

$$\sum_{j=1}^{m-1} (v_j - v_{j+1})_+ \ge (v_1 - v_m)_+ \ge w(v_1 - v_m),$$

where w is taken arbitrarily from $\{0, 1\}$. The equality holds if and only if v is monotone non-increasing. Then, for any $\ell \in [m]$ and $w_1, w_2 \in \{0, 1\}$, we have

$$\mathcal{V}_-(v) \ge \sum_{j=1}^{\ell-1} (v_j - v_{j+1})_+ + \sum_{j=\ell}^{m-1} (v_j - v_{j+1})_+ \ge w_1(v_1 - v_\ell) + w_2(v_\ell - v_m).$$

In particular, we obtain $\Gamma_i(v, \ell_i) \ge 0$. If θ is monotone non-decreasing (i.e., $w_0 = w_1 = \cdots = w_{k+1} = 0$), then the right-hand side of (44) equals to 0, and so $\Gamma_i(v, \ell_i) = 0$.

Step 2: Quantizing the tangent cone Now, let $\Pi = \{B_1, B_2, \dots, B_{k'}\}$ be a connected refinement of $\Pi_{\mathrm{const}}(\theta)$. Lemma D.7 implies that $T_{K_-(\mathcal{V})}(\theta) \cap B_n$ is

contained in the set such that $\sum_{i=1}^{k'} \|v_{B_i}\|_2^2 \le 1$ and $\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) \le \gamma$ for some $\ell_i \in B_i$ and $\gamma > 0$. From this perspective, we consider finitely many allocation patterns of the budgets for $\|v_{B_i}\|_2^2$ and $\Gamma_i(v, \ell_i)$. To be more precise, we construct a cover of the tangent cone in the following way. Consider a triple $(\mathbf{t}, \mathbf{q}, \mathbf{l})$ such that:

(a) $\mathbf{t} = (t_1, t_2, \dots, t_{k'})$ and $\mathbf{q} = (q_1, q_2, \dots, q_{k'})$ are vectors consisting of non-negative numbers, and
(b) $\mathbf{l} = (\ell_1, \ell_2, \dots, \ell_{k'})$ is a set of indices such that $\ell_i \in B_i$ for $i = 1, 2, \dots, k'$.

For such triple, we define a set

$$T(\mathbf{t}, \mathbf{q}, \mathbf{l}) = \left\{ v \in \mathbb{R}^n : \ \|v_{B_i}\|_2^2 \le t_i \quad \text{and} \quad \Gamma_i(v, \ell_i) \le q_i \gamma \quad \text{for } i = 1, 2, \dots, k' \right\}, \tag{48}$$

where γ is taken as the right-hand side of (44):

$$\gamma := \gamma(\theta, \Pi) = \left(\sum_{i=1}^{k'} \frac{1}{|B_i|} 1_{\{w_i \neq w_{i+1}\}} \right)^{\frac{1}{2}}. \tag{49}$$

Then, quantizing the allocation vectors $\mathbf{t}$ and $\mathbf{q}$, we can cover the set $T_{K_-(\mathcal{V})}(\theta) \cap B_n$ with finitely many $T(\mathbf{t}, \mathbf{q}, \mathbf{l})$s as the following lemma.

Lemma D.9. Suppose that $\Pi = (B_1, B_2, \dots, B_{k'})$ is a connected refinement of $\Pi_{\mathrm{const}}(\theta)$. Define the signs $w_1, w_2, \dots, w_{k'}$ as in Lemma D.7. Let $\mathcal{Q}$ be a set of allocation vectors satisfying the following condition; there exists an integer vector $\mathbf{m} = (m_1, m_2, \dots, m_{k'}) \in \mathbb{N}^{k'}$ such that $1 \le m_i \le k'$ $(i = 1, 2, \dots, k')$ and $\sum_{i=1}^{k'} m_i \le 2k'$, and the allocation vector $q = (q_1, q_2, \dots, q_{k'}) \in \mathcal{Q}$ can be written as

$$q_i = \frac{m_i}{k'} \quad \text{for all } i = 1, 2, \dots, k'.$$

Let $\mathcal{L}$ be a set of indices $\mathbf{l} = (\ell_1, \ell_2, \dots, \ell_{k'})$ such that $\ell_i \in B_i$ for all $i = 1, 2, \dots, k'$. Given $\mathbf{t}, \mathbf{q} \in \mathcal{Q}$ and $\mathbf{l} \in \mathcal{L}$, we define a set $T(\mathbf{t}, \mathbf{q}, \mathbf{l})$ as (48). Then, we have

$$T_{K_-(\mathcal{V})}(\theta) \cap B_n \subseteq \bigcup_{\substack{\mathbf{t}, \mathbf{q} \in \mathcal{Q}, \\ \mathbf{l} \in \mathcal{L}}} T(\mathbf{t}, \mathbf{q}, \mathbf{l}). \tag{50}$$

Proof. Fix any vector v in $T(\Pi, w) \cap B_n$. Since $\|v_{B_i}\|_2^2 \le \|v\|_2^2 \le 1$, there exists an integer $1 \le m_i \le k'$ such that

$$\frac{m_i - 1}{k'} \le \|v_{B_i}\|_2^2 \le \frac{m_i}{k'}.$$

Summing over $i = 1, 2, \dots, k'$, we have

$$\sum_{i=1}^{k'} m_i \le k' \sum_{i=1}^{k'} \|v_{B_i}\|_2^2 + k' \le 2k',$$

which implies $\mathbf{t} = (m_1/k', \dots, m_{k'}/k') \in \mathcal{Q}$.

Next, by Lemma D.7, there exist $\mathbf{l} = (\ell_1, \dots, \ell_{k'}) \in \mathcal{L}$ such that $\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) \le \gamma$. Hence, for any i, there exists an integer $1 \le l_i \le k'$ such that

$$\frac{(l_i - 1)\gamma}{k'} \le \Gamma_i(v, \ell_i) \le \frac{l_i \gamma}{k'}$$

Suppose $\gamma > 0$. Summing over $i = 1, 2, \dots, k'$, we have $\sum_{i=1}^{k'} l_i \le 2k'$ and thus $\mathbf{q} = (l_1/k', \dots, l_{k'}/k') \in \mathcal{Q}$. For the case of $\gamma = 0$, it is clear that $\mathbf{q} = (1/k', 1/k', \dots, 1/k') \in \mathcal{Q}$. □

We should note that the cardinalities of $\mathcal{Q}$ and $\mathcal{L}$ are respectively bounded as follows:

Proposition D.10. Let $\mathcal{Q}$ and $\mathcal{L}$ are the sets defined in Lemma D.9. Then, we have:

(i) $\log |\mathcal{Q}| \le 2k' \log 2\mathrm{e}$, and
(ii) $\log |\mathcal{L}| \le k' \log \frac{n}{k'}$.

Proof. For the first part, we observe that $|\mathcal{Q}|$ is not larger than the cardinality of

$$\bigcup_{M=k'}^{2k'} \left\{ \mathbf{m} = (m_1, \dots, m_{k'}) \in \mathbb{N}^{k'} : 1 \le m_i \le k', \sum_i m_i = M \right\}.$$

Then, we have

$$\begin{aligned} |\mathcal{Q}| &\le \sum_{j=0}^{k'} \binom{k' + j - 1}{k' - 1} = \sum_{j=0}^{k'} \binom{k' + j - 1}{j} \le \sum_{j=0}^{k'} \binom{2k' - 1}{j} \\ &\underset{\text{(a)}}{\le} \left(\frac{(2k' - 1)\mathrm{e}}{k'} \right)^{k'} \le (2\mathrm{e})^{k'}. \end{aligned}$$

The proof of the inequality (a) in the above can be found in Proposition 4.3 of Dudley (2014).

The second part is obtained by Jensen's inequality as

$$\log |\mathcal{L}| = \sum_{i=1}^{k'} \log |B_i| \le k' \log \left(\sum_{i=1}^{k'} \frac{|B_i|}{k'} \right) = k' \log \frac{n}{k'}.$$ □

Step 3: Controlling Gaussian widths As mentioned before, our goal is to obtain an upper bound of the Gaussian width

$$\tilde{W}(\theta) := w(T_{K_-(\mathcal{V})}(\theta) \cap B_n) = \mathbb{E} \left[\sup_{v \in T_{K_-(\mathcal{V})}(\theta) \cap B_n} \langle v, Z \rangle \right], \tag{51}$$

where we convene that $\mathbb{E} = \mathbb{E}_{Z \sim N(0, I_n)}$. Let (Π, w) is a pair of a partition and a sign vector of knots defined as in Lemma D.7. Using the decomposition in

Lemma D.9, we have

$$\tilde{W}(\theta) \le \mathbb{E}\left[\max_{\mathbf{t},\mathbf{q}\in\mathcal{Q},\ \mathbf{l}\in\mathcal{L}} \sup_{v\in T(\mathbf{t},\mathbf{q},\mathbf{l})} \langle v, Z\rangle\right].$$

Besides, leveraging a general result for Gaussian suprema (see Lemma F.4 below), we have

$$\tilde{W}(\theta) \le \max_{\mathbf{t},\mathbf{q}\in\mathcal{Q},\ \mathbf{l}\in\mathcal{L}} \mathbb{E}\left[\sup_{v\in T(\mathbf{t},\mathbf{q},\mathbf{l})} \langle v, Z\rangle\right] + 3\sqrt{k' \log \frac{\mathrm{e}n}{k'}} + \sqrt{\frac{\pi}{2}}. \tag{52}$$

Here, we used Proposition D.10 to bound the cardinality of the set $\mathcal{Q}^2 \times \mathcal{L}$. More precisely, we used the following evaluation:

$$2\log|\mathcal{Q}^2 \times \mathcal{L}| \le 4k' \log 2\mathrm{e} + 2k' \log \frac{\mathrm{e}n}{k'} \le (4\log 2\mathrm{e} + 2)k' \log \frac{\mathrm{e}n}{k'} < 8.8k' \log \frac{\mathrm{e}n}{k'}.$$

Given $\mathbf{t}, \mathbf{q} \in \mathcal{Q}$ and $\mathbf{l} \in \mathcal{L}$, we define

$$\tilde{W}(\mathbf{t},\mathbf{q},\mathbf{l}) = \mathbb{E}\left[\sup_{v\in T(\mathbf{t},\mathbf{q},\mathbf{l})} \langle v, Z\rangle\right].$$

Dividing the supremum into k' pieces $v_{B_1}, \dots, v_{B_{k'}}$, this quantity is bounded from above as $\tilde{W}(\mathbf{t},\mathbf{q},\mathbf{l}) \le \sum_{i=1}^{k'} \tilde{W}_i(t_i, q_i, \ell_i)$, where

$$\tilde{W}_i(t_i, q_i, \ell_i) := \mathbb{E}_{Z_i \sim N(0, I_{|B_i|})}\left[\sup_{v_{B_i} \in T_i(t_i,q_i,\ell_i)} \langle v_{B_i}, Z_i\rangle\right]. \tag{53}$$

Here, we write $T_i(t_i, q_i, \ell_i) := \{v_{B_i} \in \mathbb{R}^{B_i} :\ \|v_{B_i}\|_2^2 \le t_i,\ \Gamma_i(v, \ell_i) \le q_i\gamma\}$.

We now consider the quantity (53). In the set $T_i(t_i, q_i, \ell_i)$ over which the supremum taken, the lower total variation of v_{B_i} is bounded from above as

$$\mathcal{V}_-^{B_i}(v_{B_i}) \le w_i(v_{\tau_i} - v_\ell) + w_{i+1}(v_{\ell_i} - v_{\tau_{i+1}-1}) + q_i\gamma. \tag{54}$$

As mentioned in Remark D.8, the reverse inequality

$$\mathcal{V}_-^{B_i}(v_{B_i}) \ge w_i(v_{\tau_i} - v_\ell) + w_{i+1}(v_{\ell_i} - v_{\tau_{i+1}-1})$$

is always true, and the equality can hold only if two sub-vectors $(v_{\tau_i}, v_{\tau_i} + 1, \dots, \ell_i)$ and $(\ell_i, \ell_i + 1, \dots, v_{\tau_{i+1}} - 1)$ are either monotone increasing or non-increasing. From this point of view, we may consider that the meaning of the condition (54) is that v_{B_i} is approximated by two nearly monotone pieces. This suggests that the complexity of $T_i(t_i, q_i, \ell_i)$ can be evaluated by that of the class of monotone functions.

Below, we provide the upper bound of the Gaussian width of the form (53). First, the following lemma treats a special case where ℓ_i is taken as the rightmost point in B_i.

Lemma D.11. For every $n \geq 1$, $t > 0$, $w \in \{0, 1\}$ and $\gamma \geq 0$, we have

$$\mathbb{E}\left[\sup\left\{\langle v, Z\rangle \ : \ v \in \mathbb{R}^n, \|v\|_2 \leq t, \text{ and } \sum_{i=1}^{n-1}(v_i - v_{i+1})_+ \leq w(v_1 - v_n) + \gamma\right\}\right]$$
$$\leq (t + 2\gamma\sqrt{n-1})\sqrt{\log(en)}. \tag{55}$$

Proof. The proof is divided into two cases where $w = 1$ and $w = 0$.

Case 1 ($w = 1$): By scaling properly, we need only consider the case where $t = 1$. For a vector $v \in \mathbb{R}^n$, we define a monotone vector v^+ as

$$v_1^+ = 0 \quad \text{and} \quad v_i^+ = \sum_{j=2}^{i}(v_j - v_{j-1})_+ \quad \text{for } i = 2, \ldots, n.$$

We also define another monotone vector v^- as

$$v_1^- = -v_1 \quad \text{and} \quad v_i^- = v_1^- + \sum_{j=2}^{i}(v_{j-1} - v_j)_+ \quad \text{for } i = 2, \ldots, n.$$

It is easy to check that $v = v^+ - v^-$. Using these notations, we have

$$\mathcal{V}_-(v) = \sum_{i=1}^{n-1}(v_i - v_{i+1})_+ = v_n^- - v_1^-.$$

Hence, the condition $\mathcal{V}_-(v) \leq v_1 - v_n + \gamma$ is equivalent to $v_n^+ \leq \gamma$, which leads to

$$\|v^+\|_2^2 \leq (n-1)(v_n^+)^2 \leq (n-1)\gamma^2$$

and

$$\|v_-\|_2 \leq \|v\|_2 + \|v^+\|_2 \leq 1 + \gamma\sqrt{n-1}.$$

Denote by $\tilde{W}$ the left-hand side in (55) with $t = 1$. The argument in the previous paragraph implies that

$$\tilde{W} \leq \mathbb{E}\left[\sup_{v^+ \in K_n^\uparrow:\ \|v^+\|_2 \leq \gamma\sqrt{n-1}} \langle v^+, Z\rangle\right] + \mathbb{E}\left[\sup_{v^- \in K_n^\uparrow:\ \|v^-\|_2 \leq 1+\gamma\sqrt{n-1}} \langle v^-, Z\rangle\right]$$
$$\leq (1 + 2\gamma\sqrt{n-1}) \cdot \mathbb{E}\left[\sup_{v \in K_n^\uparrow:\ \|v\|_2 \leq 1} \langle v, Z\rangle\right]. \tag{56}$$

The expectation in the last line is bounded as

$$\left(\mathbb{E}\left[\sup_{v \in K_n^\uparrow:\ \|v\|_2 \leq 1} \langle v, Z\rangle\right]\right)^2 \leq \mathbb{E}\left[\left(\sup_{v \in K_n^\uparrow:\ \|v\|_2 \leq 1} \langle v, Z\rangle\right)^2\right] \leq \log(en).$$

Here, the first inequality is the Jensen's inequality, and the second inequality is a consequence of equation (D.12) in Amelunxen et al. (2014). Combining with (56), we have the desired result.

Case 2 ($w = 0$): We can assume w.l.o.g. $t = 1$. As in Case 1, and we write a vector as a difference of monotone vectors. For $v \in \mathbb{R}^n$, we define v^+ and v^- as

$$v_1^+ = v_1 \quad \text{and} \quad v_i^+ = \sum_{j=2}^{i} (v_j - v_{j-1})_+ \quad \text{for } i = 2, \dots, n$$

and

$$v_1^- = 0 \quad \text{and} \quad v_i^- = v_1^- + \sum_{j=2}^{i} (v_{j-1} - v_j)_+ \quad \text{for } i = 2, \dots, n,$$

respectively. Under this notation, the condition $\mathcal{V}_-(v) \le \gamma$ is equivalent to $v_n^- \le \gamma$, and therefore we have

$$\|v^+\|_2 \le 1 + \gamma\sqrt{n-1} \quad \text{and} \quad \|v^-\|_2 \le \gamma\sqrt{n-1}.$$

Then, a similar argument as Case 1 yields the result. □

Next, the following lemma provides an upper bound of $\tilde{W}_i$ for general choices of $\ell_i \in B_i$.

Lemma D.12. Fix $n \ge 1$, $1 \le \ell \le n$, $t > 0$ and $\gamma \ge 0$. For every $w_1, w_2 \in \{0, 1\}$, the quantity

$$\tilde{W} := \mathbb{E}\Bigg[\sup \Bigg\{ \langle v, Z \rangle \ : v \in \mathbb{R}^n, \|v\|_2 \le t, \text{ and } \mathcal{V}_-(v) \le w_1(v_1 - v_\ell) + w_2(v_\ell - v_n) + \gamma \Bigg\} \Bigg]$$

is bounded from above as

$$\tilde{W} \le \begin{cases} (t + 2\gamma\sqrt{\ell - 1})\sqrt{\log(\mathrm{e}\ell)} \\ \quad + (t + 2\gamma\sqrt{n - \ell})\sqrt{\log(\mathrm{e}(n - \ell + 1))} & \text{if } 1 < \ell < n \\ (t + 2\gamma\sqrt{n-1})\sqrt{\log(\mathrm{e}n)} & \text{if } \ell = 1 \text{ or } n. \end{cases} \tag{57}$$

In particular, we deduce a simpler bound

$$\tilde{W} \le 2(t + 2\gamma\sqrt{n-1})\sqrt{\log(\mathrm{e}n)}. \tag{58}$$

Proof. Let (A_1, A_2) be a pair of sub-vectors of $[n]$ defined as $A_1 = \{1, 2, \dots, \ell\}$ and $A_2 = \{\ell, \ell + 1, \dots, n\}$. If either $\ell = 1$ or $\ell = n$ (i.e., one of A_1 and A_2 becomes a singleton), the result is a direct consequence of Lemma D.11.

Henceforth, we assume that $1 < \ell < n$. Suppose that $v \in \mathbb{R}^n$ satisfies the assumption $\mathcal{V}_-(v) \le w_1(v_1 - v_\ell) + w_2(v_\ell - v_n) + \gamma$. Since $\mathcal{V}_-(v) \ge \mathcal{V}_-^{A_1}(v_{A_1}) + w_2(v_\ell - v_n)$, we have

$$\mathcal{V}_-^{A_1}(v_{A_1}) \le w_1(v_1 - v_\ell) + \gamma.$$

Similarly, we have

$$\mathcal{V}_{-}^{A_2}(v_{A_2}) \le \mathcal{V}_{-}(v) - w_1(v_1 - v_\ell) \le w_2(v_\ell - v_n) + \gamma.$$

Based on these observations, we reduce to $\tilde{W} \le$

$$\mathbb{E}\left[\sup_{\substack{v_{A_1}\in\mathbb{R}^\ell:\|v_{A_1}\|_2\le t,\\ \mathcal{V}_{-}^{A_1}(v_{A_1})\le w_1(v_1-v_\ell)+\gamma}} \langle v_{A_1}, Z_{A_1}\rangle\right] + \mathbb{E}\left[\sup_{\substack{v_{A_2}\in\mathbb{R}^{n-\ell+1}:\|v_{A_2}\|_2\le t,\\ \mathcal{V}_{-}^{A_2}(v_{A_2})\le w_2(v_\ell-v_n)+\gamma}} \langle v_{A_2}, Z_{A_2}\rangle\right],$$

in which both terms in the right-hand side can be bounded using Lemma D.11. □

Before going to the next step, we summarize the results in Step 3 as follows.

Proposition D.13. Fix $\theta \in \mathbb{R}^n$. Let $\Pi = (B_1, B_2, \dots, B_{k'})$ be any connected refinement of $\Pi_{\mathrm{const}}(\theta)$, and $w_1, w_2, \dots, w_{k'}$ be the signs associated with Π as in Lemma D.7. Define $\gamma \ge 0$ as (49). Then, the quantity $\tilde{W}(\theta)$ defined in (53) is bounded from above by $\tilde{W}(\theta) \le$

$$\max_{\mathbf{t},\mathbf{q}\in\mathcal{Q}} \left\{ \sum_{i=1}^{k'} 2(\sqrt{t_i} + 2q_i\gamma\sqrt{|B_i|-1})\sqrt{\log(\mathrm{e}|B_i|)} + 3\sqrt{k'\log\frac{\mathrm{e}n}{k'}} + \sqrt{\frac{\pi}{2}} \right\}. \quad (59)$$

Proof. This is a direct consequence of (52) and (58). □

Step 4: Applying Lemma D.3 We now are ready to complete the proof of Theorem 4.1.

Recall that our goal is to obtain an upper bound for $\tilde{W}(\theta)$ which is defined in (53). To this end, we will construct a suitable refinement of $\Pi_{\mathrm{const}}(\theta)$ with moderate piece lengths so that we can control the first term in (59). In fact, from an argument parallel to that in Guntuboyina et al. (2020), there exists a refinement $\Pi = (B_1, B_2, \dots, B_{k'})$ such that

$$|B_i| \le \frac{4n}{k'} \quad \text{for } i = 1, 2, \dots, k'$$

and $k(\theta) \le k' \le 2k(\theta)$. We also define the signs $w_1, w_2, \dots, w_{k'}$ in a similar way as Lemma D.6, but if the knot τ_i is not contained in the original partition $\Pi_{\mathrm{const}}(\theta)$, the corresponding sign w_i will be specified later.

We can bound the first term in (59) as the following two steps. First, from the Cauchy–Schwarz inequality and the fact that $\mathbf{t} \in \mathcal{Q}$, we have

$$\begin{aligned}\sum_{i=1}^{k'} \sqrt{t_i}\sqrt{\log(\mathrm{e}|B_i|)} &\le \left(\sum_{i=1}^{k'} t_i\right)^{1/2} \left(\sum_{i=1}^{k'} \log(\mathrm{e}|B_i|)\right)^{1/2} \\ &\le \sqrt{2}\sqrt{k'\log\frac{\mathrm{e}n}{k'}} \le 2\sqrt{k(\theta)\log\frac{\mathrm{e}n}{k(\theta)}}.\end{aligned}$$

Second, by the above construction of Π, we have

$$\begin{aligned}\sum_{i=1}^{k'} q_i\gamma\sqrt{|B_i|-1}\sqrt{\log(\mathrm{e}|B_i|)} &\le \max_{1\le i\le k'}\left[\sqrt{|B_i|\log(\mathrm{e}|B_i|)}\right]\sum_{i=1}^{k'} q_i\gamma\\ &\le 2\gamma\cdot 2(1+\log 4)\sqrt{\frac{n}{k'}\log\frac{\mathrm{e}n}{k'}}\\ &\le 10\gamma\sqrt{\frac{n}{k(\theta)}\log\frac{\mathrm{e}n}{k(\theta)}}.\end{aligned}$$

Therefore, the right-hand side in (59) can be bounded from above by

$$10\sqrt{k(\theta)\log\frac{\mathrm{e}n}{k(\theta)}}+20\gamma\sqrt{\frac{n}{k(\theta)}\log\frac{\mathrm{e}n}{k(\theta)}}. \tag{60}$$

Here, to hide the constant term $\sqrt{\pi/2}$, we have also used the fact that

$$\sqrt{m\log\frac{\mathrm{e}n}{m}}\ge 1$$

for every integer $1\le m\le n$.

Let $w_1^0, w_2^0, \dots, w_{k(\theta)+1}^0$ be the signs associated with the constant partition $\Pi_{\mathrm{const}}(\theta)=(A_1,A_2,\dots,A_{k(\theta)})$ (recall the definition (12)). Then, we can choose the values of w_i so that the following inequality holds:

$$\begin{aligned}\gamma^2 &= \sum_{i=1}^{k'}|B_i|^{-1}1_{\{w_i\ne w_{i+1}\}} \le \sum_{j=1}^{k(\theta)}\left[\min\left\{|A_j|,\ \left\lfloor\frac{2n}{k(\theta)}\right\rfloor\right\}\right]^{-1}1_{\{w_j^0\ne w_{j+1}^0\}}\\ &\le \sum_{i=1}^{k(\theta)}\left[\min\left\{|A_i|,\ \frac{n}{k(\theta)}\right\}\right]^{-1}1_{\{w_i^0\ne w_{i+1}^0\}}\\ &= M(\theta).\end{aligned} \tag{61}$$

In fact, this is possible if we choose w_i as the sign w_j^0 for the nearest knot that is to the right of τ_i. Combining (61), (60) and Proposition D.2, the statistical dimension of $T_{K_-(\mathcal{V})}(\theta)$ is bounded from above as

$$\delta(T_{K_-(\mathcal{V})}(\theta))\le\tilde{W}^2(\theta)+1\le 800n\left[\frac{k(\theta)}{n}\log\frac{\mathrm{e}n}{k(\theta)}+\frac{M(\theta)}{k(\theta)}\log\frac{\mathrm{e}n}{k(\theta)}\right]+1,$$

where we also used the elementary fact that $(a+b)^2\le 2(a^2+b^2)$. Consequently, applying Lemma D.3, we have desired result.

Remark D.14 (Non-Gaussian noises)**.** For non-Gaussian noise setting, we could prove an analogous result to Proposition D.5. We comment on a sketch of the proof for such a generalization.

The proof of Proposition D.5 consists of (i) a decomposition argument for the tangent cone and (ii) bounds for some probabilistic quantities (i.e., the statistical dimension and the Gaussian width). The former argument is completely

deterministic and independent from the distributional assumption on the noise variables. Regarding the probabilistic bounds, we used the following bound for (Gaussian) statistical dimension of $K_n^{\uparrow}$:

$$\delta(K_n^{\uparrow}) \leq \log(en).$$

Hence, if we can obtain a similar bound for non-Gaussian random variables, we can prove a analogous result to Proposition D.5.

Let $\xi_1, \ldots, x_n$ be i.i.d. random variables with $\mathbb{E}[\xi_1] = 0$ and $\mathrm{Var}(\xi_1) = \sigma^2$. For a convex cone C, we define the statistical dimension as

$$\bar{\delta}(C) = \frac{1}{\sigma^2}\mathbb{E}\left[\left(\sup_{\theta \in C : \|\theta\|_2 \leq 1} \langle \xi, \theta \rangle\right)^2\right] = \frac{1}{\sigma^2}\mathbb{E}\|\mathrm{Proj}_C(\xi)\|^2.$$

Here, we write $\mathrm{Proj}_C(x) = \mathrm{argmin}_{z \in C}\|z - x\|_2$, and the last equality holds from a deterministic relation

$$\left(\sup_{\theta \in C : \|\theta\|_2 \leq 1} \langle \xi, \theta \rangle\right)^2 = \|\mathrm{Proj}_C(\xi)\|^2.$$

(See Amelunxen et al. (2014) for details). Then, from Theorem 3.1 in Chatterjee et al. (2015), we can check that

$$\bar{\delta}(K_n^{\uparrow}) \leq 16 \log(en).$$

Therefore, by following a similar argument as the proof of Proposition D.5, we conclude that

$$\bar{\delta}(T_{K_-(\mathcal{V})}) \leq C' n \left\{ \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)} \right\}$$

for some universal constant $C' > 0$. As a consequence, we can prove the expected risk bound similar to (20) for non-Gaussian noise variables.

D.3. Proof of Corollary 4.4

Let $\alpha > 0$ be a number to be specified later. Define a vector $\theta' \in \mathbb{R}^n$ as $\theta'_1 = \theta^*_1$ and

$$\theta'_i = \theta^*_1 + \sum_{j=1}^{i-1}(\theta^*_{j+1} - \theta^*_j)_+ - \alpha \sum_{j=1}^{i-1}(\theta^*_j - \theta^*_{j+1})_+ \quad \text{for } i = 2, 3, \ldots, n.$$

Then, we have $\mathcal{V}_-(\theta') = \alpha \mathcal{V}_-(\theta^*)$. Moreover, the constant partition and the sign of θ' (defined in (12)) are the same as those of θ^*, and therefore $k(\theta') = k(\theta^*)$ and $M(\theta') = M(\theta^*)$.

Now, we set $\alpha = \mathcal{V}/\mathcal{V}_-(\theta^*)$ so that $\mathcal{V}_-(\theta') = \mathcal{V}$. Applying the upper bound (14), we have

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \le \frac{1}{n}\|\theta' - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta^*)}{n}\log\frac{en}{k(\theta^*)} + C\sigma^2\frac{M(\theta^*)}{k(\theta^*)}\log\frac{en}{k(\theta^*)}.$$

The first term in the right-hand side is bounded from above as

$$\begin{aligned}\frac{1}{n}\|\theta' - \theta^*\|_2^2 &= \frac{(1-\alpha)^2}{n}\sum_{i=2}^{n}\left(\sum_{j=1}^{i-1}(\theta_j^* - \theta_{j+1}^*)_+\right)^2 \\ &\le (1-\alpha)^2(\mathcal{V}_-(\theta^*))^2 = (\mathcal{V} - \mathcal{V}_-(\theta^*))^2.\end{aligned}$$

From the minimal length condition (18) and the definition of $M(\theta)$, we also have

$$\frac{M(\theta^*)}{k(\theta^*)}\log\frac{en}{k(\theta^*)} \le \frac{2c^{-1}(m(\theta^*)-1)}{n}\log\frac{en}{k(\theta^*)}.$$

Combining the above inequalities, we have the desired result.

D.4. Risk bounds for penalized estimators (Proof of Theorem 4.7)

We prove Theorem 4.7 as an application of Lemma D.4. Let $\partial\mathcal{V}_-(\theta)$ denote the set of subgradients (i.e., subdifferential) of the convex function $\mathcal{V}_-(\cdot)$ at $\theta \in \mathbb{R}^n$. The task is to provide a suitable upper bound for the Gaussian mean squared distance of the set $\lambda\partial\mathcal{V}_-(\theta)$. To do this, we use the technique developed in Guntuboyina et al. (2020). The idea is stated roughly as follows: Recall that the Gaussian mean squared distance of a convex cone can be written as the statistical dimension of the polar cone (Proposition D.2-(ii)). This motivates us to relate the Gaussian mean squared distance $\mathbf{D}(\lambda\partial\mathcal{V}_-(\theta))$ to that of an associated cone. In particular, we consider the conic hull of the subdifferential:

$$\mathrm{cone}(\partial\mathcal{V}_-(\theta)) := \bigcup_{\lambda\ge 0}\lambda\partial\mathcal{V}_-(\theta).$$

As we explain later, $\mathbf{D}(\mathrm{cone}(\partial\mathcal{V}_-(\theta)))$ can be evaluated by the results in the previous subsection. Then, we can complete the proof if we have an upper bound of the following form:

$$\mathbf{D}(\lambda\partial\mathcal{V}_-(\theta)) \le \mathbf{D}(\mathrm{cone}(\partial\mathcal{V}_-(\theta))) + \Delta(\theta,\lambda), \tag{62}$$

where $\Delta(\theta,\lambda)$ is a residual term that depends on θ and λ.

First, we show that $\mathbf{D}(\mathrm{cone}(\partial\mathcal{V}_-(\theta)))$ has exactly the same value as the statistical dimension of the tangent cone of $T_{K_-(\mathcal{V}_-(\theta))}(\theta)$, which we have already provided a bound in the previous part in this paper.

Proposition D.15. For any $\theta \in \mathbb{R}^n$, the following equality holds:

$$\mathbf{D}(\mathrm{cone}(\partial \mathcal{V}_-(\theta))) = \delta(T_{K_-(\mathcal{V}(\theta))}(\theta)).$$

In particular, we have the following upper bound:

$$\mathbf{D}(\mathrm{cone}(\partial \mathcal{V}_-(\theta))) \leq Cn \left\{ \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)} \right\},$$

where C is the same universal constant as in Proposition D.5.

Proof. Let us write $T := T_{K_-(\mathcal{V}(\theta))}(\theta)$. In the light of Proposition D.2-(ii), it suffices to show that T is the polar cone of $\mathrm{cone}(\partial \mathcal{V}_-(\theta))$. However, from fundamental results in convex geometry, we always have

$$\mathrm{cone}(\partial f(\theta)) = \left(T_{K(\theta)}(\theta)\right)^\circ \quad \text{with} \quad K(\theta) := \{z \in \mathbb{R}^n : f(z) \leq f(\theta)\}$$

for any convex function $f : \mathbb{R}^n \to \mathbb{R}$ (see Lemma A.5 and Lemma A.5 in Guntuboyina et al. (2020)). For the case where $f = \mathcal{V}_-$, the set $K(\theta)$ above is

$$K_-(\mathcal{V}_-(\theta)) = \{z \in \mathbb{R}^n : \mathcal{V}_-(z) \leq \mathcal{V}_-(\theta)\},$$

which implies the desired result. □

Next, we provide an inequality of the form (62). Since $\mathrm{cone}(\partial \mathcal{V}_-(\theta)) \supseteq \lambda \partial \mathcal{V}_-(\theta)$ holds for every $\lambda \geq 0$, the definition of the Gaussian mean squared distance (Definition D.1-(ii)) suggests that $\mathbf{D}(\mathrm{cone}(\partial \mathcal{V}_-(\theta))) \leq \mathbf{D}(\lambda \partial \mathcal{V}_-(\theta))$. However, we need a reverse inequality (62). To this end, we use the following result proved by Guntuboyina et al. (2020).

Lemma D.16 (Guntuboyina et al. (2020), Proposition B.5)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and $\theta \in \mathbb{R}^n$. Define a vector v_0 as

$$v_0 := \operatorname*{argmin}_{v \in \mathrm{aff}(\partial f(\theta))} \|v\|_2, \tag{63}$$

where $\mathrm{aff}(C)$ is the affine hull of the set $C \subseteq \mathbb{R}^n$. Suppose that $v_0 \neq 0$. For any $z \in \mathbb{R}^n$, define $\lambda(z) \geq 0$ as

$$\lambda(z) := \operatorname*{argmin}_{\lambda \geq 0} \mathrm{dist}(z, \lambda \partial f(\theta)).$$

Then, $\lambda(z)$ is well-defined, and has a finite expectation $\mathbb{E}_{Z \sim N(0, I_n)}[\lambda(Z)] < \infty$.

Further, define λ^* as

$$\lambda^* := \lambda^*(\theta) = \mathbb{E}_{Z \sim N(0, I_n)}[\lambda(Z)] + \frac{2}{\|v_0\|_2}.$$

Then, for every $\lambda \geq \lambda^*$ and $v^* \in \partial f(\theta)$, we have

$$\begin{aligned} &\mathbf{D}(\lambda \partial f(\theta)) \\ &\leq 4 + \left(\sqrt{\mathbf{D}(\mathrm{cone}(\partial f(\theta)))} + \frac{4\|v^*\|_2}{\|v_0\|_2} + 2 + (\lambda - \lambda^*)\|v^*\|_2 \right)^2. \end{aligned} \tag{64}$$

Before proceeding, we introduce an additional terminology: A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *weakly decomposable* if we have

$$\operatorname*{argmin}_{v \in \mathrm{aff}(\partial f(\theta))} \|v\|_2 \in \partial f(\theta) \tag{65}$$

for every $\theta \in \mathbb{R}^n$. In other words, we can choose $v_0 \equiv v^*$ in (64) if f is weakly decomposable. Under the assumption that f is weakly decomposable, the inequality (64) can be simplified as follows:

Corollary D.17. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and weakly decomposable. Under the same notation as in Lemma D.16, we have

$$\mathbf{D}(\lambda \partial f(\theta)) \le 3\mathbf{D}(\mathrm{cone}(\partial f(\theta))) + 3(\lambda - \lambda^*)^2 \|v_0\|_2^2 + 112.$$

Now, we apply Lemma D.16 to the case $f = \mathcal{V}_-$. The following proposition provides the structural information of $\partial \mathcal{V}_-(\theta)$ that we need for evaluating the upper bound (64). The proof is postponed to Appendix D.6.

Proposition D.18. (i) $\theta \mapsto \mathcal{V}_-(\theta)$ is weakly decomposable.
(ii) For any $\theta \in \mathbb{R}^n$, let us define v_0 as (63). Then, we have

$$\|v_0\|_2^2 = \sum_{i=1}^{k} \frac{1}{|A_i|} 1_{w_i \neq w_{i+1}}. \tag{66}$$

From Proposition D.18 and Corollary D.17, $\mathbf{D}(\lambda \partial \mathcal{V}_-(\theta))$ is bounded from above by

$$C' n \left\{ \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} + \frac{M(\theta)}{k(\theta)} \log \frac{en}{k(\theta)} \right\} + C'(\lambda - \lambda^*)^2 \sum_{i=1}^{k} \frac{1}{|A_i|} 1_{w_i \neq w_{i+1}}$$

provided that $\lambda \ge \lambda^*$. Here, $C' > 0$ is a universal constant. Combining this bound with Lemma D.4, we proved the desired risk bound.

Lastly, we provide an upper bound for the optimal tuning parameter λ^*. This is obtained from the following estimate of $\mathbb{E}[\lambda(Z)]$.

Proposition D.19. Suppose that $\theta \in \mathbb{R}^n$ and $\mathcal{V}_-(\theta) > 0$. For any $z \in \mathbb{R}^n$, define $\lambda(z)$ as

$$\lambda(z) := \operatorname*{argmin}_{\lambda \ge 0} \mathrm{dist}(z, \lambda \partial \mathcal{V}_-(\theta)).$$

Then, we have

$$\mathbb{E}[\lambda(Z)] \le \min \left\{ \frac{\|\theta\|_2}{\mathcal{V}_-(\theta)}, \left(\sum_{i=1}^{k} \frac{1_{\{w_i \neq w_{i+1}\}}}{|A_i|} \right)^{-1/2} \right\} [\delta(T_{K_-(\mathcal{V}_-(\theta))}(\theta))]^{1/2},$$

where $\mathbb{E}$ is the expectation with respect to $Z \sim N(0, I_n)$.

Proof. Let $C := \mathrm{cone}(\partial\mathcal{V}_-(\theta))$ be the conic hull of $\partial\mathcal{V}_-(\theta)$, and let P_C denote the orthogonal projection map onto C. By the definition of $\lambda(z)$, there exists a vector $v(z) \in \partial\mathcal{V}_-(\theta)$ such that $\lambda(z)v(z) = P_C(z)$.

First, we show a partial result

$$\mathbb{E}[\lambda(Z)] \le \frac{\|\theta\|_2}{\mathcal{V}_-(\theta)}\sqrt{\delta(T_{K_-(\mathcal{V}_-(\theta))}(\theta))}.$$

As we will see in Appendix D.6, $\mathcal{V}_-$ is the support function for a certain convex set. Then, by the fundamental fact for the support function that $\langle\theta, v\rangle = \mathcal{V}_-(\theta)$ for all $v \in \partial\mathcal{V}_-(\theta)$ (see Corollary 8.25 in Rockafeller and Wets (1998)), we have

$$\begin{aligned}\lambda(z)\mathcal{V}_-(\theta) &= \lambda(z)\langle\theta, v(z)\rangle \quad (\because v(z) \in \partial\mathcal{V}_-(\theta)) \\ &= \langle\theta, P_C(z)\rangle \quad (\because \lambda(z)v(z) = P_C(z)) \\ &= \langle\theta, z - P_T(z)\rangle.\end{aligned}$$

Here, in the last line, $T := T_{K_-(\mathcal{V}_-(\theta))}(\theta)$ is the polar cone of C (see Proposition D.15), and we used the Moreau decomposition $z = P_C(z) + P_T(z)$. Taking the expectation of both sides with respect to $z \sim N(0, I_n)$, we have

$$\begin{aligned}\mathcal{V}_-(\theta)\mathbb{E}[\lambda(z)] &= \underbrace{\mathbb{E}[\langle\theta, z\rangle]}_{=0} - \mathbb{E}[\langle\theta, P_T(z)\rangle] \\ &\le \|\theta\|_2\mathbb{E}\|P_T(z)\|_2 \\ &\le \|\theta\|_2(\mathbb{E}\|P_T(z)\|_2^2)^{1/2} \\ &= \|\theta\|_2(\delta(T))^{1/2},\end{aligned}$$

which implies the desired result. Here, we used the equality between the statistical dimension and the expected squared norm of projection (see Proposition 3.1 in Amelunxen et al. (2014)):

$$\delta(T) = \mathbb{E}_{Z\sim N(0,I_n)}\|P_T(Z)\|_2^2.$$

To prove the other inequality, we use the characterization of $\mathrm{aff}(\partial\mathcal{V}_-(\theta))$ given in (72) in Appendix D.6 below. In particular, if we take v^* as in (75), we have

$$\langle\lambda(z)v(z), v^*\rangle = \langle v^*, P_C(z)\rangle \le \|v^*\|_2(\delta(T))^{1/2},$$

and

$$\langle v(z), v^*\rangle = \|v^*\|_2^2 = \sum_{i=1}^{k}\frac{1_{\{w_i\neq w_{i+1}\}}}{|A_i|},$$

and hence the result follows. □

D.5. Proof of Corollary 4.12

First, we explain that a monotone vector satisfying the moderate growth condition is approximated by a piecewise-constant vector such that the segments

at both ends have sufficient lengths. To this end, we need the following lemma. Here, the first two statements (i) and (ii) are shown in Lemma 2 in Bellec and Tsybakov (2015). The third statement (iii) ensures that the moderate growth conditions implies the minimal length condition (18).

Lemma D.20. Let $\theta \in K_n^{\uparrow}$ be a monotone vector satisfying the moderate growth condition and $\theta_n - \theta_1 = \mathcal{V}$. Then, there exists another monotone vector $\theta' \in K_n^{\uparrow}$ satisfying the following three conditions.

(i) θ' is k-piecewise constant with

$$k = \max\left\{3, \left\lceil \left(\frac{\mathcal{V}^2 n}{\sigma^2 \log(en)}\right)^{1/3} \right\rceil\right\}. \tag{67}$$

Here, $\lceil t \rceil$ is the smallest integer that is not less than t.

(ii) We have

$$\frac{1}{n}\|\theta - \theta'\|_2^2 \le \frac{1}{4}\max\left\{\left(\frac{\sigma^2 \mathcal{V}\log(en)}{n}\right)^{2/3}, \frac{3\sigma^2\log(en)}{n}\right\} \tag{68}$$

and

$$\frac{\sigma^2 k}{n}\log\frac{en}{k} \le 2\max\left\{\left(\frac{\sigma^2 \mathcal{V}\log(en)}{n}\right)^{2/3}, \frac{3\sigma^2\log(en)}{n}\right\}. \tag{69}$$

(iii) Let $\Pi' = \{A_1, A_2, \dots, A_k\}$ be the partition on which θ' is constant. Then, we have $|A_1| \ge n/k$ and $|A_k| \ge n/k$.

Proof. Let k be an integer defined in (67). We construct a k-piecewise constant monotone vector $\theta' \in K_n^{\uparrow}$ as follows: First, define an equi-spaced partition $I_1, I_2, \dots, I_k$ of the interval $[\theta_1, \theta_n]$ as

$$I_j := \left[\theta_1 + \frac{j-1}{k}\mathcal{V},\ \theta_1 + \frac{j}{k}\mathcal{V}\right) \quad \text{for } j = 1, 2, \dots, k-1,$$

and $I_k := [\theta_1 + \frac{k-1}{k}\mathcal{V}, \theta_n]$. Next, define a partition $\Pi = (A_1, A_2, \dots, A_k)$ of $[n]$ as $A_j := \{i \in [n] : \theta_i \in I_j\}$ $(j = 1, 2, \dots, k)$. Then, let θ' be a piecewise-constant vector such that $\theta'_i := \theta_1 + \frac{j-1/2}{k}\mathcal{V}$ for $i \in A_j$. See the right panel of Figure 4 for an illustrative example for θ and its piecewise-constant approximation θ'. By a similar argument as Lemma 2 in Bellec and Tsybakov (2015), we can check (i) and (ii).

It remains to prove (iii) under the moderate growth condition. Below, we will only check that the maximal element in A_1 is not less than n/k because $|A_k| \ge n/k$ can be checked in a similar way. Let $i^* := \lceil n/k \rceil$. Note that we have $i^* \le \lceil n/2 \rceil$ since $k \ge 3$. By the moderate growth condition, we have

$$\theta_{i^*} \le \theta_1 + \frac{n/k - 1}{n-1}\mathcal{V} \le \theta_1 + \frac{\mathcal{V}}{k},$$

which means $i^* \in A_1$ and hence $|A_1| \ge \lceil n/k \rceil$. □

Now, we are ready to prove Corollary 4.12. Applying Lemma D.20 for every segments $A_1, A_2, \dots, A_m$, we have a k-piecewise constant and m-piecewise monotone vector $\theta' \in \mathbb{R}^n$ such that

$$\frac{1}{n}\|\theta^* - \theta'\|_2^2 \le \frac{1}{4}\max\left\{\left(\frac{\sigma^2\mathcal{V}\log\frac{en}{m}}{n}\right)^{2/3}, \frac{3m\sigma^2}{n}\log\frac{en}{m}\right\}$$

and

$$\frac{\sigma^2 k}{n}\log\frac{en}{k} \le 2\max\left\{\left(\frac{\sigma^2\mathcal{V}\log\frac{en}{m}}{n}\right)^{2/3}, \frac{3m\sigma^2}{n}\log\frac{en}{3m}\right\}.$$

Moreover, θ' satisfies the minimum length condition (18) with $c = 1$. Therefore, we have $M(\theta') \le 2(m-1)k/n$ and

$$\frac{\sigma^2 M(\theta')}{k}\log\frac{en}{k} \le \frac{2(m-1)\sigma^2}{n}\log\frac{en}{m},$$

where we used an obvious inequality $m \le k$. Then, Theorem 4.7 implies that there exists λ such that

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 &\le \frac{1}{n}\|\theta^* - \theta'\|_2^2 + C\frac{\sigma^2 k}{n}\log\frac{en}{k} + C\frac{\sigma^2 M(\theta')}{k}\log\frac{en}{k} \\ &\le C'\max\left\{\left(\frac{\sigma^2\mathcal{V}\log\frac{en}{m}}{n}\right)^{2/3}, \frac{m\sigma^2}{n}\log\frac{en}{m}\right\}\end{aligned}$$

for some universal constant $C' > 0$. This is the desired conclusion. Note that an upper bound for such λ is suggested by Proposition 4.8.

D.6. Subdifferential and weak decomposability

In this subsection, we discuss the structure of the subdifferential of the nearly-isotonic type penalties. The main purpose is to discuss the weak decomposability (defined in Appendix D.4) of $\mathcal{V}_-$.

D.6.1. Characterization of the subdifferential

First, we observe that $\mathcal{V}_-(\theta) = \sum_{i=1}^{n-1}(\theta_i - \theta_{i+1})_+$ can be written as a support function of a certain convex set. In fact, by Theorem 8.24 in Rockafeller and Wets (1998), we can see that

$$\mathcal{V}_-(\theta) = \max_{v\in\mathcal{B}}\langle v, \theta\rangle, \tag{70}$$

where $\mathcal{B}$ is a closed convex set. Conversely, once we have a convex function $\mathcal{V}_-$, the set $\mathcal{B}$ is specified as

$$\mathcal{B} = \{v \in \mathbb{R}^n : \forall\theta \in \mathbb{R}^n,\ \langle v, \theta\rangle \le \mathcal{V}_-(\theta)\}.$$

Many properties of the support function can be understood through the structure of the set $\mathcal{B}$; In particular, we can characterize the subdifferential and weak decomposability. Below, we investigate the more detailed structure of the set $\mathcal{B}$ in terms of submodular functions.

Let $G = (V, E)$ be a directed graph equipped with positive edge weights $\{c_{(i,j)}\}$. For any $\theta \in \mathbb{R}^n$, we define a nearly-isotonic type penalty $\mathcal{V}_G(\theta)$ for the weighted graph G as in (33). For any subset $A \subseteq [n]$, we also define $\kappa_G(A)$ by the total weights of outgoing edges:

$$\kappa_G(A) := \sum_{(i,j)\in E:\ i\in A,\ j\notin A} c_{(i,j)}. \tag{71}$$

The function $A \mapsto \kappa_G(A)$ is called the cut function of the weighted graph G.

It is well known that the cut function is a submodular function. Here, a function $F : 2^{[n]} \to \mathbb{R}$ is called submodular if $F(\emptyset) = 0$ and

$$F(A) + F(B) \geq F(A \cap B) + F(A \cup B)$$

holds for any subsets $A, B \subseteq [n]$. We refer the reader to Bach (2013) for fundamental properties of submodular functions. For any submodular function $F : 2^{[n]} \to \mathbb{R}$, we define the base polyhedron $\mathcal{B}(F) \subseteq \mathbb{R}^n$ as

$$\mathcal{B}(F) := \left\{ v \in \mathbb{R}^n : \sum_{i\in V} v_i = F(V) \text{ and } \sum_{i\in A} v_i \leq F(A) \text{ for all } A \subseteq V \right\}.$$

The Lovász extension $f : \mathbb{R}^n \to \mathbb{R}$ of F is defined as the support function of $\mathcal{B}(F)$, that is, for any $\theta \in \mathbb{R}^n$, $f(\theta) := \max_{v\in\mathcal{B}(F)} \langle v, \theta \rangle$.

We see that the nearly-isotonic type penalty (33) is actually the Lovász extension of the cut function (71).

Proposition D.21. For any directed graph G and edge weight $c_{(i,j)}$, the function $\mathcal{V}_G$ is the Lovász extension of the cut function κ_G.

Proof. This is the consequence of the well-known result so-called the greedy algorithm; see e.g., Proposition 3.2 in Bach (2013). In particular, we can find a derivation in Section 6.2 of Bach (2013). □

Now, we have the following useful characterizations of the subdifferential.

Proposition D.22. Define $F : 2^{[n]} \to \mathbb{R}$ be a submodular function and $f : \mathbb{R}^n \to \mathbb{R}$ be its Lovász extension. Suppose $\theta \in \mathbb{R}^n$.

(i) The subdifferential $\partial f(\theta)$ coincides with a face of $\mathcal{B}(F)$ given as

$$\partial f(\theta) = \operatorname*{argmax}_{v\in\mathcal{B}(F)} \langle v, \theta \rangle = \{ v \in \mathcal{B}(F) : \langle v, \theta \rangle = f(\theta) \}.$$

(ii) There is an (ordered) partition $(A_1, A_2, \ldots, A_k) \subseteq [n]$ such that

$$\operatorname{aff}(\partial f(\theta)) = \left\{ v \in \mathbb{R}^n : \sum_{j\in S_i} v_j = F(S_i) \text{ for all } i = 1, 2, \ldots, k \right\}, \tag{72}$$

where $S_i := \bigcup_{j=1}^{i} A_j$ $(i = 1, 2, \ldots, k)$. In particular, we have $\partial f(\theta) = \mathcal{B}(F) \cap \mathrm{aff}(\partial f(\theta))$.

(iii) Let v be any point in the relative interior of $\partial f(\theta)$. Then, the normal cone of $\partial f(\theta)$ at v is contained in the set of partition-wise constant vectors:

$$N_{\partial f(\theta)}(v) \subseteq \mathrm{span}\{1_{A_1}, 1_{A_2}, \ldots, 1_{A_k}\}.$$

Proof. The first statement is just a well-known property for the support function (Corollary 8.25 in Rockafeller and Wets (1998)). The second statement follows from the characterization of faces for the base polyhedron (see Proposition 4.7 in Bach (2013)). The third statement follows from (ii) and the characterization of normal cones of polyhedra (see Theorem 6.46 in Rockafeller and Wets (1998)). □

D.6.2. Weak decomposability

Here, we discuss the weak decomposability of the Lovász extension.

Before describing the result, we introduce some terminology. Let $F : 2^{[n]} \to \mathbb{R}$ be a submodular function. We say that a set $A \subseteq [n]$ is separable for F if there is a non-empty proper subset B of A such that $F(A) = F(B) + F(A \setminus B)$. We also say that A is inseparable if it is not separable. For example, if $F = \kappa_G$ is the cut function defined in (71), A is inseparable if and only if it is a connected component in the graph G. Furthermore, we define the following *agglomerative clustering condition*.

Definition D.23. We say that a submodular function $F : 2^{[n]} \to \mathbb{R}$ satisfies the agglomerative clustering (AC) condition if it has the following property: Let $A, B \subseteq [n]$ be a any disjoint pair of subsets such that $A \neq \emptyset$ and A is inseparable for the function $F_B^A : 2^A \to \mathbb{R}$ defined by $F_B^A(C) := F(B \cup C) - F(B)$. Then, for any $C \subset A$, we have

$$\frac{|C|}{|A|}(F(B \cup A) - F(B)) \leq F(B \cup C) - F(B). \tag{73}$$

Recall the definition of weak decomposability (65). The following proposition provides a sufficient condition for the weak decomposability of the Lovász extension.

Proposition D.24. Let $F : 2^{[n]} \to \mathbb{R}$ be a submodular function satisfying the AC condition in Definition D.23. Then, the Lovász extension of f of F is weakly decomposable.

Proof. Fix $\theta \in \mathbb{R}^n$. Since f is the support function of the base polyhedron $\mathcal{B}(F)$, $\partial f(\theta)$ coincides with a face of $\mathcal{B}(F)$. Let $A_1, A_2, \ldots, A_k$ be a partition of $[n]$ such that $\mathrm{aff}(\partial f(\theta))$ is represented as (72). For $i = 1, 2, \ldots, k$, we write $S_0 := \emptyset$ and $S_i := A_1 \cup A_2 \cup \cdots \cup A_i$. We should note that the above partition can be chosen so that A_i is inseparable for the function defined as

$$(A_i \supseteq)\ C \mapsto F(S_{i-1} \cup C) - F(S_{i-1}).$$

In this case, $\partial f(\theta)$ is an $n-k$ dimensional subset.

Define a vector v^* as

$$v^* := \sum_{i=1}^{k} \frac{F(S_i) - F(S_{i-1})}{|A_i|} 1_{A_i}. \tag{74}$$

Since

$$\sum_{j \in S_i} v_j^* = \sum_{j=1}^{i} (F(S_j) - F(S_{j-1})) = F(S_i)$$

holds for any $i = 1, \dots, k$, we have $v^* \in \mathrm{aff}(\partial f(\theta))$. Moreover, v^* is also contained in the normal cone of $\mathrm{aff}(\partial f(\theta))$. Hence, if we prove $v^* \in \partial f(\theta)$, we have

$$\forall v \in \partial f(\theta), \quad \langle v^*, v - v^* \rangle = 0,$$

which implies that $v^* \in \mathrm{argmin}_{v \in \partial f(\theta)} \|v\|_2^2$.

Now, our goal is to prove $v^* \in \partial f(\theta)$ under the AC condition. If $k = n$, then it is clear from (72) that $\partial f(\theta) = \{v^*\}$. Below, we assume that $k < n$. Since $v^* \in \mathrm{aff}(\partial f(\theta))$, it suffices to show that $\sum_{i \in S} v_i^* \le F(S)$ holds for any $S \subseteq [n]$ that determines a relative boundary of $\partial f(\theta)$. The relative boundary of $\partial f(\theta)$ can be written as the union of all $n-k-1$ dimensional faces of $\mathcal{B}(F)$ that have non-empty intersection with $\partial f(\theta)$. Such faces can be characterized as follows: Let $\Pi = (A_1, A_2, \dots, A_k)$ be the partition defined in the above, and choose A_i with $|A_i| \ge 2$. Let A_i' be any non-empty proper subset of A_i. We define a new ordered partition of $[n]$ by inserting $(A_i', A_i \setminus A_i')$ instead of A_i:

$$\Pi' = (A_1, A_2, \dots, A_{i-1}, A_i', (A_i \setminus A_i'), A_{i+1}, \dots, A_k).$$

Then, Π' defines an $n-k-1$ dimensional affine subspace by (72), which defines a part of the relative boundary of $\partial f(\theta)$. Therefore, we have to show that $\sum_{i \in S} v_i^* \le F(S)$ for any S that can be written as $S = S_{i-1} \cup A_i'$ with $A_i' \subset A_i$. From the AC condition, we have

$$\begin{aligned}
\sum_{i \in S} v_i^* &= \sum_{j=1}^{k} \frac{F(S_j) - F(S_{j-1})}{|A_j|} |A_j \cap S| \\
&= \sum_{j=1}^{i-1} (F(S_j) - F(S_{j-1})) + \frac{F(S_{i-1} \cup A_i') - F(S_{i-1})}{|A_i|} |A_i'| \\
&\le F(S_{i-1}) + (F(S_{i-1} \cup A_i') - F(S_{i-1})) \\
&= F(S).
\end{aligned}$$

This proves that $v^* \in \partial f(\theta)$, and hence f is weakly decomposable. □

Remark D.25. The AC condition was originally introduced in Bach (2011). In that paper, the author consider the proximal denoising estimators (37) where f is the Lovász extension of a submodular function F. The name "agglomerative

clustering" captures the following property: Let us consider the *solution path* of the minimization problem (37) parametrized by λ, that is, the solution path is the collection $\{\hat{\theta}_\lambda\}_{\lambda \geq 0}$ calculated for all $\lambda \geq 0$. In general, the solution path starts with $\hat{\theta}_\lambda = y$ for $\lambda = 0$, and $\hat{\theta}_\lambda$ shrinks toward some piecewise constant vector as λ increases. Proposition 4 of Bach (2011) showed that the solution path is agglomerative if F satisfies the AC condition.

We provide some examples of functions satisfying the AC condition:

- Let $h : \mathbb{R} \to \mathbb{R}$ be a concave function with $h(0) = 0$. A submodular function defined as $F(A) := h(|A|)$ satisfies the AC condition. Examples of solutions paths for this class can be found in Bach (2011).
- The one-dimensional fused lasso has an agglomerative solution path. The corresponding submodular function is the cut function of the undirected one-dimensional grid graph, which satisfies the AC condition. Hence, by Proposition D.24, the penalty of the one-dimensional fused lasso is weakly decomposable. This provides an alternative proof for Lemma 2.7 in Guntuboyina et al. (2020). On the other hand, the fused lasso on the two-dimensional grid does not satisfy this condition. See Bach (2011) for details.
- The nearly-isotonic regression (3) has an agglomerative solution path. A direct proof for this property is provided in Lemma 1 in Tibshirani et al. (2011). Below, we prove that the cut function for directed one-dimensional grid graph satisfies the AC condition, which provides an alternative proof for this fact.

The following proposition provides a proof for Proposition D.18.

Proposition D.26. The cut function F associated with the nearly-isotonic regression satisfies the AC condition. In particular, the lower total variation $\mathcal{V}_-(\theta)$ is weakly decomposable. Moreover, for any $\theta \in \mathbb{R}^n$, the minimum value of the ℓ_2-norm in $\partial \mathcal{V}_-(\theta)$ is given by (66).

Proof. For any $A \subseteq V := [n]$, $F(A)$ is given by the number of connected components in A that does not contains the rightmost point n. Let $A \subseteq [n]$ be a connected subset, and $B \subseteq [n] \setminus A$. The value of $F(B \cup A) - F(B)$ depends on whether one or both of two endpoints of A are adjacent to B.

We will check the AC condition by considering all patterns of adjacency as Table 1. Here, C represents any proper subset of A, and "None" means that A contains 1 or n. In each case, we can easily check that the inequality (73) is satisfied. Hence, F satisfies the AC condition.

The second statement is a consequence of Proposition D.24.

The last statement follows from fact that the minimizer of $\|v\|_2^2$ in $\partial f(\theta)$ coincides with that in $\mathrm{aff}(\partial f(\theta))$, which is given as (74). In this case, we can choose $A_1, A_2, \ldots, A_k$ as the constant partition of θ that is sorted by the values of θ. Thus, we have

$$v^* = \sum_{i=1}^{k} \frac{F(S_i) - F(S_{i-1})}{|A_i|} 1_{A_i} = \sum_{i=1}^{k} \frac{1_{w_i \neq w_{i+1}}}{|A_i|} 1_{A_i} \tag{75}$$

TABLE 1
The values of F_B^A for the cut function F of one-dimensional grid graph.

Node left to A	Node right to A	$F(B \cup A) - F(B)$	$F(B \cup C) - F(B)$
None	None	0	≥ 0
None	B	0	≥ 0
None	$V \setminus B$	1	$\geq 1_{\{C \neq \emptyset\}}$
B	None	-1	≥ 0
B	B	-1	≥ 0
B	$V \setminus B$	0	≥ 0
$V \setminus B$	None	0	≥ 0
$V \setminus B$	B	0	≥ 0
$V \setminus B$	$V \setminus B$	1	$\geq 1_{\{C \neq \emptyset\}}$

which proves the desired result. □

Remark D.27 (Missing part in the proof of Proposition A.1)**.** With a slight modification of the above argument, we can show the AC condition for the cut function of weighted graph

$$F(A) = \sum \{c_j : j \in A, j+1 \notin A\},$$

where $c_j > 0$ $(j = 1, \ldots, n-1)$ are edge weights. As mentioned in Proposition A.1, we need this result to prove the validity of the modified PAVA algorithm (Algorithm 1). Here, we prove that (31) provides a sufficient condition for the AC condition, and hence the solution path of the weighted nearly-isotonic regression (30) is agglomerative.

Let $A \subseteq [n]$ be a non-empty connected subset, B be a subset of $[n] \setminus A$, and C be a proper subset of A. Recall that our goal is to check the inequality (73). For clarity, we write $A = \{j_L, j_L + 1, \ldots, j_R\}$. As in the proof of Proposition D.26, we consider all adjacency patterns of A, B and C. Then, we can easily check the following case statement:

1. Suppose that either "$j_L = 1$ and $j_R + 1 \notin B$" or "$j_L - 1 \notin B$ and $j_R + 1 \notin B$" holds. Then, we have $F(B \cup A) - F(B) = F(A) = c_{j_R}$ and $F(B \cup C) - F(B) = F(C)$. Now, we will check (73) under the concavity condition (31). First, (73) trivially holds when $j_R \in C$ because in this case $F(C) \geq c_{j_R} = F(A)$. Next, we assume $j_R \notin C$. Let i be the largest element in C. Then, we have $F(C) \geq c_i$, $|C| \leq i - j_L + 1$. Under the assumption (31), we have

$$\begin{aligned} \frac{|C|}{|A|} F(A) &\leq \frac{i - j_L + 1}{j_R - j_L + 1} c_{j_R} \\ &\leq \frac{i}{j_R} c_{j_R} \quad (\because j_L \leq i < j_R) \\ &\leq c_i \quad (\because (31)) \\ &\leq F(C), \end{aligned}$$

which implies (73).

2. Suppose that $j_L - 1 \in B$ and $j_R + 1 \notin B$. Then, we have $F(B \cup A) - F(B) = c_{j_R} - c_{j_L - 1}$ and $F(B \cup C) - F(B) \geq F(C) - c_{j_L - 1}$. By a similar argument above, (73) trivially holds when $j_R \in C$. Let $j_R \notin C$ and let i be the largest element in C. Then, under the assumption (31), we have

$$\begin{aligned}\frac{|C|}{|A|}(F(B \cup A) - F(B)) &\leq \frac{i - j_L + 1}{j_R - j_L + 1}(c_{j_R} - c_{j_L - 1}) \\ &\leq c_i - c_{j_L - 1} \quad (\because (31)) \\ &\leq F(C) - c_{j_L - 1} \\ &\leq F(B \cup C) - F(B).\end{aligned}$$

3. For other case, we have $F(B \cup A) - F(B) \leq F(B \cup C) - F(B)$, which implies (73).

Appendix E: Proofs in Section 5

The goal of this section is to prove Theorem 5.1. The outline of the proof is essentially the same as the framework of Theorem 4.18 in Massart (2007). We explain this framework in Section E.1. To complete the proof, we have to control the maximum value of a certain normalized Gaussian process. For this, we provide an upper bound in Section E.2.

E.1. Proof overview

Let $(\hat{\Pi}, \hat{\mathbf{V}})$ be the selected pair in (27). Fix any connected partition Π and $\mathbf{V} \in \mathscr{V}(|\Pi|)$. By the definition of the estimator, we have

$$\begin{aligned}\|y - \hat{\theta}_{\hat{\Pi}, \hat{\mathbf{V}}}\|_2^2 + \mathrm{pen}(\hat{\Pi}, \hat{\mathbf{V}}) &\leq \|y - \hat{\theta}_{\Pi', \mathbf{V}'}\|_2^2 + \mathrm{pen}(\Pi', \mathbf{V}') \\ &\leq \|y - \theta'\|_2^2 + \mathrm{pen}(\Pi', \mathbf{V}')\end{aligned}$$

for any vector θ' that belongs to $K^{\uparrow}_{\Pi'}(\mathbf{V}')$. In particular, we can choose θ' as

$$\theta' = \theta^*_{\Pi', \mathbf{V}'} := \operatorname*{argmin}_{\theta' \in K^{\uparrow}_{\Pi'}(\mathbf{V}')} \|\theta' - \theta^*\|_2.$$

Substituting $y = \theta^* + \xi$, we can deduce that

$$\|\theta^* - \hat{\theta}_{\hat{\Pi}, \hat{\mathbf{V}}}\|_2^2 \leq \|\theta^* - \theta^*_{\Pi', \mathbf{V}'}\|_2^2 - \mathrm{pen}(\hat{\Pi}, \hat{\mathbf{V}}) + \mathrm{pen}(\Pi', \mathbf{V}') + 2\langle \hat{\theta}_{\hat{\Pi}, \hat{\mathbf{V}}} - \theta^*_{\Pi', \mathbf{V}'},\ \xi \rangle. \tag{76}$$

Here, recall that ξ is a random variable drawn from $N(0, \sigma^2 I_n)$.

Let $z > 0$ be a positive number and $c \in (0, 1)$. Suppose that an inequality

$$\max_{\Pi} \sup_{\mathbf{V} \in \mathscr{V}(|\Pi|)} \sup_{\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})} \frac{\langle \theta - \theta^*_{\Pi', \mathbf{V}'},\ \xi \rangle}{(\|\theta - \theta^*\|_2 + \|\theta' - \theta^*\|_2)^2 + \eta(\Pi, \mathbf{V}, z)} \leq \frac{c}{4} \tag{77}$$

holds on some event Ω_z that occurs with probability at least $1 - \mathrm{e}^{-z}$. Here, $\eta(\Pi, \mathbf{V}, z) > 0$ is a positive constant that can depend on $\Pi, \mathbf{V}, z$. Combining this inequality with (76), we have on the same event

$$(1-c)\|\theta^* - \hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}}\|_2^2 \le (1+c)\|\theta^* - \theta^*_{\Pi',\mathbf{V}'}\|_2^2 - \mathrm{pen}(\hat{\Pi},\hat{\mathbf{V}}) + \mathrm{pen}(\Pi',\mathbf{V}') + c\eta(\hat{\Pi},\hat{\mathbf{V}},z), \tag{78}$$

where we used the elementary inequality $(a+b)^2 \le 2(a^2+b^2)$.

E.2. Controlling the normalized process

Now, our goal is to provide an inequality of the form (77). Below, we fix $\theta' := \theta^*_{\Pi',\mathbf{V}'}$.

First, we fix a partition Π and $\mathbf{V} \in \mathscr{V}(|\Pi|)$. For any $\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})$, we define

$$\omega(\theta) = \omega_{\Pi,\mathbf{V}}(\theta) := (\|\theta - \theta^*\|_2 + \|\theta' - \theta^*\|_2)^2 + \eta,$$

where $\eta > 0$ is a positive constant which will be specified later. Define a random variable $Z_{\Pi,\mathbf{V}}$ as

$$Z_{\Pi,\mathbf{V}} := \sup_{\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})} \frac{\langle \theta - \theta', \xi \rangle}{\omega(\theta)}.$$

Note that $Z_{\Pi,\mathbf{V}}$ is the supremum of a sample-continuous Gaussian process. By the concentration inequality for Gaussian processes (Lemma F.1), we have

$$\Pr\left\{ Z_{\Pi,\mathbf{V}} - \mathbb{E}[Z_{\Pi,\mathbf{V}}] \ge \sqrt{2v(x+z)} \right\} \le \exp(-(x+z)) \tag{79}$$

for any $x > 0$ and $z > 0$. Here, the variance v is bounded as

$$v := \sup_{\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})} [Z^2_{\Pi,\mathbf{V}}] \le \frac{\sigma^2}{4\eta}$$

because $\omega(\theta) \ge \|\theta - \theta'\|_2^2 + \eta \ge 2\eta^{1/2}\|\theta - \theta'\|_2$, and $\langle u, \xi \rangle$ is distributed according to $N(0, \sigma^2\|u\|_2^2)$ for any $u \in \mathbb{R}^n$.

We will provide an upper bound for $\mathbb{E}[Z_{\Pi,\mathbf{V}}]$. Let $\theta^*_{\Pi,\mathbf{V}}$ be the orthogonal projection of θ^* onto $K^{\uparrow}_{\Pi}(\mathbf{V})$. Note that

$$\mathbb{E}[Z_{\Pi,\mathbf{V}}] \le \underbrace{\mathbb{E}\left[\sup_{\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})} \frac{\langle \theta - \theta^*_{\Pi,\mathbf{V}},\ \xi \rangle}{\omega(\theta)} \right]}_{\text{(a)}} + \underbrace{\mathbb{E}\left[\frac{|\langle \theta^*_{\Pi,\mathbf{V}} - \theta',\ \xi \rangle|}{\inf_{\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})} \omega(\theta)} \right]}_{\text{(b)}}. \tag{80}$$

The second term (b) in the right-hand side of (80) is bounded from above by $\sigma\eta^{-1/2}$. Indeed, since

$$\inf_{\theta \in K^{\uparrow}_{\Pi}(\mathbf{V})} \omega(\theta) = (\|\theta^*_{\Pi,\mathbf{V}} - \theta^*\|_2 + \|\theta' - \theta^*\|_2)^2 + \eta \ge 2\eta^{1/2}\|\theta^*_{\Pi,\mathbf{V}} - \theta'\|_2,$$

we have

$$\text{(b)} \le \frac{1}{2\sqrt{\eta}} \mathbb{E}_{u \sim N(0,\sigma^2)}[|u|] = \frac{\sigma}{\sqrt{2\pi\eta}}.$$

To bound the term (a) in (80), we use the following lemma:

Lemma E.1. Let $\Pi = (A_1, A_2, \dots, A_m)$ be any partition and $\mathbf{V} = (\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_m)$. Fix any $\bar{\theta} \in K_\Pi^\uparrow(\mathbf{V})$. For any $t > 0$, we have

$$\begin{aligned} &\mathbb{E}\left[\sup_{\theta \in K_\Pi^\uparrow(\mathbf{V}): \|\theta - \bar{\theta}\|_2 \le t} \langle \xi, \theta - \bar{\theta} \rangle\right] \\ &\le C\sigma t^{1/2} \left(\sum_{i=1}^m |A_i|^{1/3} \mathcal{V}_i^{2/3}\right)^{3/4} + C\sigma t \sqrt{m \log \frac{\mathrm{e}n}{m}}, \end{aligned} \tag{81}$$

where $C > 0$ is a universal constant. Futhermore, for any $\eta > 0$, we have

$$\begin{aligned} &\mathbb{E}\left[\sup_{\theta \in K_\Pi^\uparrow(\mathbf{V})} \frac{\langle \theta - \bar{\theta},\ \xi \rangle}{\|\theta - \bar{\theta}\|_2 + \eta}\right] \\ &\le 4C\sigma \left\{ \eta^{-3/4} \left(\sum_{i=1}^m |A_i|^{1/3} \mathcal{V}_i^{2/3}\right)^{3/4} + \eta^{-1/2} \sqrt{m \log \frac{\mathrm{e}n}{m}} \right\}, \end{aligned} \tag{82}$$

where C is the same constant as in (81).

Proof. We will prove the first inequality (81). Let $W := W(\Pi, \mathbf{V})$ denote the left-hand side of (81). We consider a collection of finitely many sets $S(\mathbf{q})$ as follows: Let $\mathcal{Q} := \mathcal{Q}(m)$ be a collection of vectors $\mathbf{q} = (q_1, q_2, \dots, q_m)$ that can be written as $\mathbf{q} = t^2 \mathbf{a}/m$ for some integer vector $\mathbf{a} = (a_1, a_2, \dots, a_m)$ such that $1 \le a_i \le m$ and $\sum_{i=1}^m a_i \le 2m$. Note that, by Proposition D.10, the cardinality of $\mathcal{Q}$ is bounded by $(2\mathrm{e})^m$. For any $\mathbf{q} \in \mathcal{Q}$, define the set

$$S(\mathbf{q}) := \left\{ \theta \in \mathbb{R}^n : \|\theta_{A_i}\|_2^2 \le q_i,\ \mathcal{V}^{A_i}(\theta_{A_i}) \le 2\mathcal{V}_i \text{ for all } A_i \in \Pi \right\}.$$

Then, we can easily check that

$$K_\Pi^\uparrow(\mathbf{V}) \cap \{\theta \in \mathbb{R}^n : \|\theta - \bar{\theta}\|_2 \le t\} \subseteq \bigcup_{\mathbf{q} \in \mathcal{Q}} S(\mathbf{q}).$$

From Lemma F.3 below, there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left[\sup_{\theta \in S(\mathbf{q})} \langle \theta, \xi \rangle\right] \le C\sigma \sum_{i=1}^m \left\{ \sqrt{2} q_i^{1/4} |A_i|^{1/4} \mathcal{V}_i^{1/2} + q_i^{1/2} \sqrt{\log \mathrm{e}|A_i|} \right\}. \tag{83}$$

Here, by Hölder's inequality, we have

$$\begin{aligned} \sum_{i=1}^m q_i^{1/4} |A_i|^{1/4} \mathcal{V}_i^{1/2} &\le \left(\sum_{i=1}^m q_i\right)^{1/4} \left(\sum_{i=1}^m (|A_i|^{1/4} \mathcal{V}_i^{1/2})^{4/3}\right)^{3/4} \\ &\le 2^{1/4} t^{1/2} \left(\sum_{i=1}^m |A_i|^{1/3} \mathcal{V}_i^{2/3}\right)^{3/4}, \end{aligned}$$

and by the Cauchy–Schwarz inequality, we also have

$$\sum_{i=1}^{m} 2q_i^{1/2}\sqrt{\log \mathrm{e}|A_i|} \le 2\sqrt{2}t\left(\sum_{i=1}\log \mathrm{e}|A_i|\right)^{1/2} \le 2\sqrt{2}t\sqrt{m\log\frac{\mathrm{e}n}{m}}.$$

Then, by Lemma F.4 below, we have

$$\begin{aligned} W &\le \max_{\mathbf{q}\in\mathcal{Q}} \mathbb{E}\left[\sup_{v\in S(\mathbf{q})}\langle \xi, v\rangle\right] + 2t\sigma\left(\sqrt{2\log|\mathcal{Q}|}+\sqrt{\frac{\pi}{2}}\right) \\ &\le C\sigma\left\{2^{3/4}t^{1/2}\left(\sum_{i=1}^{m}|A_i|^{1/3}\mathcal{V}_i^{2/3}\right)^{3/4} + 2\sqrt{2}t\sqrt{m\log\frac{\mathrm{e}n}{m}}\right\} \\ &\quad + 2t\sigma\left(\sqrt{4m\log 2\mathrm{e}}+\sqrt{\frac{\pi}{2}}\right) \\ &\le C'\sigma\left\{t^{1/2}\left(\sum_{i=1}^{m}|A_i|^{1/3}\mathcal{V}_i^{2/3}\right)^{3/4} + t\sqrt{m\log\frac{\mathrm{e}n}{m}}\right\} \end{aligned}$$

for some $C' > 0$. Thus, (81) has been proved.

The second inequality (82) is a consequence of the peeling lemma (Lemma F.2 below). □

Combining (79), (80) and (82), we conclude that

$$\begin{aligned} Z_{\Pi,\mathbf{V}} &\le 4C\sigma\eta^{-3/4}\left(\sum_{i=1}^{m}|A_i|^{1/3}\mathcal{V}_i^{2/3}\right)^{3/4} \\ &\quad + \sigma\eta^{-1/2}\left\{4C\sqrt{m\log\frac{\mathrm{e}n}{m}} + (2\pi)^{-1/2} + 2^{-1/2}\sqrt{x+z}\right\} \end{aligned} \tag{84}$$

holds with probability at least $1-\exp(-(x+z))$, where C is the constant in (82). Now, we choose the two constant $\eta := \eta(\Pi, \mathcal{V}, z)$ and $x := x(\Pi, \mathcal{V})$ as

$$\eta(\Pi,\mathcal{V},z) := 2^8(4C+1)^{4/3}\sum_{i=1}^{m}\sigma^{4/3}|A_i|^{1/3}\mathcal{V}_i^{2/3} + 2^8(4C+2)^2\sigma^2 m\log\frac{\mathrm{e}n}{m} + 2^8\sigma^2 z$$

and

$$x(\Pi,\mathcal{V}) := \sum_{i=1}^{m}\sigma^{-2/3}|A_i|^{1/3}\mathcal{V}_i^{2/3} + 2m\log\frac{\mathrm{e}n}{m},$$

respectively. Then, it is elementary to check that the right-hand side of (84) is not larger than 1/8.

Applying the union bound over all pairs $(\Pi, \mathbf{V})$, we have

$$\Pr\left\{\max_{\Pi}\sup_{\mathbf{V}\in\mathscr{V}(|\Pi|)} Z_{\Pi,\mathbf{V}} > \frac{1}{8}\right\} \le \exp(-z)\sum_{\Pi}\sum_{\mathbf{V}}\exp(-x(\Pi,\mathbf{V})).$$

Here, we can show that

$$\sum_{\Pi}\sum_{\mathbf{V}}\exp(-x(\Pi,\mathbf{V}))\le 1, \tag{85}$$

and hence we conclude that (77) holds with $c=1/2$. Indeed, (85) follows from the fact that, for any Π,

$$\begin{aligned}\sum_{\mathbf{V}\in\mathscr{V}(\Pi)}\exp\left(-\sum_{i=1}^{m}\sigma^{-2/3}|A_i|^{1/3}\mathcal{V}_i^{2/3}\right)&=\prod_{i=1}^{m}\exp\left(-\sigma^{-2/3}|A_i|^{1/3}\right)\left(\sum_{j_i=1}^{\infty}\mathrm{e}^{-j_i}\right)\\&\le\exp\left(-\sum_{i=1}^{m}\sigma^{-2/3}|A_i|^{1/3}\right)\le 1\end{aligned}$$

and

$$\begin{aligned}\sum_{\Pi}\exp\left(-2|\Pi|\log\frac{\mathrm{e}n}{|\Pi|}\right)&=\sum_{m=1}^{n}\sum_{\Pi:|\Pi|=m}\exp\left(-2m\log\frac{\mathrm{e}n}{m}\right)\\&\le\sum_{m=1}^{n}\sum_{\Pi:|\Pi|=m}\exp\left(-m-\log\binom{n-1}{m-1}\right)\\&=\sum_{m=1}^{n}\mathrm{e}^{-m}\le 1.\end{aligned}$$

E.3. Proof of Theorem 5.1

Now, we are ready to complete the proof of Theorem 5.1. Define $\mathrm{pen}(\Pi,\mathbf{V})$ as

$$2^7(4C+1)^{4/3}\sum_{i=1}^{m}\sigma^{4/3}|A_i|^{1/3}\mathcal{V}_i^{2/3}+2^7(4C+2)^2\sigma^2 m\log\frac{\mathrm{e}n}{m},$$

where C is the constant in (82). Let $(\Pi',\mathbf{V}')$ be the pair that minimizes

$$(\Pi,\mathbf{V})\mapsto\frac{3}{2}\|\theta^*-\theta^*_{\Pi,\mathcal{V}}\|_2^2+\mathrm{pen}(\Pi,\mathbf{V})$$

among all possible pairs. Applying (78) and (77) for this choice of $(\Pi',\mathbf{V}')$, we conclude that

$$\|\hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}}-\theta^*\|_2^2\le\min_{\Pi}\min_{\mathbf{V}\in\mathscr{V}(|\Pi|)}\left\{3\mathrm{dist}^2(\theta^*,K_{\Pi}^{\uparrow}(\mathbf{V}))+2\mathrm{pen}(\Pi,\mathbf{V})\right\}+2^8\sigma^2 z$$

holds with probability at least $1-\exp(-z)$. Moreover, by integrating both sides with respect to z, we have

$$\mathbb{E}_{\theta^*}\|\hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}}-\theta^*\|_2^2\le\min_{\Pi}\min_{\mathbf{V}\in\mathscr{V}(|\Pi|)}\left\{3\mathrm{dist}^2(\theta^*,K_{\Pi}^{\uparrow}(\mathbf{V}))+2\mathrm{pen}(\Pi,\mathbf{V})\right\}+2^8\sigma^2.$$

Appendix F: Auxiliary lemmas

Here, we present several auxiliary lemmas that are used in the proofs in the previous sections.

Lemma F.1 (Borel–Tsirelson–Ibragimov–Sudakov inequality; see Proposition 3.19 in Massart (2007))**.** Suppose that $(X_t)_{t\in T}$ is a Gaussian process on a totally bounded metric space (T,d) such that $\mathbb{E}[X_t]=0$ for any $t\in T$ and the sample path $t\mapsto X_t$ is almost surely continuous. Let $v:=\sup_{t\in T}\mathbb{E}[X_t^2]$. Then, for any $z>0$, we have

$$\Pr\left\{\sup_{t\in T}X_t-\mathbb{E}\left[\sup_{t\in T}X_t\right]\geq\sqrt{2vz}\right\}\leq\exp(-z).$$

Lemma F.2 (Peeling lemma; see e.g. Lemma 4.23 in Massart (2007))**.** Let K be a set in $\mathbb{R}^n$ and $\bar{\theta}\in K$. Assume that there is a function $\psi:[0,\infty)\to\mathbb{R}$ such that $\psi(t)/t$ is non-increasing and

$$\mathbb{E}_{\xi\sim N(0,I_n)}\left[\sup_{\theta\in K:\|\theta-\bar{\theta}\|_2\leq t}\langle\xi,\theta-\bar{\theta}\rangle\right]\leq\psi(t)$$

for any $t\geq\bar{t}\geq 0$. Then, for any $x\geq\bar{t}$, we have

$$\mathbb{E}_{\xi\sim N(0,I_n)}\left[\sup_{\theta\in K:\|\theta-\bar{\theta}\|_2\leq t}\frac{\langle\xi,\theta-\bar{\theta}\rangle}{\|\theta-\bar{\theta}\|_2^2+x^2}\right]\leq\frac{4\psi(x)}{x^2}.$$

Lemma F.3 (Guntuboyina et al. (2020), Lemma B.1)**.** For any $t>0$ and $\mathcal{V}>0$, let

$$S(V,t):=\{\theta\in\mathbb{R}^n:\mathcal{V}(\theta)\leq\mathcal{V}\text{ and }\|\theta\|_2\leq t\}.$$

There exists a universal constant $C>0$ such that

$$\mathbb{E}_{\xi\sim N(0,\sigma^2I_n)}\left[\sup_{\theta\in S(V,t)}\langle\theta,\xi\rangle\right]\leq C\sigma t^{1/2}n^{1/4}\mathcal{V}^{1/2}+C\sigma t\sqrt{\log en}.$$

Lemma F.4 (Guntuboyina et al. (2020), Lemma D.1)**.** Suppose $p,n\geq 1$ and let $\Theta_1,\dots,\Theta_p$ be subset of $\mathbb{R}^n$ each containing the origin and each contained in the closed Euclidean ball of radius D centered at the origin. Then, for $\xi\sim N(0,\sigma^2I)$, we have

$$\mathbb{E}\left[\max_{1\leq i\leq p}\sup_{\theta\in\Theta_i}\langle\xi,\theta\rangle\right]\leq\max_{1\leq i\leq p}\mathbb{E}\left[\sup_{\theta\in\Theta_i}\langle\xi,\theta\rangle\right]+D\sigma\left(\sqrt{2\log p}+\sqrt{\frac{\pi}{2}}\right).\tag{86}$$

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP17J06640. The author would like to thank three anonymous reviewers for their valuable comments and suggestions. The author also thanks Hiromichi Nagao for suggesting the example of a seismological phenomenon, and Fumiyasu Komaki and Keisuke Yano for helpful discussions.

References

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transition in convex programs with random data. *Information and Inference: A Journal of IMA*, 3:224–294, 2014. MR3311453

M. Ayer, H. D. Brunk, G. M. Ewing, W.T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26:641–647, 1955. MR0073895

F. Bach. Shaping level sets with submodular functions. In *NIPS*, 2011.

F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2–3):143–373, 2013.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. MR2486527

P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018. MR3782383

P. C. Bellec and A. B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16:1879–1892, 2015. MR3417801

L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268, 2001. MR1848946

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. MR3185193

H. D. Brunk. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26:607–616, 1955. MR0073894

S. Chatteejee, A. Guntuboyina, and B. Sen. On matrix estimation under monotonicity constraints. *Bernoulli*, 24(2):1072–1100, 2018. MR3706788

S. Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42:2340–2381, 2014. MR3269982

S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43:1774–1800, 2015. MR3357878

X. Chen, Q. Lin, and B. Sen. On degrees of freedom of projection estimators with applications to multivariate shape restricted regression. *arxiv preprint* arXiv:1509.01877, 2015. MR4078455

A. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1), 2017. MR3556784

D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B*, 54(1):41–81, 1992. MR1157714

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, second edition, 2014. MR3445285

Z. Fan and L. Guan. Approximate l_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46(6B):3217–3245, 2018. MR3852650

G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18:30–55, 1989. MR0978165

C. Gao, F. Han, and C.-H. Zhang. On estimation of isotonic piecewise constant signals. arXiv:1705.06386, 2017.

P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, 2014. MR3445293

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594, 2018. MR3881209

A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics*, 48(1):205–229, 2020. MR4065159

Q. Han and J. A. Wellner. Multivariate convex regression: global risk bounds and adaptation. arXiv:1601.06844, 2016.

Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019. MR3988762

K. Hattori and T. Hattori. Sales ranks, Burgers-like equations, and least-recently-used caching. In *RIMS Kokyuroku Bessatsu*, pages 149–162, 2010. MR2792027

C. Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49:598–619, 1954. MR0065093

D. S. Hochbaum and M. Queyranne. Minimizing a convex cost closure set. *SIAM Journal of Discrete Mathematics*, 16(2):192–207, 2003. MR1982135

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *SIAM Review, problems and techniques section*, 51(2):339–360, 2009. MR2505584

R. Kyng, A. Rao, and S. Sachdeva. Fast, provable algorithms for isotonic regression in all ℓ_p-norms. In *NIPS*, 2015. MR3631020

Y. T. Lee, A. Sidford, and S. S. Vempala. Efficient convex optimization with membership oracles. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1292–1294, 2018.

K. Lin, J. L. Sharpnack, A. Rinaldo, and R. J. Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *NIPS*, 2017.

R. Luss and S. Rosset. Bounded isotonic regression. *Electronic Journal of Statistics*, 11:4488–4514, 2017. MR3724487

J. Mairal, R. Janatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011. MR2845676

E. Mammen and S. van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25:387–413, 1997. MR1429931

P. Massart. *Concentration Inequalities and Model Selection*. Springer, 2007. MR2319879

T. I. Melbourne, W. M. Szeliga, M. Santillan, T. A. Herring, M. A. Floyd, and R. W. King. GAGE processing GPS plate boundary observatory expanded analysis product for 2017: Final position time series; constrained position time series from Central Washington University (analysis center) in NAM08

and IGS08 reference frames produced by the Massachusetts Institute of Technology (analysis center coordinator), 2018.

M. Meyer and M. Woodroofe. On the degrees of freedom in shape-restricted regression. *The Annals of Statistics*, 28(4):1083–1104, 2000. MR1810920

H. Nagao, T. Higuchi, S. Miura, and D. Inazu. Time-series modeling of tide gauge records for monitoring of the crustal activities related to oceanic trench earthquakes around Japan. *The Computer Journal*, 56(3):355–364, 2013.

G. Obozinski and F. Bach. A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. hal-01412385, 2016.

S. Oymak and B. Hassibi. Sharp MSE bound for proximal denoising. *Foundations of Computational Mathematics*, 16:965–1029, 2016. MR3529131

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011. MR2882274

T. Robertson and F. T. Wright. Consistency in generalized isotonic regression. *The Annals of Statistics*, 3:350–362, 1975. MR0365871

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley, 1988. MR0961262

T. Rockafeller and R. Wets. *Variational Analysis*. Springer, 1998. MR1491362

G. Roggers and H. Dragert. Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip. *Science*, 300(5627):1942–1943, 2003.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012001030, 2007. MR2341696

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. MR3363401

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005. MR2136641

R. J. Tibshirani, H. Höfling, and R. Tibshirani. Nearly-isotonic regression. *Technometrics*, 53:54–61, 2011. MR2791946

S. van de Geer. *Estimation and Testing Under Sparsity*. Springer, 2015. MR3526202

C. van Eeden. Maximum likelihood estimation of ordered probabilities. *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings. Series A*, 59(18):444–455, 1956.

R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. MR3837109

M. Woodroofe and J. Sun. A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing. *Statistica Sinica*, 3(2):501–515, 1993. MR1243398

J. Wu, M. C. Meyer, and J. D. Opsomer. Penalized isotonic regression. *Journal of Statistical Planning and Inference*, 161:12–24, 2015. MR3316548

C.-H. Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002. MR1902898