博士論文

Statistical Learning with Structured Low-Dimensionality

(構造的低次元性を伴う統計的学習)

南 賢太郎

Abstract

A widely accepted principle in statistical learning is that a good estimator is obtained through a good control of model complexity. Typically, low-complexity models are obtained as low-dimensional sub-structures of a single large model. In this thesis, we focus on the problem of selecting the best model among a large number of candidates for low-dimensional structures. In particular, we study statistical properties of regularization methods that can induce complex low-dimensional structures.

Our first contribution is about risk estimation in regularization methods for structured sparsity. Over the last two decades, structured sparsity have been a major research topic in high-dimensional statistics and machine learning. Since penalized model selection criteria for structured sparsity can involve hard combinatorial optimization problems, various regularization methods have been developed as their computationally tractable convex relaxations. However, regularization methods can produce their own stochastic errors, although the original intension of penalized model selection criteria is to cancel the effect of stochastic fluctuations associated with each low-dimensional model. In this thesis, we study submodular regularization, which is a wide class of regularization based estimators related to submodular functions. We derive unified formulae for unbiased risk estimates of submodular regularization. Our formulae can be applied for any submodular regularization estimators and any design matrices. Our results also recover some existing results for regularization and projection type estimators, such as the lasso, the fused lasso, and the isotonic regression. Moreover, we show that, in submodular regularization, the computational complexity of calculating unbiased risk estimates can be much faster than other general class of polyhedral convex regularizers. We also provide some numerical experiments that show reasonable effectiveness of our formulae as selection criteria of regularization parameter.

As our second contribution, we study the problem of estimating piecewise monotone vectors. This problem can be seen as a generalization of the classical isotonic regression that allows a small number of order-violating changepoints. We mainly focus on the performance of the nearly-isotonic regression, which can be regarded as an example of submodular regularization. We derive risk bounds for the nearly-isotonic regression estimators that are adaptive to piecewise monotone signals. Under a weak assumption, the estimator achieves a nearly minimax convergence rate over certain classes of piecewise monotone signals. We also 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.

Contents

Chapter 1

Introduction

1.1 Background: Statistical learning, model complexity, and dimensionality

A fundamental goal in data-driven science is to understand what we can learn from observed data. Until today, a variety of models for learning have been developed and studied in statistics, machine learning, information theory, artificial intelligence and related fields. Notably, many successful applications in modern artificial intelligence have widely adopted the idea that learning can be formulated as statistical procedures.

Statistical learning theory is a research field devoted to providing statistical guarantees for various learning algorithms. In statistical learning theory, the algorithms are typically analyzed in a decision-theoretic framework. An informal explanation of this framework is as follows; The observed data is assumed to be drawn from an unknown "true" probability distribution. The purpose of the statistician (or perhaps decision makers, data scientists, machine learning algorithms, and so on) is to estimate the unobserved population quantity that depends on the true distribution. To this end, the statistician makes an estimator as an observable counterpart of the desired population quantity. The performance of such an estimator is measured by a certain loss function, which is to be minimized. Therefore, a possible goal for the statistician is to find an optimal estimator that minimizes the loss function.

A widely accepted principle in statistical learning is that a good estimator is obtained through a good control of the **model complexity**. Here, the *model* means a set within which the statisticians search estimators. While the precise definition of the model complexity depends on the context, it can roughly be regarded as the size or the capacity of the model.

We here provide some intuition behind how the model complexity affects the statistical performance of the estimator. Suppose that $\underline{X} = (X_1, X_2, \ldots, X_n)$ are n observations from an unknown distribution P. Given a model Θ, a typical candidate of the estimator is given by the minimizer of an empirical objective $R_{\underline{X}}(\theta)$ among Θ, that is, $\hat{\theta}(\underline{X})$ is defined as

$$\hat{\theta}(\underline{X}) \in \operatorname*{argmin}_{\theta \in \Theta} R_{\underline{X}}(\theta).$$

Here, we expect that the quantity $R_{\underline{X}}(\theta)$ mimics the true population objective $R_P(\theta)$ in some sense. For example, $R_{\underline{X}}(\theta)$ is often taken as an unbiased or asymptotically unbiased estimator of $R_P(\theta)$. This formulation covers many important estimators in statistics and machine learning literature, such as the maximum likelihood estimators, the M-estimators, and the empirical risk minimization. Then, the statistical performance of the estimator $\hat{\theta}$ is affected by the size of the model Θ as follows. In principle, the model should be taken large enough so that it contains a sufficiently accurate approximation of the population quantity θ_P. If we enlarge the model Θ, the empirical objective $R_{\underline{X}}(\theta)$ can be made small, which means that the data is well fitted by the estimator. On the other hand, due to the randomness of the observation $\underline{X}$, a larger model allows a larger fluctuation of the empirical objective $R_{\underline{X}}$. Therefore, using too large models may cause a large gap between the performance of the estimator and the optimal one. In this sense, there is a trade-off between the approximation error (i.e., **bias**) and the stochastic error (i.e., **variance**).

In many cases, the true objective $R_P(\theta)$ can be decomposed as the sum of two terms contributed by the bias and the variance respectively:

$$R_P(\theta) = \underbrace{\mathrm{Bias}(\theta)}_{\approx R_{\underline{X}}(\theta)} + \underbrace{\mathrm{Variance}(\theta)}_{\approx \text{model complexity}} .$$

Moreover, in the above decomposition, the bias term is often estimated by the empirical objective $R_{\underline{X}}(\hat{\theta})$ itself, while the variance term is estimated by some complexity measure of the model. Such decomposition phenomena have been found and studied ubiquitously in many different frameworks, and the idea has been implemented as model selection criteria penalized with model complexity measures. The following is a partial list of such frameworks:

- **AIC:** The Akaike Information Criterion (AIC) (Akaike 1973) is a model selection criterion for parametric statistical models. In the AIC, the model complexity is measured by the dimension of the model. Under some regularity conditions for the model, the AIC is shown to be an asymptotically unbiased estimator of the Kullback–Leibler risk.
- **Mallows' C_p:** Mallows' C_p (Mallows 1973) is a criterion for variable selection in the linear regression model. The penalty term in the C_p statistics is given as the number of variables (or equivalently the dimension) of the linear model. In fact, Mallows' C_p coincides with the AIC when the noise distribution is assumed to be isotropic Gaussian with a known variance, while the derivation can be justified in a non-asymptotic sense.
- **BIC:** The Bayesian Information Criterion (BIC) (Schwarz 1978) is another criterion based on Bayesian testing procedure. The BIC is obtained as an asymptotic leading term of the marginal likelihood, and eventually the penalty term is proportional to the dimension of the model.
- **MDL:** The Minimum Description Length (MDL) principle (Rissanen 1978) is an information theoretic model selection criterion originated in compression theory. The MDL chooses a model that has minimal code length of a certain two-part coding model. In some regular models, the MDL criterion involves a dimension penalty which is very similar as the BIC, although the underlying philosophy seems quite different.

A sensible criterion for a model to have low-complexity is **low-dimensionality**. That is, models with low-complexity are typically obtained just as low-dimensional sub-models of a single large model. Indeed, the aforementioned five frameworks serve examples that the model complexities are determined by model dimensions. For a more concrete example, let us consider the classical variable selection problem in linear regression. Suppose that the target variable can be represented as a linear combination of a given collection of explanatory variables. If only a few number of explanatory variables are actually needed to represent the target variable, the best model in prediction tasks should contain only a small number of explanatory variables. It is because, by increasing the number of explanatory variables, the approximation errors are not much improved while the model complexity terms get larger.

But how can we believe that our data can be represented by low-dimensional models? Nowadays, data and models addressed in the statistical learning are getting more and more complicated. Machine learning algorithms have come to address high-dimensional and large-scale datasets such as natural images, raw audio signals, and explosively increasing logs generated by web services. Meanwhile, the models are also getting more complicated. For example, deep neural networks have become a gold standard model in various modern machine learning applications (Goodfellow et al. 2016), while the theoretical analysis of their statistical and computational aspects remains largely uninvestigated. Still, we have a brief, or an *inductive bias*, that meaningful data should have meaningful structure. Natural images should be far more structured than pure noise images with uncorrelated pixels. Natural language should have much more structural constraints than word sequences generated completely at random. These structures themselves can be quite complicated, and we may not be able to write down them explicitly. However, such structures, albeit unknown, may enforce the data to have small degrees of freedom, and thus the low-dimensionality assumption on models can be still valid under some appropriate justifications.

1.2 Thesis goal: Structured low-dimensionality

The goal of this thesis is to investigate how we can exploit the **structured low-dimensionality** in statistical learning. We have hitherto explained the importance low-dimensional structures in learning theory. Hereafter, setting aside general problems in learning theory, we focus on the problem of selecting the best model among a large number of candidates for low-dimensional structures.

Let $\{\Theta_m : m \in \mathcal{M}\}$ be a collection of models. To choose a data-dependent model $\Theta_{\hat{m}}$, we consider a penalized model selection criterion as follows:

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left(\min_{\theta \in \Theta_m} R_{\underline{X}}(\theta) + \operatorname{pen}(\Theta_m, \underline{X}) \right) \tag{1.1}$$

Here, for each model Θ_m, $\operatorname{pen}(\Theta_m, \underline{X})$ is a penalty term that depends on the (effective) dimensionality of the model. As we mentioned before, model selection criteria of this form have been ubiquitously used in statistical learning theory. As an optimization problem, the above model selection is often computationally difficult due to the **combinatorial** aspect of the collection of models $\mathcal{M}$. Here, we give two examples:

- **Variable selection:** Let us consider the problem of choosing a best subset among a given set of explanatory variables $X_1, \ldots, X_p$. In this case, a model m corresponds to some subset of indices $[p] := \{1, \ldots, p\}$, and thus the cardinality of $\mathcal{M}$ can grow exponentially with p.
- **Change-point models:** Let $\theta = (\theta_1, \ldots, \theta_n)$ be a discrete signal. Suppose that there is a connected partition $A_1, \cdots A_k$ of $[n]$ such that each piece θ_{A_j}, $j \in \{1, \ldots, k\}$ belongs to a certain model of smooth signals (e.g., constant signals, linear signals, and Sobolev ellipsoids). In this case, $\mathcal{M}$ can contain the set of all connected partitions with cardinality 2^{n-1}.

In modern applications, the learning problem is often of large-scale in the sense that the sample size n or the data dimensionality p (or both) are extremely large. Hence, it can be quite hard to solve the above combinatorial model selection problem directly. To overcome such a curse of dimensionality in model selection, researchers have proposed

regularization methods based on convex relaxation of the discrete penalty $\mathrm{pen}(\Theta_m, \underline{X})$. Generally speaking, such regulization based estimators have the following form:

$$\hat{\boldsymbol{\theta}}_\lambda \in \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \left(R_{\underline{X}}(\boldsymbol{\theta}) + \lambda \Omega(\boldsymbol{\theta}) \right), \tag{1.2}$$

where the **regularizer** $\Omega : \mathbb{R}^p \to \mathbb{R}$ corresponds to the convex relaxation of $\mathrm{pen}(\Theta_m, \underline{X})$. This formulation includes the lasso (Tibshirani 1996), the group lasso (Yuan and Lin 2006), the fused lasso (Rudin et al. 1992, Tibshirani et al. 2005), and many other important regularization methods that have been proposed to date.

However, the regularization method (1.2) can produce its own stochastic error (i.e., variance), although the original intension of the criterion (1.1) is to cancel the effect of the stochastic error associated with each model. This motivates the following two research questions, which are main goals in this thesis:

(Q1) Can we obtain a general formula for stochastic errors (or variances) of the regularization based estimators (1.2)? How can we estimate them from data? Besides, if we can construct estimators for stochastic errors, a possible application is to use them for data-dependent criteria for selecting the tuning parameter λ. How effective are such criteria in practice?

(Q2) When the true distribution satisfies some structured low-dimensionality property, can the regularization based estimators (1.2) achieve optimal rate of statistical risks? If not, how much they are suboptimal compared to the best possible estimation by the (discrete) model selection procedures (1.1)?

The above questions will be addressed in Chapter 4 and Chapter 5, respectively. Below, we introduce particular problems addressed in this thesis.

1.2.1 Risk estimate in structured sparsity

In Chapter 4, we study a wide class of regularization based estimators that induces **structured sparsity** in linear regression problems. Roughly speaking, structured sparsity means that the sparsity pattern of the coefficient vector is determined by some combinatorial structure on the index set. For example, the coordinate sparsity (or the vanilla sparsity) stands for the property that each coordinate can independently shrink to zero, while the group sparsity requires that the support of the vector cannot be finer than some predefined partition. See Chapter 4 as well as Section 3.3.2 for more detailed reviews of structured sparsity.

We focus on **submodular regularization**, which is a class of regularization based estimators defined by convex relaxations of **submodular functions**. It has been pointed out that submodular regularization contains a wide class of existing penalties that induce structured sparsity (Bach 2010, 2011, Obozinski and Bach 2016). Besides, submodular regularization has computational advantage that there are efficient algorithms for calculating proximal operators (Bach 2013, Obozinski and Bach 2016). However, their statistical risk behaviours have not been fully understood.

We derive unified formulae for the **degrees of freedom**, a covariance penalty related to **Stein's Unbiased Risk Estimate** (SURE) (Stein 1981, Efron 2004). Our formulae can be applied for any submodular regularization estimators and any design matrices. As particular applications, our results provide new formulae for risk unbiased estimators for the SLOPE estimator (Bogdan et al. 2015) and the hypergraph total variation (Hein et al. 2013). Moreover, we point out that, in submodular regularization, the computational complexity of calculating SURE can be much faster than other general class of polyhedral convex regularizers. We also provide some numerical experiments that show reasonable effectiveness of the degrees of freedom as a selection criterion of regularization parameter.

1.2.2 Piecewise shape restricted regression

Estimation of monotone signals has a long history in statistics. The most fundamental estimator for this problem is the **isotonic regression**, which is the least squares estimator onto the set of monotone signals (Ayer et al. 1955, Brunk 1955, van Eeden 1956). The isotonic regression is known to be minimax optimal for the problem of estimating monotone signals with bounded total variations (Zhang 2002). Roughly speaking, the optimal rate of monotone signal estimation has the same order as that of nonparametric regression with bounded first derivatives (see e.g., Chapter 9 of van de Geer (2009)) or bounded total variations (Mammen and van de Geer 1997). Moreover, Chatterjee et al. (2015) and Bellec (2018) recently showed that the isotonic regression also achieves an adaptive minimax rate for the piecewise constant monotone signals. Thus, the isotonic regression naturally combines aspects of smooth signal estimation and combinatorial model selection.

Broadly speaking, the isotonic regression is an example of shape restricted regression. Shape restricted regression is a subfield of nonparametric regression where the true regression functions are assumed to satisfy some shape constraints (e.g., monotonicity, unimodality and convexity). See Groeneboom and Jongbloed (2014) and Guntuboyina and Sen (2017) for detailed reviews of this field.

In Chapter 5, we study the signal denoising problem when the true signal is **piecewise monotone**, that is, the signal is obtained by concatenating a few number of monotone signals. In our terms, this is a good example of structured low-dimensionality. In particular, estimating piecewise monotone signals can be regarded as a hierarchical selection procedure that consists of (i) selecting a partition on which the restricted signals are monotone and (ii) estimating monotone (and piecewise constant) signal on each segment in the partition.

We focus on the performance of the **nearly-isotonic regression** proposed by Tibshirani et al. (2011). In fact, the nearly-isotonic regression is an example of submodular regularization that we study in Chapter 4. In Chapter 5, we first provide a minimax lower bound for piecewise monotone signal estimation. Then, we derive risk bounds for the nearly-isotonic regression estimators that are adaptive to piecewise monotone signals. Under a weak assumption, the estimator achieve a nearly minimax convergence rate over certain classes of piecewise monotone signals. We also provide some simulation results suggesting that the nearly-isotonic regression performs as well as the ideal estimator that knows the true positions of changepoints.

1.3 Structure of this thesis

The reminder of this thesis is organized as follows. In Chapter 2, we provide some mathematical background preliminaries required in the later chapters. In Chapter 3, we review the theory of statistical estimation under sparsity, which aims to provide a minimum background required for our main contributions in Chapter 4. Chapter 4 presents our first contribution on unbiased risk estimator for submodular regularization. Chapter 5 presents our second contribution on statistical analysis of piecewise monotone signal estimation. In Chapter 6, we conclude this thesis.

Chapter 2

Preliminaries

This chapter provides some mathematical background preliminaries required in the later chapters. Readers can skip this chapter and only refer back when they are needed.

2.1 Convex analysis

In this section, we provide some machinery of convex analysis. For basic definitions and terminologies, we refer readers to Rockafeller and Wets (1998) and Rockafellar (1970).

2.1.1 Subgradient calculus

Recall that a function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is convex if it satisfies $h(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq th(\boldsymbol{x}) + (1-t)h(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and $t \in [0,1]$. Below, we always assume that h is proper, i.e., there exists $\boldsymbol{x} \in \mathbb{R}^p$ such that $h(\boldsymbol{x}) < \infty$.

For a function $h : \mathbb{R}^p \to \mathbb{R}$, a vector $\boldsymbol{v} \in \mathbb{R}^p$ is called a **subgradient** of h at a point $\boldsymbol{x} \in \mathbb{R}^p$ if it satisfies $h(\boldsymbol{x}) - h(\boldsymbol{x}') \leq \boldsymbol{v}^\top(\boldsymbol{x} - \boldsymbol{x}')$ for any $\boldsymbol{x}' \in \mathbb{R}^p$. The set of all subgradients of h at $\boldsymbol{x}$ is called the **subdifferential** of h at $\boldsymbol{x}$, and denoted by $\partial h(\boldsymbol{x})$. Note that $\partial h(\boldsymbol{x})$ is nonempty for any $\boldsymbol{x} \in \mathbb{R}^p$ if h is convex.

The notion of subdifferential is important because it is related the first-order optimality condition of convex optimization problems. For our purpose, we use the following facts.

Lemma 2.1. Let $h, g : \mathbb{R}^p \to \mathbb{R}$ be convex functions.

(i) Suppose that h is differentiable at $\boldsymbol{x} \in \mathbb{R}^p$. Then, $\partial h(\boldsymbol{x}) = \{\nabla h(\boldsymbol{x})\}$.

(ii) For any $\boldsymbol{x} \in \mathbb{R}^p$, $\partial(h+g)(\boldsymbol{x}) = \partial h(\boldsymbol{x}) + \partial g(\boldsymbol{x})$ holds. Here, the summation in the right-hand side is understood as the Minkowski sum.

(iii) $\boldsymbol{x} \in \mathbb{R}^p$ minimizes h over $\mathbb{R}^p$ if and only if $\boldsymbol{0} \in \partial h(\boldsymbol{x})$.

The following corollary provides the first-order optimality condition for regularization type problems, which is used in Chapter 4.

Corollary 2.2. Let $\ell : \mathbb{R}^p \to \mathbb{R}$ be a differentiable convex function, and $\psi : \mathbb{R}^p \to \mathbb{R}$ be a convex function. Consider a problem of minimizing a function $\ell + \lambda\psi$ ($\lambda > 0$) over $\mathbb{R}^p$. Then, a necessary and sufficient condition for $\boldsymbol{\theta} \in \mathbb{R}^p$ to be globally optimal is that

$$-\nabla \ell(\boldsymbol{\theta}) \in \lambda \partial \psi(\boldsymbol{\theta}). \tag{2.1}$$

2.1.2 Special convex sets

For a closed convex set $C \subseteq \mathbb{R}^p$ and a point $\boldsymbol{x}$ on it, the **tangent cone** of C at x is defined as

$$T_C(\boldsymbol{x}) := \text{closure}(\{t(\boldsymbol{x}' - \boldsymbol{x}) : t \geq 0, \boldsymbol{x}' \in C\}).$$

We also define the **normal cone** of C at $\boldsymbol{x}$ as

$$N_C(\boldsymbol{x}) := \{\boldsymbol{z} \in \mathbb{R}^p : \langle \boldsymbol{z} - \boldsymbol{x}, \boldsymbol{x}' - \boldsymbol{x} \rangle \leq 0 \text{ for all } \boldsymbol{x}' \in C\}.$$

Let $C \subseteq \mathbb{R}^p$ be a **polyhedron**, that is, a set that can be written as an intersection of finitely many (say k) half spaces. By definition, there exist a matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)^\top \in \mathbb{R}^{k \times p}$ and a vector $\boldsymbol{b} \in \mathbb{R}^k$ such that $C = \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$. Given a point $\boldsymbol{x} \in C$, define a set of indices of satisfied equality constraints as $I(\boldsymbol{x}) = \{i \in [k] : \boldsymbol{a}_i^\top \boldsymbol{x} = b_i\}$. Then, we have a parametric representation of the normal cone of C at $\boldsymbol{x}$:

$$N_C(\boldsymbol{x}) = \left\{ \sum_{i \in I(\boldsymbol{x})} c_i \boldsymbol{a}_i : \ c_i \geq 0 \text{ for } i \in I(\boldsymbol{x}) \right\}$$

(see Theorem 6.46 of Rockafeller and Wets (1998)).

A **face** F of a polyhedron C is a nonempty subset of C such that there exists $\boldsymbol{x} \in \mathbb{R}^p$ satisfying $F = \text{argmax}_{\boldsymbol{z} \in C} \boldsymbol{z}^\top \boldsymbol{x}$. Any polyhedron has a finite number of faces.

2.1.3 Support functions

For any nonempty set $C \subseteq \mathbb{R}^p$, we define the **support function** $\Omega_C : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ as

$$\Omega_C(\boldsymbol{x}) := \sup_{\boldsymbol{z} \in C} \boldsymbol{z}^\top \boldsymbol{x}.$$

For any C, the support function Ω_C is convex and positive homogeneous, i.e.,

$$\Omega_C(\lambda \boldsymbol{x}) = \lambda \Omega_C(\boldsymbol{x}) \quad \text{for all } \lambda > 0, \boldsymbol{x} \in \mathbb{R}^p.$$

Moreover, if C is a bounded set, then $\Omega_C(\boldsymbol{x})$ is bounded for all $\boldsymbol{x} \in \mathbb{R}^p$.

The following lemma provides two equivalent expressions of the subdifferentials of support functions.

Lemma 2.3 (Rockafeller and Wets (1998), Corollary 8.25)**.** Let C be a closed, bounded and convex set. For any $\boldsymbol{x} \in \mathbb{R}^p$, the subdifferential of Ω_C is given as

$$\partial \Omega_C(\boldsymbol{x}) = \underset{\boldsymbol{z} \in C}{\text{argmax}}\, \boldsymbol{z}^\top \boldsymbol{x} = \{\boldsymbol{z} \in C : \boldsymbol{x} \in N_C(\boldsymbol{z})\}.$$

For a bounded polyhedron C and $\boldsymbol{x} \in C$, Lemma 2.3 implies that the subdifferential of the support function $\partial \Omega_C(\boldsymbol{x})$ coincides with one of the faces of C.

2.2 Submodular analysis

In this section, we review some definitions and properties related to submodular functions. Submodular functions are an important class of set functions in combinatorial optimization. We refer readers to Fujishige (2005) and Bach (2013) for general introduction to this concept.

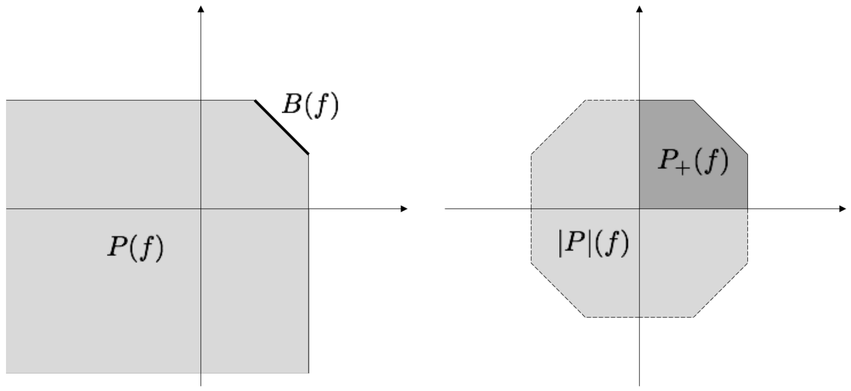

Fig. 2.1: **Examples of submodular polyhedra for** $p = 2$**.** (Left) The base polyhedron $B(f)$ is a face of the submodular polyhedron $P(f)$. The dimension of $B(f)$ is at most $p-1$. (Right) The positive submodular polyhedron $P_+(f)$ is the intersection of $P(f)$ and non-negative orthant $\mathbb{R}^p_+$. The symmetric submodular polyhedron $|P|(f)$ can be obtained as the union of 2^p axisymmetric copies of $P_+(f)$.

Let $V = [p] := \{1, \ldots, p\}$ be a finite set. We say that a function $f : 2^V \to \mathbb{R}$ is **submodular** if it satisfies the submodular inequality

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

for any $A, B \subseteq V$. Below, we always assume $f(\emptyset) = 0$. We say that a function $f : 2^V \to \mathbb{R}$ is non-decreasing if $f(A) \leq f(B)$ holds for any $A \subseteq B$.

2.2.1 Submodular polyhedra

Here, we introduce some polyhedra related to submodular functions.

Definition 2.4. Let $f : 2^V \to \mathbb{R}$ be a submodular function. We define the **submodular polyhedron** $P(f)$ and the **base polyhedron** $B(f)$ as follows:

$$P(f) := \{\boldsymbol{x} \in \mathbb{R}^p : \mathbf{1}_A^\top \boldsymbol{x} \leq f(A) \text{ for all } A \in 2^V\} \tag{2.2}$$

$$B(f) := \{\boldsymbol{x} \in P(f) : \mathbf{1}_V^\top \boldsymbol{x} = f(V)\}. \tag{2.3}$$

If f is non-decreasing, we also define the positive submodular polyhedron $P_+(f) := P(f) \cup \mathbb{R}^p_+$ and the **symmetric submodular polyhedron**

$$|P|(f) := \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}| \in P(f)\}. \tag{2.4}$$

If f is non-decreasing, $B(f)$ is included in the non-negative orthant $\mathbb{R}^p_+$. Thus, $P_+(f)$ shares all extremal points of $P(f)$. The symmetric submodular polyhedron $|P|(f)$ can be obtained as the union of axisymmetric copies of $P_+(f)$, i.e., $|P|(f) = \bigcup_{\gamma \in \{-1,1\}^p} \gamma \odot P_+(f)$. See Figure 2.1 for illustrative examples for $p = 2$.

Let $\boldsymbol{s}$ be any point in $P(f)$. A subset $A \subseteq V$ is said **tight** at $\boldsymbol{s}$ if it satisfies the equality constraint in (2.2) (i.e., $\mathbf{1}_A^\top \boldsymbol{s} = f(A)$). Let $\mathcal{D}(\boldsymbol{s})$ denote the collection of all tight sets at $\boldsymbol{s}$:

$$\mathcal{D}(\boldsymbol{s}) := \{A \subseteq V : \mathbf{1}_A^\top \boldsymbol{s} = f(A)\}.$$

It is known that $\mathcal{D}(\boldsymbol{s})$ becomes a distributive lattice[*1].

[*1] Here, a collection of sets $\mathcal{D} \subseteq 2^V$ is called a distributive lattice if, for any $A, B \in \mathcal{D}$, both $A \cap B$ and $A \cup B$ are contained in $\mathcal{D}$.

Lemma 2.5 ((Fujishige 2005), Lemma 2.2)**.** For any $\boldsymbol{s} \in P(f)$, the collection of tight sets $\mathcal{D}(\boldsymbol{s})$ is a distributive lattice with $\emptyset \in \mathcal{D}(\boldsymbol{s})$. Moreover, if $\boldsymbol{s} \in B(f)$, then $V \in \mathcal{D}(\boldsymbol{s})$.

The above lemma plays an important role in Chapter 4 for the following reasons. From a geometrical perspective, there is a one-to-one correspondence between the distributive lattices obtained by Lemma 2.5 and the faces of the base polyhedron $B(f)$. From a combinatorial perspective, a distributive lattice naturally defines a partition, which is important in the representation of the degrees of freedom discussed in Section 4.4.2. In particular, the following Birkhoff's representation theorem is important (see (Fujishige 2005,Section 3.2)).

Theorem 2.6 (Birkhoff's representation theorem)**.** Let $\mathcal{D} \subseteq 2^V$ be a distributive lattice with $\emptyset, V \in \mathcal{D}$. Then, there exists a pair $(\Pi(\mathcal{D}), \preceq_{\mathcal{D}})$ that consists of a partition $\Pi(\mathcal{D})$ of V and a partial order $\preceq_{\mathcal{D}}$ defined on $\Pi(\mathcal{D})$ satisfying the following condition: for any $A \in \mathcal{D}$, there exists an ideal $J \subseteq \Pi(\mathcal{D})$ with respect to $\preceq_{\mathcal{D}}$ such that $A = \bigcup\{S : S \; inJ\}$. Conversely, for any ideal J of $\preceq_{\mathcal{D}}$, the preceding set is an element of $\mathcal{D}$. Here an ideal J of a partial order $\preceq_{\mathcal{D}}$ is a subset of $\Pi(\mathcal{D})$ such that $S \in J, S' \preceq_{\mathcal{D}} S \Rightarrow S' \in J$.

An increasing sequence in $\mathcal{D}$ is called a maximal chain if its length is maximal. Partition $\Pi(\mathcal{D})$ in the above theorem can be constructed from any maximal chain as follows.

Corollary 2.7 ((Fujishige 2005), Corollary 3.10)**.** Let $\mathcal{D} \subseteq 2^V$ be a distributive lattice with $\emptyset, V \in \mathcal{D}$. Let $\emptyset = A_0 \subset A_1 \subset \cdots \subset A_k = V$ be an arbitrary maximal chain in $\mathcal{D}$. Then, the partition in Theorem 2.6 is given as $\Pi(\mathcal{D}) = \{A_i - A_{i-1} : i \in \{1, \ldots, k\}\}$.

2.2.2 Lovász extension

The **Lovász extension** is a natural convex relaxation of a submodular function. For general introduction to this concept, we refer readers to Chapter 3 of Bach (2013) and Section 6.3 of Fujishige (2005). Formally, the Lovász extension of a submodular function $f : 2^V \to \mathbb{R}$ is a function $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ defined as follows: for any $\boldsymbol{\theta} \in \mathbb{R}^p$, let $\tau : V \to V$ be any permutation such that $\theta_{\tau(1)} \geq \cdots \geq \theta_{\tau(p)}$. Define an increasing sequence of sets $A_0 = \emptyset$ and $A_i = \{\tau(1), \ldots, \tau(i)\}$, $i \in \{1, \ldots, p\}$. Then, the value of the Lovász extension $\hat{f}(\theta)$ is defined as

$$\hat{f}(\boldsymbol{\theta}) := \sum_{i=1}^{p} \boldsymbol{\theta}_{\tau(i)}(f(A_i) - f(A_{i-1})).$$

An important fact is that the Lovász extension is the support function of the base polyhedron:

$$\hat{f}(\boldsymbol{\theta}) = \max_{\boldsymbol{z} \in B(f)} \boldsymbol{z}^\top \boldsymbol{\theta}. \tag{2.5}$$

From this expression, it is clear that $\hat{f}$ is convex, piecewise-linear, and positively homogeneous.

For any vector $\boldsymbol{\theta} \in \mathbb{R}^p$, we write $|\boldsymbol{\theta}| = (|\theta_1|, \ldots, |\theta_p|)^\top$. If f is non-decreasing, we have the following:

$$\hat{f}(|\boldsymbol{\theta}|) = \max_{\boldsymbol{z} \in B(f)} \boldsymbol{z}^\top |\boldsymbol{\theta}| = \max_{\boldsymbol{z} \in P_+(f)} \boldsymbol{z}^\top |\boldsymbol{\theta}| = \max_{\boldsymbol{z} \in |P|(f)} \boldsymbol{z}^\top \boldsymbol{\theta}.$$

Hence, the mapping $\boldsymbol{\theta} \mapsto \hat{f}(|\boldsymbol{\theta}|)$ is the support function of the symmetric submodular polyhedron $|P|(f)$. Bach (2010) showed that if f is non-decreasing and strictly positive for all singletons, this mapping defines a norm in $\mathbb{R}^p$.

2.3 Basic probability facts

In this section, we provide some useful results in probability theory.

2.3.1 Stein's lemma

We introduce Stein's lemma (Stein 1981), which plays a central role in Chapter 4. A function $h : \mathbb{R}^n \to \mathbb{R}$ is said to be weakly differentiable if there exist n locally integrable functions $g_1, \dots, g_n$ such that, for any $i \in \{1, \dots, n\}$,

$$\int_{\mathbb{R}^n} h(\boldsymbol{x}) \frac{\partial \varphi}{\partial x_i} \,\mathrm{d}\boldsymbol{x} = -\int_{\mathbb{R}^n} g_i(\boldsymbol{x}) \varphi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$$

for any $\varphi \in C^\infty(\mathbb{R}^n)$. If h is continuously differentiable, this equality usually holds with $g_i = \frac{\partial h}{\partial x_i}$. Hence, for notation convenience, we write $g_i = \frac{\partial h}{\partial x_i}$ whenever h is weakly differentiable. For a function $f : \mathbb{R}^n \to \mathbb{R}^n$ such that the coordinates $f_i : \mathbb{R}^n \to \mathbb{R}$ ($i \in \{1, \dots, n\}$) are weakly differentiable, we define the divergence of f as

$$(\nabla \cdot f)(\boldsymbol{x}) = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(\boldsymbol{x}).$$

Lemma 2.8 (Stein's lemma)**.** Let $\boldsymbol{z}$ be a random variable drawn from n-dimensional isotropic normal distribution $N(\boldsymbol{0}, \boldsymbol{I}_n)$. For any weakly differentiable function $f : \mathbb{R}^n \to \mathbb{R}^n$, the following equality holds:

$$\mathbb{E}[\boldsymbol{z}^\top f(\boldsymbol{z})] = \mathbb{E}\left[(\nabla \cdot f)(\boldsymbol{z})\right].$$

In Chapter 4, Stein's lemma is used for deriving unbiased estimators of the mean squared errors. We comment on some basic usage of Stein's lemma for such purposes. For any $n \times n$ matrix $\boldsymbol{A}$, we have $\mathbb{E}[\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z}] = \mathrm{tr}(\boldsymbol{A})$. Let C be a closed convex polyhedron in $\mathbb{R}^n$, and let Proj_C be the orthogonal projection map onto C. Then, Proj_C is not differentiable, but continuous and locally affine, and thus weakly differentiable. In particular, there is a partition $S_1, \dots, S_M$ of $\mathbb{R}^n$ such that, for each $j \in \{1, \dots, M\}$,

(i) S_j is a closed convex polyhedron, and
(ii) the restriction of Proj_C onto S_j is an affine map and written as $\mathrm{Proj}_C(\boldsymbol{z}) = \boldsymbol{A}_j \boldsymbol{z} + \boldsymbol{b}_j$.

Then, by Stein's lemma, we have

$$\mathbb{E}[\boldsymbol{z}^\top \mathrm{Proj}_C(\boldsymbol{z})] = \sum_{j=1}^M \mathrm{tr}(\boldsymbol{A}_j) \Pr(\boldsymbol{z} \in S_j).$$

In particular, $\boldsymbol{z} \mapsto \sum_j \mathrm{tr}(\boldsymbol{A}_j) 1_{\{\boldsymbol{z} \in S_j\}}$ evidently provides an unbiased estimator of the quantity $\mathbb{E}[\boldsymbol{z}^\top \mathrm{Proj}_C(\boldsymbol{z})]$.

2.3.2 Some probability inequalities

Here, we present several auxiliary probability inequalities used in Chapter 5.

Lemma 2.9 (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 2.10 (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

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

Chapter 3

Statistical Learning with Sparsity

In this chapter, we review the theory of statistical estimation under sparsity. The goal of this chapter is to provide a minimum background required for our main contributions in Chapter 4 and a part of Chapter 5.

3.1 Fixed design regression models

Throughout the main part of this thesis, Chapter 4 and Chapter 5, we stick to the statistical analysis of the **prediction error in fixed design regression models**. Here, we provide a precise definition of the fixed design regression as well as several examples covered by this setting.

Suppose that $y_1, y_2, \ldots, y_n$ be n independent observations drawn according to the following model:

$$y_i = f^*(\boldsymbol{x}_i) + \xi_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are *known* vectors in $\mathbb{R}^p$ and $\xi_1, \xi_2, \ldots, \xi_n$ are mutually independent random variables with $\mathbb{E}[\xi_i] = 0$ and $\mathrm{Var}[\xi_i] \leq \sigma^2$. The task is to estimate the unknown function $f^* : \mathbb{R}^p \to \mathbb{R}$ under the **prediction error** (or the **mean squared error**) defined as follows

$$R(f, f^*) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{f^*}[(f(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i))^2],$$

where the expectation $\mathbb{E}_{f^*}$ is taken with respect to the model (3.1).

Remark 3.1 (Random design regression: What this thesis is *not* for)**.** The terminology "prediction error" is commonly used in the study of high-dimensional statistics (see e.g., van de Geer (2015)). However, this definition of the prediction error is somewhat different from the prediction performance that is usually intended in the statistic and machine learning literature. In these literature, *prediction* generally means a statistical problem to forecast the behavior of future observations. For this purpose, it is common to assume that the training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ and the future observation $(\boldsymbol{x}_{\mathrm{new}}, y_{\mathrm{new}})$ are generated from the same distribution P, and a typical objective is defined as

$$R_{\mathrm{random}}(f, f^*) := \|f - f^*\|_{L^2(P_X)}^2 = \int (f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 \, \mathrm{d}P_X(\boldsymbol{x}).$$

In our setting, the future observations are allowed to occur at the same design points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ as in the given data (3.1). Hence, this setting is often referred to as **fixed design regression**. On the other hand, the situation where the future observation occurs at a new design point $\boldsymbol{x}_{\mathrm{new}} \in \mathbb{R}^p$ is referred to as the **random design regression**. The theoretical results and applications in the fixed and the random design settings are closely related to each other, but may have independent interest.

3.2 Linear regression

Linear regression corresponds the case where f^* is a linear function, i.e., there exists a regression coefficient $\boldsymbol{\theta}^* \in \mathbb{R}^p$ such that $f^*(\boldsymbol{x}) = \langle \boldsymbol{\theta}^*, \boldsymbol{x} \rangle$. In this case, we can rewrite (3.1) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\xi},$$

where we write $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$. For any estimator $\hat{\boldsymbol{\theta}}$, the prediction error is written as

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}^*}\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2.$$

If the design matrix rank$\boldsymbol{X} = p$ (i.e., full column rank), the ordinary least squares (OLS) estimator that minimizes

$$R_n(\boldsymbol{\theta}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$$

is uniquely determined as $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. Here, the assumption that rank$\boldsymbol{X} = p$ corresponds to the classical low-dimensional setting with $n \leq p$. For high-dimensional settings $p > n$, this is not the case because $\boldsymbol{X}^\top\boldsymbol{X}$ is not invertible. In Section 3.3, we will discuss more about the high-dimensional linear regression.

Below, we review some basic results for the OLS estimators and variable selection problems in low-dimensional linear regression. First, we can easily see that the prediction error of the OLS estimator is bounded by $\mathrm{O}(\frac{p}{n})$.

Proposition 3.2. Suppose rank$\boldsymbol{X} = p$. Suppose also that the noise variables ξ_i, $i \in \{1, \ldots, n\}$ are uncorrelated and $\mathrm{Var}[\xi_i] \leq \sigma^2$. Then, for any $\boldsymbol{\theta}^* \in \mathbb{R}^p$, the prediction error of the OLS estimator is bounded as $R(\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}, \boldsymbol{\theta}^*) \leq \frac{\sigma^2 p}{n}$.

Suppose that $\boldsymbol{\theta}^*$ is k-sparse, i.e., the number of non-zero coordinates in $\theta_1^*, \ldots, \theta_p^*$ is not larger than k. If we know the true sparsity pattern $A^* = \{i \in [p] : \theta_i^* \neq 0\}$, Proposition 3.2 suggests that the OLS estimator for restricted matrix $\boldsymbol{X}_{A^*} := [\boldsymbol{X}_i]_{i \in A^*}$ has the prediction error of order $\mathrm{O}(\frac{k}{n})$. When $k \ll p$, this can be much smaller than that of the full model, and thus motivates data-dependent variable selection.

Mallows (1973) proposed the following criterion for selecting variables for the OLS estimator:

$$C_p(A) := \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\theta}}_A\|_2^2 + \frac{2\sigma^2 p}{n}.$$

Here, for any subset $A \subseteq [p]$, $\hat{\boldsymbol{\theta}}_A$ is defined as the OLS estimator with respect to selected variables $\boldsymbol{X}_A = [\boldsymbol{X}_i]_{i \in A}$. The C_p criterion can be seen as minimizing an unbiased estimator of the prediction risk. The following fact is well-known in the literature (see e.g., Stein (1981) and Efron (2004)).

Proposition 3.3. Suppose that ξ_i, $i \in \{1, \ldots, n\}$ are independently drawn from Gaussian distribution $N(0, \sigma^2)$. Then,

$$\hat{R}(\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}\|_2^2 + \frac{2\sigma^2 p}{n} - \sigma^2$$

gives an unbiased estimator for the prediction error of $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$.

The next question is how close the minimizer of Mallows' C_p criterion from the best one. In some situations, the optimality of Mallows' C_p criterion has been proved. Shibata (1981) and Li (1987) showed that the following asymptotic optimality:

$$\frac{\|\boldsymbol{X}^{(n)}(\hat{\boldsymbol{\theta}}_{\hat{A}_n} - \boldsymbol{\theta}^*_n)\|_2^2}{\inf_{A \in \mathcal{A}_n} \|\boldsymbol{X}^{(n)}(\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}^*_n)\|_2^2} \overset{n\to\infty}{\longrightarrow} 1 \quad \text{in probability.}$$

Here, for $n \in \{1, 2, \ldots\}$, $\boldsymbol{X}^{(n)}$ is a deterministic sequence of $n \times p_n$ matrices, $\mathcal{A}_n$ is a collection of subsets of $[p_n]$, and $\boldsymbol{\theta}^*_n \in \mathbb{R}^{p_n}$ is the sequence of true coefficients. Under some regularity assumptions[*1], the above result says that the prediction loss of the selected model $\hat{A}_n$ asymptotically optimal among model candidates $\mathcal{A}_n$. Also, Baraud (2000) showed a non-asymptotic oracle inequality.

Mallows' C_p can be regarded as a special case of Stein's Unbiased Risk Estimate (SURE) (Efron 2004), which can be defined for a more general class of estimators. SURE-tuned estimators have been shown to be optimal in some particular situations (see e.g. Donoho and Johnstone (1995)), but providing a unified theoretical guarantee in general settings remains as an open question (Tibshirani and Rosset 2019). In Chapter 4, we will derive SUREs for a wide class of regularization based estimators.

3.3 Sparse linear regression

In the previous section, we reviewed the variable selection problem when the OLS estimators exist. In this section, we review methods for linear regression with possibly high-dimensional design matrices. For general introduction to high-dimensional statistics, see Bühlmann and van de Geer (2011), Giraud (2015), Hastie et al. (2015), and van de Geer (2015).

In modern high-dimensional statistics, we have to take care of statistical performance as well as **computational efficiency**. Below, we review some typical estimators for sparse linear regression and discuss their statistical/computational efficiency.

Let $\boldsymbol{X}$ be $n \times p$ design matrix with $p \geq n$. Suppose that $\boldsymbol{\theta}^*$ is k-sparse with $k < n$, but we do not know the true position of non-zero coordinates of $\boldsymbol{\theta}^*$. Let us define the **best k-subset estimator $\hat{\boldsymbol{\theta}}_{\mathrm{BS}}$** as any solution of the following constrained problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_0 \leq k.$$

We can check that, under suitable assumptions, the best subset selection estimator satisfies

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}^*}\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_{\mathrm{BS}} - \boldsymbol{\theta}^*)\|_2^2 \leq C\frac{\sigma^2 k \log(\mathrm{e}p/k)}{n}$$

for some universal constant $C > 0$. The rate $\mathrm{O}(\frac{k\log(\mathrm{e}p/k)}{n})$ often appears in the minimax optimal rate for k-sparse regression (Raskutti et al. 2011). However, a statistical drawback of the best subset selection is that it is by definition not adaptive to the sparsity level k, which is often unknown in practice. Moreover, a practically more serious issue is that computing the best subset estimator can involve intractable combinatorial optimization due to the cardinality constraint $\|\boldsymbol{\theta}\|_0 \leq k$. In fact, this problem is shown to be NP-hard (Natarajan 1995), and even the state-of-the-art solvers can handle problems with sizes up to $n \approx 100$ and $p \approx 2000$ (Bertsimas et al. 2016).

[*1] The regularity assumption may contain conditions on the designs $(\boldsymbol{X}^{(n)}, \boldsymbol{\theta}^*_n, \mathcal{A}_n)$ and the moment of noise variables. See Li (1987) as well as Baraud (2000) for details.

For the statistical issue, we can alternatively consider the ℓ_0-regularization estimator as follows:

$$\hat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_0 \right),$$

where $\lambda > 0$ is a regularization parameter. This formulation contains Mallows' C_p type regularization ($\lambda = \sigma^2$) and BIC type regularization ($\lambda = \sigma^2 \log n$). The result in Birgé and Massart (2001) suggests that, if the noises are Gaussian, the above ℓ_0-regularization estimator with $\lambda = 2\sigma^2(1 + \sqrt{2\log p})^2$ satisfies

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}^*} \|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\|_2^2 \le C \frac{\sigma^2(1 + k\log(p))}{n},$$

where $C > 0$ is a universal constant (see also Theorem 2.2 of Giraud (2015) for a more precise statement). Therefore, we can conclude that there exists an estimator that achieves the nearly-minimax rate for any sparsity level k. However, computing ℓ_0-regularization estimators is still a hard optimization problem.

3.3.1 The lasso estimator

For a regularization parameter $\lambda > 0$, the well-known **lasso** estimator (Tibshirani 1996) is defined as follows:

$$\hat{\boldsymbol{\theta}}_{\text{Lasso},\lambda} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right).$$

Here, the ℓ_1-regularization term $\lambda\|\boldsymbol{\theta}\|_1$ can be regarded as a convex relaxation of the cardinality penalty $\lambda\|\boldsymbol{\theta}\|_0$. Indeed, it is known that $\boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|_1$ is the tightest convex lower bound of the ℓ_0-norm $\boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|_0$ (Bach 2010). Thus, the optimization problem for the lasso is convex function minimization, and there are several practical algorithms to obtain (approximate) solutions. The coordinate descent (Friedman et al. 2007) is a standard approach for solving the lasso optimization problems, and its implementations are available as famous software libraries such as `glmnet` (Friedman et al. 2010) and `scikit-learn` (Pedregosa et al. 2011). Least angle regression algorithm (LARS, Efron et al. (2004)) calculates the entire solution path as a function of λ. Proximal methods, especially FISTA (Beck and Teboulle 2009), reasonably perform well in large-scale settings (Bach et al. 2012). For a more detailed review of optimization methods for lasso, see Chapter 5 of Hastie et al. (2015).

Regarding the prediction errors, Bickel et al. (2009) showed that the lasso estimator can achieve the rate of $\mathrm{O}(\frac{k\log p}{n})$ under some spectral condition on the design matrix $\boldsymbol{X}$. More precisely, they proved risk bounds that hold under the following **restricted eigenvalue condition**:

Definition 3.4. Let $1 \le k \le p$ be an integer and $\alpha > 0$. We say that an $n \times p$ matrix $\boldsymbol{X}$ satisfies the restricted eigenvalue condition $\mathrm{RE}(k, \alpha)$ if it satisfies

$$\kappa(k,\alpha) := \min_{A\subset[p]:|A|\le k} \min_{\substack{\boldsymbol{\theta}\neq\boldsymbol{0}:\\ \|\boldsymbol{\theta}_{A^c}\|_1 \le \alpha\|\boldsymbol{\theta}_A\|_1}} \frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2}{\sqrt{n}\|\boldsymbol{\theta}_A\|_2} > 0.$$

It is known that the restricted eigenvalue condition holds with high probability for Gaussian random matrices (Raskutti et al. 2010). The following risk bound is adapted from Corollary 4.3 of Giraud (2015).

Proposition 3.5. Let $k = \|\boldsymbol{\theta}^*\|_0$. Assume that $\boldsymbol{X}$ satisfies $\mathrm{RE}(k, 5)$ and all the columns of $\boldsymbol{X}$ have norm 1. Assume also that the noise variables $\xi_1, \ldots, \xi_n$ are independently drawn from $N(0, \sigma^2)$. Then, for any $\delta \in (0, 1)$, we can choose the tuning parameter $\lambda = \lambda_\delta$ such that

$$\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_{\mathrm{Lasso},\lambda} - \boldsymbol{\theta}^*)\|_2^2 \leq \inf_{\boldsymbol{\theta} \neq \mathbf{0}} \left(\frac{1}{n}\|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|_2^2 + \frac{C}{\kappa(k,5)^2} \cdot \frac{\sigma^2 k \log(p/\delta)}{n} \right),$$

holds with probability at least $1 - \delta$. Here, $C > 0$ is a universal constant.

We conclude this subsection with some comments on recent developments in the theory of lasso. First, in the above risk bound, the choice of the tuning parameter $\lambda = \lambda_\delta$ happens to depend on the deviation δ in the risk bound. For the Gaussian noise setting, Bellec et al. (2018) refined this result so that λ can be taken independently from δ. Second, it has been shown that the constant $1/\kappa(k, \alpha)^2$ appeared in the upper bound is unavoidable for any polynomial time methods (Zhang et al. 2014), while the risk bounds for some non-polynomial time methods (e.g., ℓ_0-regularization and aggregation) do not involve this term.

3.3.2 Structured sparsity

Over the last two decades, structured sparsity has been a major research topic in high-dimensional statistics and machine learning (Jenatton et al. 2011, Wainwright 2014, van de Geer 2015, Obozinski and Bach 2016). Given a convex regularizer $\Omega : \mathbb{R}^p \to \mathbb{R}$, we consider the following estimator:

$$\hat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \Omega(\boldsymbol{\theta}) \right). \tag{3.2}$$

Lasso corresponds the case $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, which promotes the "coordinate sparsity" of solutions. Here, coordinate sparsity stands for the property that each individual coordinate $\hat{\theta}_{\lambda,i}$ can be non-zero independently from the other coordinates, and thus possible sparsity patterns are all the subsets of the index set. On the other hand, different choices of the regularizer Ω allow sparsity patterns to be more structured. To name a few, regularization methods for structured sparsity include the group lasso (Yuan and Lin 2006) and its generalizations (Jacob et al. 2009, Jenatton et al. 2011, Obozinski and Bach 2016), the fused lasso (Rudin et al. 1992, Tibshirani et al. 2005), and methods based on directed graphs (Tibshirani et al. 2011) and hypergraphs (Hein et al. 2013, Takeuchi et al. 2015). For comprehensive lists of existing regularization methods for structured sparsity, see e.g., Wainwright (2014), Obozinski and Bach (2016) and Hastie et al. (2015).

While there are many statistical analyses for specific choices of regularizers, some researchers study theoretical risk bounds for regularization methods in the general form (3.2). Notably, Negahban et al. (2012) developed consistency results for general M-estimators of the form (3.2) by introducing the notion of decomposability and restricted strong convexity. van de Geer (2014) developed risk bounds based on more generalized conditions, i.e., weak decomposability and compatibility. Here, weak decomposability is a regularity condition for Ω that generalizes the property of ℓ_1-norm that $\|\boldsymbol{\theta}\|_1 = \|\boldsymbol{\theta}_A\|_1 + \|\boldsymbol{\theta}_{A^c}\|_1$, which determines allowed sparsity patterns for the structured sparsity regularizer Ω. On the other hand, compatibility is a condition on the interaction between the regularizer Ω and the loss $\boldsymbol{\theta} \mapsto \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$, which generalizes the restricted eigenvalue condition. Since providing a precise statement of the risk bound is slightly out of scope of this thesis, we refer interested readers to Chapter 6 of van de Geer (2015) for a good introduction to this topic.

Chapter 4

Degrees of Freedom in Submodular Regularization

Degrees of freedom is a covariance penalty related to penalized model selection procedures such as Mallows' C_p and AIC. We study the degrees of freedom of two polyhedral convex regularization classes defined through submodular functions called the Lovász extension regularization and submodular norm regularization. It has been pointed out that submodular regularization contains many existing penalties that induce structured sparsity. In this chapter, we show that the degrees of freedom of submodular regularization estimators can be represented in terms of partitions induced by the estimators. Our formula does not depend on the choice of the design matrix and the penalty function. Moreover, if the design matrix has full column rank, calculating an unbiased estimator of the degrees of freedom requires an additional computational cost of only $\mathrm{O}(p \log p)$ after a solution for the estimator is obtained, where p is the dimension of the parameter. Existing results for some regularization and projection type estimators, such as the lasso, the fused lasso, and the isotonic regression, are also recovered.

This chapter is based on Minami (2020).

4.1 Overview

Consider the following linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}, \tag{4.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the target variable, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is an arbitrary design matrix, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is the regression coefficient, and $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ is a Gaussian noise vector. Throughout this chapter, we consider the following regularization based estimator:

$$\hat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \Omega(\boldsymbol{\theta}) \right). \tag{4.2}$$

Here, $\Omega : \mathbb{R}^p \to \mathbb{R}$ is a convex regularizer that imposes some low-dimensional structure on the estimated parameter $\hat{\boldsymbol{\theta}}_\lambda$. An important example of low-dimensional structure is sparsity. Suppose that we have prior knowledge that the true parameter is sparse (i.e., having only a few non-zero elements). Then, a typical choice of Ω is the ℓ_1-norm $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, with which the estimator (4.2) admits a sparse solution (Tibshirani 1996). To introduce further combinatorial structures on the selected variables, structured sparsity has received notable attention over the last decade (see Bach (2010, 2013), Obozinski and Bach (2016), and Chapter 6 in van de Geer (2015), for example). It has been pointed out that a broad class of regularizers inducing structured sparsity can be obtained as convex relaxations of combinatorial penalty functions (Bach 2010, 2013, Obozinski and Bach 2016). We focus on the following two classes of structured regularization estimators defined through submodular functions:

1. (Lovász extension regularization (LER)) Let $f : 2^V \to \mathbb{R}$ be a submodular function and $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ be its Lovász extension (defined in Section 4.3). The *Lovász extension regularization estimator* (LERE) is defined as

$$\hat{\boldsymbol{\theta}}_\lambda = \hat{\boldsymbol{\theta}}_{\mathrm{LERE},\lambda} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \hat{f}(\boldsymbol{\theta}) \right), \tag{4.3}$$

where $\lambda > 0$ is a regularization parameter. An important and motivating example of LERE is the generalized fused lasso (Tibshirani et al. 2005). Given an undirected graph $G = (V, E)$, the generalized fused lasso regularizer is defined as $\hat{f}(\boldsymbol{\theta}) = \sum_{(i,j)\in E} |\theta_i - \theta_j|$. Other examples of LERE will be introduced in Section 4.5.
2. (Submodular norm regularization (SNR)) Let f be a monotone submodular function. The *submodular norm regularization estimator* (SNRE) is defined as

$$\hat{\boldsymbol{\theta}}_\lambda = \hat{\boldsymbol{\theta}}_{\mathrm{SNRE},\lambda} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \hat{f}(|\boldsymbol{\theta}|) \right), \tag{4.4}$$

where, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$, we write $|\boldsymbol{\theta}| = (|\theta_1|, \ldots, |\theta_p|)^\top$. As suggested by the name, the regularizer in SNR $\boldsymbol{\theta} \mapsto \hat{f}(\boldsymbol{\theta})$ forms a norm, and a wide class of sparsity-inducing norms including the lasso, SLOPE (Bogdan et al. 2015), the group lasso (Yuan and Lin 2006) and its overlapping and hierarchical extensions (Jacob et al. 2009, Jenatton et al. 2011, Obozinski and Bach 2016) can be represented by SNRE. We will discuss further examples of SNRE in Section 4.5.

A practically important issue is to determine the tuning parameter $\lambda \geq 0$ using the observed data $\boldsymbol{y}$. To be precise, let us consider that a "good" tuning parameter λ minimizes the risk $R(\hat{\boldsymbol{\theta}}_\lambda) := \mathbb{E}_{\boldsymbol{\theta}_0} \|\boldsymbol{X}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_\lambda)\|_2^2$, where $\mathbb{E}_{\boldsymbol{\theta}_0}$ is the expectation with respect to the model (4.1). Many existing methods to tackle this problem are based on minimizing some estimator of $R(\hat{\boldsymbol{\theta}}_\lambda)$. The most popular model selection methods certainly fall into two categories: the cross-validation methods (see Yang (2007) and references therein) and methods based on information criteria. Since the cross-validation methods contain a data splitting procedure, it can be applied for the setting in which the matrix $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)^\top$ is interpreted as n observations of p dimensional explanatory variables. On the other hand, information criteria, including AIC (Akaike 1973), Mallows' C_p (Mallows 1973), and BIC (Schwarz 1978), use theoretically derived estimators of the risk R or some other statistical objectives. In the fields such as image processing and graph signal processing, $\boldsymbol{y}$ is interpreted as a noisy observation of a single structured signal $\boldsymbol{\theta}_0$ (i.e., $\boldsymbol{X} = \boldsymbol{I}_n$). For this case, information criteria are more suitable because the data cannot be split into validation samples in natural ways. Therefore, deriving an information criterion for submodular regularization estimators is important.

In this chapter, we focus on a certain unbiased estimate of the risk, which is known as Stein's Unbiased Risk Estimate (SURE). Suppose that $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\xi}$ is a Gaussian observation with the unknown mean parameter $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. For any estimator $\hat{\boldsymbol{\mu}}$, the risk is decomposed into the bias-variance form as

$$R(\hat{\boldsymbol{\mu}}) = \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 = \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{y}\|_2^2 + 2 \sum_{i=1}^n \mathrm{Cov}(\hat{\mu}_i, y_i) - n\sigma^2,$$

where $\mathrm{Cov}(u, v) = \mathbb{E}_{\boldsymbol{\mu}}[(u - \mathbb{E}_{\boldsymbol{\mu}}[u])(v - \mathbb{E}_{\boldsymbol{\mu}}[v])]$. The first term on the right-hand side can be regarded as the bias term, which is the expectation of the empirical goodness-of-fit to data $\boldsymbol{y}$. On the other hand, the second and the third terms corresponds to the variance term, which is often interpreted as the model complexity or the effective number of parameters. In the spirit of Efron (2004), the degrees of freedom of $\hat{\boldsymbol{\mu}}$ is defined as

$$\mathrm{df}(\hat{\boldsymbol{\mu}}) := \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{\mu}_i, y_i). \tag{4.5}$$

Suppose that $\boldsymbol{y} \mapsto \hat{\boldsymbol{\mu}}(\boldsymbol{y})$ is continuous and weakly differentiable. By Stein's lemma (Stein 1981), the degrees of freedom can be written as

$$\mathrm{df}(\hat{\boldsymbol{\mu}}) = \mathbb{E}_{\boldsymbol{\mu}}[(\nabla \cdot \hat{\boldsymbol{\mu}})(\boldsymbol{y})] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}(\boldsymbol{y})\right], \tag{4.6}$$

where $\nabla \cdot \hat{\boldsymbol{\mu}}$ is the divergence of $\hat{\boldsymbol{\mu}}$ that is defined almost everywhere. Hence, we can use $\hat{\mathrm{df}}(\hat{\boldsymbol{\mu}}) := \nabla \cdot \hat{\boldsymbol{\mu}}$ as an unbiased estimate of the degrees of freedom, and thus SURE defined as

$$\hat{R}_{\mathrm{SURE}}[\hat{\boldsymbol{\mu}}](\boldsymbol{y}) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{y}\|_2^2 + 2\sigma^2 \hat{\mathrm{df}}[\hat{\boldsymbol{\mu}}](\boldsymbol{y}) - n\sigma^2$$

is an unbiased estimate of the risk $R(\hat{\boldsymbol{\mu}})$. With a slight abuse of the terminology, we also refer to the unbiased estimate $\hat{\mathrm{df}}$ as the degrees of freedom.

Our problem is to characterize the degrees of freedom of the submodular regularization estimators described above. To clarify the desired form of the statements, we first review the existing results on the degrees of freedom of typical regularization estimators. For the lasso $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, the degrees of freedom is given by

$$\hat{\mathrm{df}}(\boldsymbol{X}\hat{\boldsymbol{\theta}}_{\mathrm{Lasso},\lambda}) = (\text{Number of non-zero elements in } \hat{\boldsymbol{\theta}}_{\mathrm{Lasso},\lambda}). \tag{4.7}$$

This result was first proved for a full-rank design matrix $\boldsymbol{X}$ by Zou et al. (2007) and justified for general low-rank matrices by Tibshirani and Taylor (2012). Since the solution of the lasso is known to become automatically sparse, the above result (4.7) seems natural in the sense that the degrees of freedom provides the effective number of parameters. Another example is the fused lasso (Tibshirani et al. 2005) on a given undirected graph $G = (V, E)$. It is known that the solution $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{Fused},\lambda}$ of the fused lasso becomes "piecewise constant" on connected components in G. We refer to a connected component $A \subseteq V$ on which $\hat{\boldsymbol{\theta}}$ is constant as a fused group. In Tibshirani and Taylor (2011, 2012), the authors proved that the degrees of freedom of the fused lasso is given as

$$\hat{\mathrm{df}}(\boldsymbol{X}\hat{\boldsymbol{\theta}}_{\mathrm{Fused},\lambda}) = (\text{Number of fused groups in } \hat{\boldsymbol{\theta}}_{\mathrm{Fused},\lambda}). \tag{4.8}$$

As mentioned before, the above two example are instances of SNRs and LERs, respectively. Therefore, the question is whether representations similar to (4.7) or (4.8) hold for general submodular regression estimators.

4.1.1 Contribution

Here, we describe our main theorems for the degrees of freedom of submodular regularization estimators. For the sake of notation simplicity, we provide the statements that hold only for a full-rank design matrix $\boldsymbol{X}$. The complete statements that hold for a general matrix $\boldsymbol{X}$ will be provided in Section 4.4.

Theorem 4.1 (Degrees of freedom of full-rank LERE)**.** Let $f : 2^V \to \mathbb{R}$ be any submodular function and $\lambda \geq 0$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be any design matrix with $\mathrm{rank}(\boldsymbol{X}) = p$. Then, the degrees of freedom of the LERE (4.3) is given as

$$\hat{\mathrm{df}}(\hat{\boldsymbol{\theta}}) = (\text{Number of unique values in } \hat{\theta}_1, \ldots, \hat{\theta}_p). \tag{4.9}$$

Theorem 4.2 (Degrees of freedom of full-rank SNRE)**.** Let $f : 2^V \to \mathbb{R}$ be any monotone submodular function, and $\lambda \geq 0$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be any design matrix with $\mathrm{rank}(\boldsymbol{X}) = p$. Then, the degrees of freedom of the SNRE (4.4) is given as

$$\hat{\mathrm{df}}(\hat{\boldsymbol{\theta}}) = (\text{Number of non-zero unique values in } |\hat{\theta}_1|, \ldots, |\hat{\theta}_p|). \tag{4.10}$$

The existing results on the lasso and the fused lasso can be recovered from the above theorems. For the lasso, the expression (4.10) is equivalent to (4.7) almost surely because the event that $\hat{\theta}_i = \hat{\theta}_j$ for some $i \neq j$ occurs with probability 0. For the fused lasso, the expression (4.9) actually coincides with (4.8) almost surely. We provide a detailed discussion on this equivalence in Section 4.5.

We can interpret Theorem 4.1 and Theorem 4.2 from an algorithmic perspective. Once we have a solution $\hat{\boldsymbol{\theta}}_\lambda$ of LERE (4.3), the degrees of freedom is calculated by sorting the elements of $\hat{\theta}_\lambda$ and counting "jumps" in the sorted values. Similary, the degrees of freedom of SNRE (4.4) is calculated by sorting the elements of $|\hat{\boldsymbol{\theta}}_\lambda|$ and counting "jumps" between non-zero values. In both cases, the computational complexity of calculating the degrees of freedom is $\mathrm{O}(p \log p)$. We should note that it is substantially faster compared to the case of the general polyhedral convex regularization method. The Lovász extension of a submodular function on $[p]$ is defined as the support function of a polytope that is defined through $M = \mathrm{O}(2^p)$ linear inequality constraints. As we discuss in Section 4.4, a naïve implementation of calculating the degrees of freedom involves an enumeration of satisfied linear equalities that requires $\mathrm{O}(M)$ runtime. Hence, we can say that the submodularity helps to reduce the complexity of the degrees of freedom calculation.

4.1.2 Organization

The remainder of the present chapter is organized as follows. In Section 4.2, we review related work on the degrees of freedom. In Section 4.3, we present notation and definitions that are needed to describe our main results. Section 4.4 is the main part of this chapter. In Section 4.4.1, we explain the relationship between submodular regularization estimators and a certain class of (anti-)projection estimators. In Section 4.4.2, we state our main theorem on the degrees of freedom of submodular regularization estimators. In particular, Theorem 4.1 and Theorem 4.2 will be obtained as corollaries of Theorem 4.7 and 4.11, respectively. In Section 4.5, we provide specific examples of submodular regularization estimators and discuss the relationship between the existing results on the degrees of freedom and our results. We conduct some numerical simulations in Section 4.6. Finally, we provide some additional discussion in Section 4.7.

4.2 Related work

Explicit formulae for the degrees of freedom have been developed for various models. A parallel approach to the convex regularization estimator (4.2) is the projection estimator onto a convex set C:

$$\hat{\boldsymbol{\theta}}_C(\boldsymbol{y}) = \mathrm{Proj}_C(\boldsymbol{y}) := \operatorname*{argmin}_{\boldsymbol{z} \in C} \|\boldsymbol{y} - \boldsymbol{z}\|_2^2.$$

The degrees of freedom of a projection estimator is given as the dimension of the face associated with the projected point $\mathrm{Proj}_C(\boldsymbol{y})$. This characterization was proved essentially in Meyer and Woodroofe (2000), who studies the 1-dimensional isotonic regression. Kato (2009) extended the result to general convex sets with smooth boundaries. For polyhedral convex sets, Chen et al. (2019) pointed out that it suffices to consider the rank of the sub-matrix defining the face. The regularization type estimator (4.2) can often be regarded as an anti-projection estimator $\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{y} - \mathrm{Proj}_C(\boldsymbol{y})$. Hence, the degrees of freedom is given by the codimension of a face of the corresponding convex set (Tibshirani and Taylor 2012,Section 2.2). A specific example of the results can be found for the lasso (Zou et al. 2007, Tibshirani and Taylor 2012), the generalized lasso (Tibshirani and Taylor 2011, 2012), and a directed graph regularization estimator (Tibshirani et al. 2011). It is important to note that calculating the degrees of freedom directly from the above face (co)dimension characterization often involves the enumeration of linear equality constraints, which can be computationally expensive in general.

If the design matrix $\boldsymbol{X}$ is not an identity matrix, it would be beneficial to obtain a formula for calculating the degrees of freedom in terms of the value of the estimator $\hat{\boldsymbol{\theta}}$ itself. However, if the rank of $\boldsymbol{X}$ is less than p, the solution of (4.2) is not unique, and the correspondence $\boldsymbol{y} \mapsto \hat{\boldsymbol{\theta}}$ cannot be regarded as a differential map. For example, the number of non-zero coordinates of the lasso estimator (4.7) does not necessarily coincide with the face codimension unless $\mathrm{rank}(\boldsymbol{X}) = p$. The seminal paper of Tibshirani and Taylor (2012) established a characterization of the degrees of freedom via the concept of the active set, which allows to prove that the expression (4.7) holds almost everywhere for any design matrix.

Besides the modeling flexibility, submodular regularization has another advantage that the estimators can be calculated using computationally efficient convex optimization procedures. In particular, if $\Omega(\theta)$ is either the Lovász extension $\hat{f}(\theta)$ or the submodular norm $\hat{f}(|\theta|)$, calculating the proximal operator $\mathrm{prox}_{\lambda\Omega}(\theta) := \mathrm{argmin}_{z\in\mathbb{R}^p} \left\{\frac{1}{2}\|z-\theta\|_2^2 + \lambda\Omega(\theta)\right\}$ is reduced to solving the minimum norm point problem over an associated polytope, which is equivalent to (parametrized) submodular function minimization (Bach 2013,Chapter 9). Generally, submodular function minimization can be solved in strongly polynomial time (Schrijver 2000, Iwata et al. 2001). Moreover, if the submodular function has a certain graph-representable property (Jegelka et al. 2011), the minimum norm point problem can be solved by practically fast network flow algorithms (Mairal et al. 2011b, Mairal and Yu 2013, Takeuchi et al. 2015). If the proximal operators are obtained, the minimization problems in (4.3) and (4.4) can be solved by the accelerated proximal gradient algorithms (a.k.a. FISTA; Beck and Teboulle (2009)). It is reported in Bach et al. (2012) that the proximal gradient algorithms have good empirical performance especially in large-scale instances.

4.3 Preliminary

Let $\boldsymbol{X}$ be an arbitrary $n\times p$ matrix. $\mathrm{col}(\boldsymbol{X})$, $\mathrm{row}(\boldsymbol{X})$, and $\mathrm{null}(\boldsymbol{X})$ denote column, row, and null spaces of a matrix $\boldsymbol{X}$, respectively. $\boldsymbol{X}^+$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{X}$. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$, we use $\boldsymbol{u} \odot \boldsymbol{v}$ to denote the element-wise product $\boldsymbol{u} \odot \boldsymbol{v} = (u_1v_1, \dots, u_pv_p)^\top \in \mathbb{R}^p$. For a number $\lambda \in \mathbb{R}$, a vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ and a set $C \subseteq \mathbb{R}^p$, we define $\lambda C = \{\lambda \boldsymbol{x} : \boldsymbol{x} \in C\}$ and $\boldsymbol{\gamma} \odot C = \{\boldsymbol{\gamma} \odot \boldsymbol{x} : \boldsymbol{x} \in C\}$.

Let $V = [p] := \{1, \dots, p\}$ be a finite set. For any subset $A \subseteq V$, $\mathbf{1}_A \in \mathbb{R}^p$ is the characteristic vector of A. For a vector $\boldsymbol{x} \in \mathbb{R}^p$ and a set $A = \{i_1, \dots, i_k\} \subseteq V$, we use $\boldsymbol{x}_A$ to denote a subvector $\boldsymbol{x}_A = (x_{i_1}, \dots, x_{i_k})^\top \in \mathbb{R}^A$.

Let $C \subseteq \mathbb{R}^d$ be a closed convex set. We use $\mathrm{Proj}_C : \mathbb{R}^d \to C$ to denote the orthogonal projection map $\mathrm{Proj}_C(x) \in \mathrm{argmin}_{\boldsymbol{z}\in C} \|\boldsymbol{x}-\boldsymbol{z}\|_2^2$. If $C = L$ is a linear space, we also use $\boldsymbol{P}_L$ to denote the orthogonal projection matrix. A normal vector of C at $\boldsymbol{x} \in C$ is a vector $\boldsymbol{v}$ such that $\boldsymbol{v}^\top(\boldsymbol{x}'-\boldsymbol{x}) \geq 0$ holds for any $\boldsymbol{x}' \in C$. The normal cone $N_C(\boldsymbol{x})$ at $\boldsymbol{x}$ is a set of all normal vectors. We use $\mathrm{relint}(C)$ to denote the relative interior in the sense of convex analysis, i.e., the interior of C with respect to the relative topology induced by the smallest affine subspace containing C.

Let $V = [p]$ be a finite set. A set of subsets $\mathcal{D} \subseteq 2^V$ is a distributive lattice if, for any $A, B \in \mathcal{D}$, $A \cup B$ and $A \cap B$ are contained in $\mathcal{D}$. A partition Π of V is a collection of nonempty subsets $\Pi = \{A_1, \ldots, A_k\}$ that are disjoint and satisfy $\bigcup_i A_i = V$. A partition Π_1 is a refinement of partition Π_2 if, for any element $A \in \Pi_1$, there is an element $B \in \Pi_1$ such that $A \subseteq B$. In this case, Π_2 is called a cover of Π_1.

4.4 Degrees of freedom in submodular regularization

In this section, we present our main theorems on the degrees of freedom of the submodular regularization.

4.4.1 Projection and anti-projection estimators

Our goal is to provide formulae for the degrees of freedom of submodular regularization estimators. Before that, we review some properties of two generalized classes of estimators, i.e., projection estimators and anti-projection estimators.

Let $C \subseteq \mathbb{R}^n$ be a closed convex polyhedron. We define the projection estimator $\hat{\boldsymbol{\mu}}^{\mathrm{P}}_C(\boldsymbol{y}) := \mathrm{Proj}_C(\boldsymbol{y})$ and the anti-projection estimator $\hat{\boldsymbol{\mu}}^{\mathrm{A}}_C(\boldsymbol{y}) := (\boldsymbol{I}_n - \mathrm{Proj}_C)(\boldsymbol{y}) = \boldsymbol{y} - \mathrm{Proj}_C(\boldsymbol{y})$. These two classes of estimators are closely related to regularization estimators such as LERE (4.3) or SNRE (4.4) in the following sense:

(i) (Lagrange duality). Let $\Omega : \mathbb{R}^p \to \mathbb{R}_+$ be a convex function. For any $t \geq 0$, let $\hat{\boldsymbol{\mu}}^{\mathrm{P}}_{C_t} = \boldsymbol{X}\hat{\boldsymbol{\theta}}^{\mathrm{P}}_{C_t}$ be the projection estimator onto the set $C_t := \{\boldsymbol{\theta} \in \mathbb{R}^p : \Omega(\boldsymbol{\theta}) \leq t\}$. In other words, $\hat{\boldsymbol{\theta}}^{\mathrm{P}}_{C_t}$ is a solution of the following convex optimization problem:

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \quad \text{subject to} \quad \Omega(\boldsymbol{\theta}) \leq t.$$

It is well-known in convex analysis that, under very mild assumptions, there exists the Lagrange multiplier $\lambda \geq 0$ such that $\hat{\boldsymbol{\theta}}^{\mathrm{P}}_{C_t}$ is a solution of regularization type problem (4.2). In particular, the solution paths in the primal and dual formulations are equivalent.

(ii) (Limiting solution as $\lambda \to \infty$). In some cases, projection estimators are obtained as a limit of regularization estimators. Suppose that $\Omega : \mathbb{R}^p \to \mathbb{R}_+$ is a lower semi-continuous convex function with the minimum value of 0. Then, the set of minimizers $C_{\min} = \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \Omega(\boldsymbol{\theta}) = \{\boldsymbol{\theta} \in \mathbb{R}^p : \Omega(\boldsymbol{\theta}) = 0\}$ is a nonempty closed convex set. Taking the limit as $\lambda \to \infty$, the solution $\hat{\boldsymbol{\theta}}_\lambda$ of (4.2) converges to the solution of the constrained problem:

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \quad \text{subject to} \quad \boldsymbol{\theta} \in C_{\min}.$$

As a specific example, we will discuss the isotonic regression estimator in Section 4.5.

(iii) (Support function). The regularization estimator (4.2) has a direct connection to an anti-projection estimator. For any (nonempty) closed convex set $C \subseteq \mathbb{R}^p$, the support function of C is a function $\Omega_C : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ defined as $\Omega_C(\boldsymbol{\theta}) :=$

$\sup_{\boldsymbol{z}\in C} \boldsymbol{z}^\top \boldsymbol{\theta}$ for any $\boldsymbol{\theta} \in \mathbb{R}^p$. If the regularization function $\Omega = \Omega_C$ is a support function of a convex set C, we can see that $\boldsymbol{X}\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{y} - \mathrm{Proj}_C(\boldsymbol{y})$ holds for any choice of the solution $\hat{\theta}_\lambda$. As we have already seen in Section 2.2.2, Lovász extensions and submodular norms are examples of support functions.

For the purpose of deriving the degrees of freedom, the third perspective is important. The following lemma provides explicit representations of submodular regularization estimators as anti-projection estimators.

Lemma 4.3 (Anti-projection representations of LERE and SNRE)**.** Let $f : 2^V \to \mathbb{R}$ be a submodular function, and $\lambda > 0$ be a regularization parameter. Let $\boldsymbol{y}$ be an arbitrary point in $\mathbb{R}^n$.

(i) Let $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}_{\mathrm{LERE}}$ be an arbitrary solution of LERE (4.3). Then, the regression fit $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ is written as $\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{I}_n - \mathrm{Proj}_C)(\boldsymbol{y})$, which does not depend on the choice of $\hat{\boldsymbol{\theta}}$. Here, $C \subset \mathbb{R}^n$ is a polyhedron defined as

$$C := \{\boldsymbol{z} \in \mathbb{R}^n : \lambda^{-1}\boldsymbol{X}^\top \boldsymbol{z} \in B(f)\}. \tag{4.11}$$

Moreover, $\hat{\boldsymbol{\theta}}$ is contained in the normal cone of the base polyhedron $B(f)$:

$$\hat{\boldsymbol{\theta}} \in N_{B(f)}(\lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})). \tag{4.12}$$

(ii) Assume also that f is a non-decreasing function. Let $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}_{\mathrm{SNRE}}$ be an arbitrary solution of SNRE (4.4). Then, we have $\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{I}_n - \mathrm{Proj}_C)(\boldsymbol{y})$ with

$$C := \{\boldsymbol{z} \in \mathbb{R}^n : \lambda^{-1}\boldsymbol{X}^\top \boldsymbol{z} \in |P|(f)\}. \tag{4.13}$$

Moreover, $\hat{\boldsymbol{\theta}}$ is contained in the normal cone of the symmetric submodular polyhedron $|P|(f)$:

$$\hat{\boldsymbol{\theta}} \in N_{|P|(f)}(\lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})). \tag{4.14}$$

Remark 4.4. If the design matrix $\boldsymbol{X}$ has full column rank, the solution $\hat{\boldsymbol{\theta}}$ in optimization problem (4.3) (or (4.4)) is unique. On the other hand, the solution is not determined uniquely for general low-rank matrices that appear in the high-dimensional scenario (i.e., $p > n$). However, Lemma 4.3 implies that the regression fit $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ is always unique.

To derive analytic expressions of the degrees of freedom, we are interested in the divergence map of $\boldsymbol{y} \mapsto \boldsymbol{X}\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{y} - \mathrm{Proj}_C(\boldsymbol{y})$. For any polyhedron C, it is known that the divergence of the projection map $\mathrm{Proj}_C(\boldsymbol{y})$ is given by the codimension of the face of C (Meyer and Woodroofe 2000, Kato 2009, Chen et al. 2019).

Proposition 4.5 (e.g., Proposition 1 of (Meyer and Woodroofe 2000), Theorem 2.3 of (Chen et al. 2019))**.** Let $C \subseteq \mathbb{R}^n$ be a polyhedron. Then, the projection map $\boldsymbol{y} \mapsto \mathrm{Proj}_C(\boldsymbol{y})$ is differentiable almost everywhere, and the divergence is given as $(\nabla \cdot \mathrm{Proj}_C)(\boldsymbol{y}) = \dim(F(\boldsymbol{y}))$, where $F(\boldsymbol{y})$ is the face of C such that $\mathrm{Proj}_C(\boldsymbol{y}) \in \mathrm{relint}(F(\boldsymbol{y}))$. Moreover, if C is represented by a linear inequality system as

$$C = \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{a}_i^\top \boldsymbol{\theta} \le b_i,\ i \in \{1, \ldots, m\}\} \tag{4.15}$$

for some $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$, then we have $\dim(F(\boldsymbol{y})) = n - \mathrm{rank}\boldsymbol{A}_0(\boldsymbol{y})$. Here, $\boldsymbol{A}_0(\boldsymbol{y})$ is a matrix whose columns consist of all vectors $\boldsymbol{a}_i$ satisfying $\boldsymbol{a}_i^\top \mathrm{Proj}_C(\boldsymbol{y}) = b_i$.

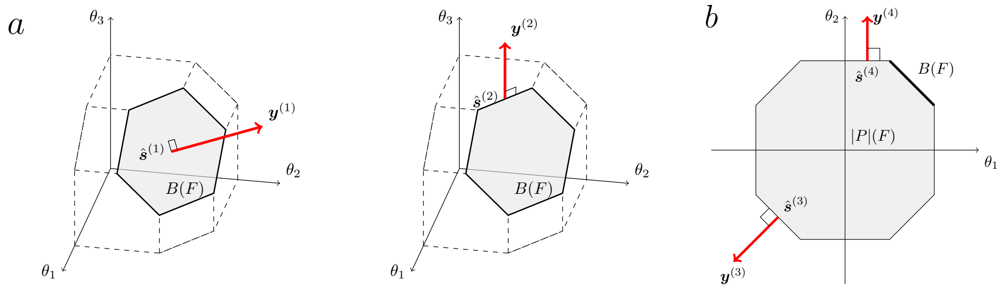

Fig. 4.1: Submodular polyhedra and submodular regularization estimators. a: An LERE $\hat{\boldsymbol{\theta}}$ is an anti-projection estimator with respect to $B(f)$, which implies that $\hat{\boldsymbol{\theta}}$ is a normal vector of $B(f)$. Left: Suppose that $\hat{\boldsymbol{s}}^{(1)} = \mathrm{Proj}_{B(f)}(\boldsymbol{y}^{(1)})$ is contained in the relative interior a two-dimensional face. Then, the LERE solution $\hat{\theta}^{(1)} = \boldsymbol{y}^{(1)} - \hat{\boldsymbol{s}}^{(1)}$ must be parallel to $\mathbf{1}_V = (1,1,1)^\top$. This is because there are only two tight sets $\emptyset$ and V. Right: If $\hat{\boldsymbol{s}}^{(2)} = \mathrm{Proj}_{B(f)}(\boldsymbol{y}^{(2)})$ is contained in one-dimensional face, the dimension of the normal cone at $\hat{\boldsymbol{s}}^{(2)}$ is two. In the example shown in the figure, we have $\mathcal{D}(\hat{\boldsymbol{s}}^{(2)}) = \{\emptyset, \{1,2\}, \{1,2,3\}\}$. Hence, for any normal vector $\hat{\boldsymbol{\theta}}^{(2)}$ at $\hat{\boldsymbol{s}}^{(2)}$, the first and the second coordinates must be equal. b: Similarly, an SNRE is an anti-projection estimator with respect to $|P|(f)$. Unlike the case of LEREs, $|\hat{\boldsymbol{\theta}}|$ becomes sparse. See Section 4.4.2 for a detailed explanation.

Proposition 4.5 implies that an unbiased estimate of the degrees of freedom of an anti-projection estimator $\hat{\boldsymbol{\mu}}_C^{\mathrm{A}}(\boldsymbol{y}) = \boldsymbol{y} - \mathrm{Proj}_C(\boldsymbol{y})$ is given as $\hat{\mathrm{df}}(\hat{\boldsymbol{\mu}}_C^{\mathrm{A}})(\boldsymbol{y}) = \mathrm{rank}\boldsymbol{A}_0(\boldsymbol{y})$. To calculate $\mathrm{rank}\boldsymbol{A}_0$ for a general polyhedron C of the form (4.15), we have to enumerate the equality constraints. Importantly, such an enumeration can be computationally inefficient if the number of inequality constraints m is large. For the case of submodular regularization estimators, Lemma 4.3 shows that m grows exponentially with the dimensionality p. In the next subsection, we will derive other formulae for the degrees of freedom, which can be calculated in $\mathrm{O}(p \log p)$ for design matrices of full column rank.

4.4.2 Main results

We now explain our main results. In particular, Theorem 4.1 and Theorem 4.2 in the Section 4.1 are obtained as corollaries of Theorem 4.7 and Theorem 4.11, respectively.

To illustrate the relationship between a particular solution $\hat{\boldsymbol{\theta}}$ and the effective number of parameters, we first provide an example of LERE with $\boldsymbol{X} = \boldsymbol{I}_n$ and $\lambda = 1$. According to Lemma 4.3, the LERE is given as $\hat{\boldsymbol{\theta}} = \boldsymbol{y} - \mathrm{Proj}_{B(f)}(\boldsymbol{y})$, which is a normal vector of the base polyhedron $B(f)$ at $\boldsymbol{s} = \mathrm{Proj}_{B(f)}(\boldsymbol{y})$. Let $\mathcal{D}(\boldsymbol{s}) = \{A \subseteq V : \mathbf{1}_A^\top \boldsymbol{s} = f(A)\}$ be a collection tight sets at $\boldsymbol{s}$. From the general theory of polyhedra, any normal vector $\hat{\boldsymbol{\theta}}$ can be expressed by a linear combination of characteristic vectors $\{\mathbf{1}_A : A \in \mathcal{D}(\boldsymbol{s})\}$ (see 2.1.3). Suppose that there are two distinct indices $i, j \in V$ that cannot be separated by sets in $\mathcal{D}(\boldsymbol{s})$. Then, the i-th and j-th components of $\hat{\boldsymbol{\theta}}$ must be equal. In fact, if $\boldsymbol{s}$ is contained in the relative interior of a face with more than one dimension, then there exists a partition Π of V such that components of $\hat{\boldsymbol{\theta}}$ are constant for each element of Π. Therefore, the effective number of parameters is considered to be given by the size of the partition $|\Pi|$. Fig. 4.1a shows an illustrative example with $n = p = 3$.

Degrees of freedom of LERE

Here, we derive the degrees of freedom of LERE (4.3). The key observation is that every LERE solution is partition-wise constant. Our task is to characterize the degrees of freedom in terms of constant partitions of LERE solutions.

We first introduce some terminologies related to partitions. Let $\Pi = \{A_1, \dots, A_k\}$ be any partition of $V = [p]$. We say that $\boldsymbol{\theta} \in \mathbb{R}^p$ is (partition-wise) constant on Π if there exist numbers $t_1, \dots, t_k$ such that $\theta_i = t_k$ for all $i \in A_k$. Denote by $L(\Pi)$ the set of all partition-wise constant vectors on Π, i.e., $L(\Pi)$ is a linear subspace of $\mathbb{R}^p$ spanned by characteristic vectors $\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_k}$. Note that the orthogonal projection onto $L(\Pi)$ is given by the partition-wise average as

$$\mathrm{Proj}_{L(\Pi)}(\boldsymbol{\theta}) = \boldsymbol{P}_{L(\Pi)}\boldsymbol{\theta} = \sum_{j=1}^{k} \bar{\theta}_{A_j} \mathbf{1}_{A_j},$$

where $\bar{\theta}_{A_j} = \frac{1}{|A_j|}\sum_{i \in A_i} \theta_i$ for each $j \in \{1, \dots, k\}$. For any vector $\boldsymbol{\theta} \in \mathbb{R}^p$, we define the constant partition of $\boldsymbol{\theta}$ as the unique minimal partition Π satisfying $\boldsymbol{\theta} \in L(\Pi)$. We use $\Pi_{\mathrm{const}}(\boldsymbol{\theta})$ to denote the constant partition of $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{\theta}}$ be an arbitrary LERE solution. We will construct a data-dependent partition Π on which $\hat{\boldsymbol{\theta}}$ becomes constant. By Lemma 4.3 (i), $\hat{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ is included in the base polyhedron $B(f)$. Recall that the collection of tight sets $\mathcal{D}(\hat{\boldsymbol{s}})$ forms a distributive lattice containing $\emptyset$ and V (see Lemma 2.5), and that there is a partition $\Pi(\hat{\boldsymbol{s}})$ such that every element A in $\mathcal{D}(\hat{\boldsymbol{s}})$ is obtained as a union of some elements in $\Pi(\hat{\boldsymbol{s}})$ (see Theorem 2.6). Since $\hat{\boldsymbol{s}}$ is uniquely determined by $\boldsymbol{y}$, so is the distributive lattice $\mathcal{D}(\hat{\boldsymbol{s}})$. We define the following terminology:

Definition 4.6. For any $\boldsymbol{y} \in \mathbb{R}^n$, we define the boundary lattice at $\boldsymbol{y}$ as

$$\mathcal{D}_{\mathrm{bound}}(\boldsymbol{y}) := \mathcal{D}(\hat{\boldsymbol{s}}) = \{A \subseteq V : \langle \mathbf{1}_A, \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})\rangle = \lambda f(A)\}.$$

We also define the boundary partition $\Pi_{\mathrm{bound}}(\boldsymbol{y})$ as the partition associated with $\mathcal{D}_{\mathrm{bound}}(\boldsymbol{y})$ in the sense of Theorem 2.6.

Now, in fact, we can see that $\hat{\boldsymbol{\theta}}$ is constant on partition $\Pi(\hat{\boldsymbol{s}})$. From Lemma 4.3 (i), $\hat{\boldsymbol{\theta}}$ is contained in the normal cone $N_{B(f)}(\hat{\boldsymbol{s}})$. Since any normal vector at $\hat{\boldsymbol{s}}$ is spanned by the normal vectors corresponding to the equality constraints, we have $\hat{\boldsymbol{\theta}} \in \mathrm{span}\{\mathbf{1}_A : A \in \mathcal{D}(\hat{\boldsymbol{s}})\}$. Here, this set actually coincides with $L(\Pi(\hat{\boldsymbol{s}}))$. Indeed, by taking a maximal chain $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_k = V$ in $\mathcal{D}(\hat{\boldsymbol{s}})$, we have $L(\Pi(\hat{\boldsymbol{s}})) = \mathrm{span}\{\mathbf{1}_{S_i} - \mathbf{1}_{S_{i-1}} : i \in \{1, \dots, k\}\} \subseteq \mathrm{span}\{\mathbf{1}_A : A \in \mathcal{D}(\hat{\boldsymbol{s}})\}$. The opposite inclusion is clear from Theorem 2.6. Combining these, we have $\hat{\boldsymbol{\theta}} \in L(\Pi(\hat{\boldsymbol{s}}))$, which implies that $\hat{\boldsymbol{\theta}}$ is partition-wise constant on $\Pi(\hat{\boldsymbol{s}})$.

For any matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ and partition Π of V, we write

$$\boldsymbol{X}L(\Pi) := \{\boldsymbol{X}\boldsymbol{\theta} : \boldsymbol{\theta} \in L(\Pi)\} = \mathrm{span}\left\{\sum_{i \in A} \boldsymbol{x}_i : A \in \Pi\right\}.$$

We are now ready to present the degrees of freedom result for LERE.

Theorem 4.7 (The degrees of freedom of the LER)**.** Suppose that $f : 2^V \to \mathbb{R}$ is a submodular function and $\lambda > 0$. Then, the following statements are true for the degrees of freedom of the regression fit $\boldsymbol{y} \mapsto \boldsymbol{X}\hat{\boldsymbol{\theta}}$ of LERE (4.3).

(i) (Representation by the boundary partition). The degrees of freedom of the LERE is given as

$$\mathrm{df}(\boldsymbol{X}\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}_0}[\dim \boldsymbol{X}L(\Pi_{\mathrm{bound}}(\boldsymbol{y}))]. \tag{4.16}$$

In particular, if rank$\boldsymbol{X} = p$, then $\mathrm{df}(\boldsymbol{X}\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}_0}[|\Pi_{\mathrm{bound}}(\boldsymbol{y})|]$.

(ii) (Representation by the constant partition). For any $\boldsymbol{y} \in \mathbb{R}^n$, choose an LERE solution $\hat{\boldsymbol{\theta}}$ arbitrarily. Then, the value of $\dim \boldsymbol{X}L(\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}}))$ is uniquely determined for almost all $\boldsymbol{y} \in \mathbb{R}^n$. Moreover, the degrees of freedom is given as

$$\mathrm{df}(\boldsymbol{X}\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}_0}[\dim \boldsymbol{X}L(\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}}))] = \mathbb{E}_{\boldsymbol{\theta}_0}\left[\dim \mathrm{span}\left\{\sum_{i \in A} \boldsymbol{x}_i : A \in \Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})\right\}\right]. \tag{4.17}$$

In particular, if rank$\boldsymbol{X} = p$, an unbiased estimate of the degrees of freedom is given by (4.9).

Remark 4.8. Since the rank of a matrix is nor larger than the number of columns, (4.9) always gives an upper bound of the unbiased estimator based on (4.17). Similarly, for SNREs, (4.10) is an upper bound of the unbiased estimator derived in Theorem 4.11 (ii) below. Interestingly, in many practical cases, the simplified estimators output the same values as the exact unbiased estimators even if $\boldsymbol{X}$ is degenerated. See Section 4.6 for experimental results.

Proof sketch of Theorem 4.7.

The proof of Theorem 4.7 is based on the same framework as Tibshirani and Taylor (2012). As we defer a formal proof to 4.9, we provide a brief sketch of the proof. Let us write $L_1 := L(\Pi_{\mathrm{bound}}(\boldsymbol{y}))$, $L_2 := L(\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}}))$, and $\boldsymbol{P}_\ell := \boldsymbol{P}_{L_\ell}$, $\ell \in \{1, 2\}$. From the first-order optimality condition of the convex optimization problem (4.3), we have

$$\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+\boldsymbol{y} + \boldsymbol{v}_\ell, \qquad \ell \in \{1, 2\}.$$

Here, $\boldsymbol{v}_\ell \in \mathbb{R}^n$, $\ell \in \{1, 2\}$, possibly depends on $\boldsymbol{y}$. We want to show that the boundary partition $\Pi_{\mathrm{bound}}(\boldsymbol{y})$, the constant partition $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$, and vectors $\boldsymbol{v}_\ell$ can be taken as locally invariant. If this is true, the regression fit $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ is locally affine. Thus, the derivative is (almost surely) given as follows:

$$(\nabla \cdot \boldsymbol{X}\hat{\theta})(\boldsymbol{y}) = \mathrm{tr}[(\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+] = \dim \boldsymbol{X}L_\ell, \quad \ell \in \{1, 2\},$$

where the second equality holds because $\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^+)$ equals to the number of non-zero singular values of $\boldsymbol{A}$.

The local invariance property of the boundary lattice and the constant partitions holds from the following two lemmas, which are proved in 4.9.

Lemma 4.9. The boundary lattice $\mathcal{D}_{\mathrm{bound}}(\boldsymbol{y})$ is locally invariant for almost all $\boldsymbol{y} \in \mathbb{R}^n$. To be more precise, there exists a measure zero set $\mathcal{M}_1 \subset \mathbb{R}^n$ with the following property: for any $\boldsymbol{y} \notin \mathcal{M}_1$, there exists an open neighborhood U of $\boldsymbol{y}$ such that $\hat{\mathcal{D}}(\boldsymbol{y}') = \hat{\mathcal{D}}(\boldsymbol{y})$ for all $\boldsymbol{y}' \in U$.

Lemma 4.10. There exists a measure-zero set $\mathcal{M}_2 \subset \mathbb{R}^n$ with the following property: for any $\boldsymbol{y} \notin \mathcal{M}_2$, choose an arbitrary LERE solution $\hat{\boldsymbol{\theta}}$ in (4.3). Then, there exists a neighborhood U of $\boldsymbol{y}$ such that for any $\boldsymbol{y}' \in U$ we can choose a LERE solution $\hat{\boldsymbol{\theta}}'$ with the same constant partition $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}}') = \Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$.

Degrees of freedom of SNRE

Next, we derive the degrees of freedom of SNRE (4.4). Deriving the degrees of freedom of SNRE is slightly more complicated than LERE. The difference is that, in SNR, typical solutions of (4.4) are sparse.

Here, we explain why solutions become sparse. For simplicity, let $\boldsymbol{X} = \boldsymbol{I}_n$ and $\lambda = 1$. By Lemma 4.3 (ii), an SNRE solution $\hat{\boldsymbol{\theta}}$ is a normal vector of the symmetric submodular polyhedron $|P|(f)$ at a projected point $\hat{\boldsymbol{s}} = \mathrm{Proj}_{|P|(f)}(\boldsymbol{y})$. The dimension of the normal cone $N_{|P|(f)}(\hat{\boldsymbol{s}})$ coincides with the codimension of the minimal face of $|P|(f)$ containing $\hat{\boldsymbol{s}}$. Observe that $|P|(f)$ is the convex hull of the union of the sign-inverted base polyhedra:

$$|P|(f) = \mathrm{conv}\left(\bigcup_{\boldsymbol{\gamma} \in \{-1,1\}^p} \boldsymbol{\gamma} \odot B(f) \right).$$

Hence, every face F of $|P|(f)$ can be obtained by one of the following two cases:

(i) A sign inversion of a face of the base polyhedron, i.e., $F = \boldsymbol{\gamma} \odot F'$ for some $\boldsymbol{\gamma} \in \{-1,1\}^p$ and a face F' of $B(f)$.

(ii) A face generated by the convex hull operation. For example, in the case of $n = p = 2$, this is a line segment that connects a vertex and its reflection in the horizontal or the vertical axis.

Fig. 4.1b shows an example with $n = p = 2$. Suppose that $\hat{\boldsymbol{s}}^{(3)} = \mathrm{Proj}_{|P|(f)}(\boldsymbol{y}^{(3)})$ is contained in an sign-inverted base polyhedron $\binom{-1}{-1} \odot B(f)$. Then, the SNRE solution $\hat{\boldsymbol{\theta}}^{(3)}$ is parallel to the vector $\binom{-1}{-1}$. On the other hand, suppose that $\hat{\boldsymbol{s}}^{(4)} = \boldsymbol{P}_{|P|(f)}(\boldsymbol{y}^{(4)})$ is on the edge connecting $B(f)$ and $\binom{-1}{1} \odot B(f)$. In this case, the first coordinate of $\hat{\boldsymbol{\theta}}^{(4)}$ becomes zero. Generally speaking, the sparsity of SNRE solutions comes from the fact that normal vectors of the faces of type (ii) are sparse.

To give a formal statement for our theorem, we provide notations related to sparse and partition-wise absolute constant vectors.

First, we define the notion of sparse constant partitions. Given a triple $(Z, \Pi, \boldsymbol{\gamma})$, where $Z \subseteq V$ is a subset, Π is a partition of $V - Z$ and $\boldsymbol{\gamma} \in \{-1,1\}^p$ is a sign vector, we define the following linear subspace:

$$L_0(Z, \Pi, \boldsymbol{\gamma}) := \boldsymbol{\gamma} \odot \mathrm{span}\{\mathbf{1}_A : A \in \Pi\}.$$

Here, we define that $L_0(V, \emptyset, \boldsymbol{\gamma}) = \{\mathbf{0}\}$ if $Z = V$. $L_0(Z, \Pi, \boldsymbol{\gamma})$ is a set of vectors $\boldsymbol{v}$ such that $\boldsymbol{\theta}_Z = \mathbf{0}$ and $|\boldsymbol{\theta}| = \boldsymbol{\gamma} \odot \boldsymbol{v}$ is constant on partition Π. Note that the sign vector $\boldsymbol{\gamma}$ indicates only whether the signs of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ $(i, j \in V - Z)$ are equal, and it is clear by definition that $L_0(Z, \Pi, \boldsymbol{\gamma}) = L_0(Z, \Pi, -\boldsymbol{\gamma})$. For a vector $\boldsymbol{\theta} \in \mathbb{R}^p$, we write $Z(\boldsymbol{\theta}) = \{i \in V : \theta_i = 0\}$ for the set of indices of zero components. We also write $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \mathrm{sign}(\boldsymbol{\theta})$. We define $\Pi_{\mathrm{const},0}(\boldsymbol{\theta})$ as the smallest partition Π of $V - Z(\boldsymbol{\theta})$ such that $\boldsymbol{\theta} \in L_0(Z(\boldsymbol{\theta}), \Pi, \gamma(\boldsymbol{\theta}))$. An important fact is that the number of distinct non-zero absolute values in $|\theta_1|, \ldots, |\theta_p|$ equals to $|\Pi_{\mathrm{const},0}(\boldsymbol{\theta})|$. Based on these definitions, we define the sparse constant partition of $\boldsymbol{\theta}$ as triple $(Z(\boldsymbol{\theta}), \Pi_{\mathrm{const},0}(\boldsymbol{\theta}), \boldsymbol{\gamma}(\boldsymbol{\theta}))$.

Next, we define the sparse boundary lattice, which plays a role that corresponds to the boundary lattice discussed in Section 4.4.2. Let $f : 2^V \to \mathbb{R}$ be a monotone non-decreasing submodular function, and $\hat{\boldsymbol{\theta}}$ be a solution of (4.4). By Lemma 4.3 (ii), $\hat{s} = \lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ is included in $|P|(f)$, which implies $|\hat{\boldsymbol{s}}| \in P(f)$. Let $\mathcal{D}_{\mathrm{bound},0}(\hat{\boldsymbol{s}}) = \mathcal{D}(|\hat{\boldsymbol{s}}|)$ be a distributive lattice of the equality constraints obtained according to Lemma 2.5. We also use $\gamma_{\text{bound},0}(\boldsymbol{y}) = \gamma(\hat{\boldsymbol{s}})$ to denote the sign vector of $\hat{\boldsymbol{s}}$. Then, we refer to the pair $(\mathcal{D}_{\text{bound},0}(\boldsymbol{y}), \gamma_{\text{bound},0}(\boldsymbol{y}))$ as the sparse boundary lattice at $\boldsymbol{y}$.

Since $\mathcal{D}_{\text{bound},0}(\boldsymbol{y})$ is union-closed, there exists a unique maximal element $A_0 = A_0(\boldsymbol{y}) \in \hat{\mathcal{D}}_0(\boldsymbol{y})$. By regarding $\mathcal{D}_{\text{bound},0}(\boldsymbol{y})$ as a sublattice of 2^{A_0}, we obtain a partition $\Pi_{\text{bound},0}(\boldsymbol{y}) = \Pi(\mathcal{D}_{\text{bound},0}(\boldsymbol{y}))$ of A_0 by Theorem 2.6. Suppose that A_0 is a proper subset of V, and that $Z_0(\boldsymbol{y}) := V - A_0(\boldsymbol{y})$. Then, for any vector $\boldsymbol{\theta}$ in $\text{span}\{\boldsymbol{1}_A : A \in \Pi_{\text{bound},0}(\boldsymbol{y})\}$, the subvector of $\boldsymbol{\theta}$ restricted on $Z_0(\boldsymbol{y})$ is zero. Thus, we have a canonical linear subspace $L_0(Z_0(\boldsymbol{y}), \Pi_{\text{bound},0}(\boldsymbol{y}), \gamma_{\text{bound},0}(\boldsymbol{y}))$ associated with the sparse boundary lattice at $\boldsymbol{y}$.

We can now state the degrees of freedom result for SNRE that corresponds to Theorem 4.7 for the LERE.

Theorem 4.11 (The degrees of freedom of the SNR)**.** Suppose that $f : 2^V \to \mathbb{R}$ is a non-decreasing submodular function and $\lambda > 0$. Then, the following statements are true for the degrees of freedom of the regression fit $\boldsymbol{y} \mapsto \boldsymbol{X}\hat{\boldsymbol{\theta}}$ of SNRE (4.4).

1. (Representation by the sparse boundary lattice). Let
$$\mathcal{L}_1(\boldsymbol{y}) := L_0(Z_0(\boldsymbol{y}), \Pi_{\text{bound},0}(\boldsymbol{y}), \gamma_{\text{bound},0}(\boldsymbol{y}))$$
be a linear subspace associated with the sparse boundary lattice. The degrees of freedom of the SNRE is given as:
$$\text{df}(\boldsymbol{X}\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}_0}[\dim \boldsymbol{X}\mathcal{L}_1(\boldsymbol{y})]. \tag{4.18}$$
In particular, if $\text{rank}\boldsymbol{X} = p$, then $\text{df}(\boldsymbol{X}\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}_0}[|\Pi_{\text{bound},0}(\boldsymbol{y})|]$.
2. (Representation by the sparse constant partition). For any $\boldsymbol{y} \in \mathbb{R}^n$, choose an SNRE solution $\hat{\boldsymbol{\theta}}$ arbitrarily. Let $\mathcal{L}_2(\hat{\boldsymbol{\theta}}) := L_0(Z(\hat{\boldsymbol{\theta}}), \Pi_{\text{const},0}(\hat{\boldsymbol{\theta}}), \gamma(\hat{\theta}))$. Then, the value of $\dim \boldsymbol{X}\mathcal{L}_2(\hat{\boldsymbol{\theta}})$ is uniquely determined for almost all $\boldsymbol{y} \in \mathbb{R}^n$. Moreover, the degrees of freedom is given as
$$\begin{aligned}\text{df}(\boldsymbol{X}\hat{\boldsymbol{\theta}}) &= \mathbb{E}_{\boldsymbol{\theta}_0}[\dim \boldsymbol{X}\mathcal{L}_2(\hat{\boldsymbol{\theta}})] \\ &= \mathbb{E}_{\theta_0}\left[\dim \text{span}\left\{\sum_{i \in A} \text{sign}(\hat{\theta}_i)\boldsymbol{x}_i : A \in \Pi_{\text{const}}(|\hat{\boldsymbol{\theta}}|), \hat{\boldsymbol{\theta}}_A \neq 0\right\}\right].\end{aligned} \tag{4.19}$$
In particular, if $\text{rank}\boldsymbol{X} = p$, an unbiased estimate of the degrees of freedom is given by (4.10).

Proof sketch of Theorem 4.11.

The proof follows basically the same process as that of Theorem 4.7. Let us write $\boldsymbol{P}_\ell = \boldsymbol{P}_{\mathcal{L}_\ell}$, $\ell \in \{1, 2\}$, with
$$\mathcal{L}_1 = \mathcal{L}_1(\boldsymbol{y}) := L_0(Z_0(\boldsymbol{y}), \Pi_{\text{bound},0}(\boldsymbol{y}), \gamma_{\text{bound},0}(\boldsymbol{y})), \tag{4.20}$$
and
$$\mathcal{L}_2 = \mathcal{L}_2(\hat{\boldsymbol{\theta}}) := L_0(Z(\hat{\boldsymbol{\theta}}), \Pi_{\text{const},0}(\hat{\boldsymbol{\theta}}), \gamma(\hat{\boldsymbol{\theta}})), \tag{4.21}$$
respectively. By definition, $\boldsymbol{P}_2\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ holds. We can also show that $\boldsymbol{P}_1\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ according to the following lemma.

Lemma 4.12. Any solution $\hat{\boldsymbol{\theta}}$ of SNRE is contained in $\mathcal{L}_1$.

A similar argument for LERE yields a locally affine representation of the regression fit $\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+\boldsymbol{y} + \boldsymbol{v}_\ell$, $\ell \in \{1, 2\}$. Thus, if we prove the local invariance results of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, the assertion follows from Stein's lemma.

The precise statements for the local invariances of the sparse boundary lattice and the sparse constant partition are as follows.

Lemma 4.13. The sparse boundary lattice is locally invariant for almost all $\boldsymbol{y} \in \mathbb{R}^n$. To be more precise, there exists a measure zero set $\mathcal{M}_3 \subset \mathbb{R}^n$ with the following property; for any $\boldsymbol{y} \notin \mathcal{M}_3$, there exists an open neighborhood U of $\boldsymbol{y}$ such that $\mathcal{D}_{\mathrm{bound},0}(\boldsymbol{y}') = \hat{\mathcal{D}}_0(\boldsymbol{y})$ and $\gamma_{\mathrm{bound},0}(\boldsymbol{y}') = \hat{\gamma}(\boldsymbol{y})$ for all $\boldsymbol{y}' \in U$.

Lemma 4.14. There exists a measure zero set $\mathcal{M}_4 \subset \mathbb{R}^n$ with the following property: for any $\boldsymbol{y} \notin \mathcal{M}_4$, choose an arbitrary SNRE solution $\hat{\boldsymbol{\theta}}$ in (4.4). Then, there exists a neighborhood U of $\boldsymbol{y}$ such that for any $\boldsymbol{y}' \in U$ we can choose a SNRE solution $\hat{\boldsymbol{\theta}}'$ with the same sparse constant partition as $\hat{\boldsymbol{\theta}}$.

4.5 Examples of submodular regularization

In this section, we provide examples of submodular regularization and their degrees of freedom results. We will also discuss the connections to the previous study.

4.5.1 LER

Fused lasso

The fused lasso is a typical example of LERE. Let $G = (V, E, w)$ be a connected undirected graph with a non-negative edge weight $w : E \to \mathbb{R}$. The cut function $f_{\mathrm{cut}}(A) := \sum_{(i,j)\in E:\ i\in A, j\notin A} w_{i,j}$ is known to be submodular, and its Lovász extension is given as

$$\hat{f}_{\mathrm{cut}}(\boldsymbol{\theta}) = \sum_{(i,j)\in E} w_{i,j}|\theta_i - \theta_j|.$$

$\hat{f}$ coincides with the regularization term of the (generalized) fused lasso (Tibshirani et al. 2005). As mentioned above, this regularization term acts as a smoother based on the adjacency on graph G.

The fused lasso can also be interpreted as a special case of the generalized lasso (Tibshirani and Taylor 2011). Given a matrix $\boldsymbol{D} \in \mathbb{R}^{m\times p}$, the generalized lasso estimator is defined as

$$\hat{\boldsymbol{\theta}}_{\mathrm{genlasso}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{D}\boldsymbol{\theta}\|_1.$$

In particular, if $\boldsymbol{D}$ is the incidence matrix of the weighted graph G, the regularization term $\|\boldsymbol{D}\boldsymbol{\theta}\|_1$ equals to the fused regularization term $\hat{f}_{\mathrm{cut}}(\boldsymbol{\theta})$. Tibshirani and Taylor (2011) showed that, if $\boldsymbol{X}$ has full column rank, the degrees of freedom is given by the number of fused groups. Here, a fused group means a connected component of G on which the values of $\hat{\boldsymbol{\theta}}$ are identical. For a general low rank matrix $\boldsymbol{X}$, the degrees of freedom is given by the same authors (Tibshirani and Taylor 2012).

An unbiased estimator derived in our theory is essentially equivalent to the above. From Theorem 4.7, an unbiased estimator of the degree of freedom for the full-rank fused lasso is given by the number of elements in the constant partitions $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$. Theorem 4.7 alone does not give the information that each element in $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$ is connected. However, the following property holds:

Proposition 4.15. Suppose that $\boldsymbol{X}$ has full column rank. For almost all $\boldsymbol{y} \in \mathbb{R}^n$, every element $A \in \Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$ is a connected component on the graph G.

Nearly-isotonic type regularization

The isotonic regression has a long history in statistics (Robertson et al. 1988). Let $(V, \preceq)$ be a partially ordered set. The isotonic regression is defined as a least squares estimator onto the set of monotone vectors with respect to $\preceq$:

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{\theta}\|_2 \quad \text{subject to} \quad \forall i \preceq j, \quad \theta_i \leq \theta_j. \tag{4.22}$$

There is an alternative formulation of the isotonic regression using directed acyclic graphs (DAGs). Let $G = (V, E)$ be a DAG. Then, G induces a partial order $\preceq_G$ on V by defining $i \preceq_G j$ when there exists a directed path from i to j. Conversely, given a partially ordered set $(V, \preceq)$, there is a DAG $G = (V, E_\preceq)$ whose induced order $\preceq_G$ coincides with the original order $\preceq$. The isotonic regression estimator on the graph G is defined as the solution of the following least squares problem:

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{\theta}\|_2 \quad \text{subject to} \quad \forall (i,j) \in E, \quad \theta_i \leq \theta_j. \tag{4.23}$$

Indeed, the above two formulations (4.22) and (4.23) define the same class of projection estimators onto polyhedral convex cones of the form $K_G^\uparrow := \{\boldsymbol{\theta} \in \mathbb{R}^n : \forall (i,j) \in E, \quad \theta_i \leq \theta_j\}$.

The nearly-isotonic regression (Tibshirani et al. 2011) is a regularization type variant of the one-dimensional isotonic regression estimator

$$\hat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^n} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{n-1} \max\{\theta_i - \theta_{i+1}, 0\} \right). \tag{4.24}$$

More generally, we can define a directed graph regularization estimator on a weighted graph $G = (V, E, w)$ as

$$\hat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2 + \lambda \sum_{(i,j) \in E} w_{i,j} \max\{\theta_i - \theta_j, 0\} \right). \tag{4.25}$$

This is an example of LERE. Indeed, the regularization term in (4.25) is the Lovász extension of the cut function of G. For each directed edge $(i,j) \in E$, the regularization term in (4.25) imposes penalty $w_{i,j}|\theta_i - \theta_j|$ if $\theta_i > \theta_j$. If G is a DAG, taking the limit as $\lambda \to \infty$ yields the same solution as the graph isotonic regression estimator (4.23).

Proposition 4.16. Let $G = (V, E)$ be a DAG, and $\{w_{i,j}\}_{(i,j) \in E}$ be positive weights on edges. Fix $\boldsymbol{y} \in \mathbb{R}^n$. Then, there exists $\lambda_+ \in (0, \infty)$ such that the directed graph regularization (4.25) with tuning parameter $\lambda \geq \lambda_+$ admits the same solution as the projection estimator (4.23).

For the projection estimators (4.23), Meyer and Woodroofe (2000) showed that an unbiased estimate of the degrees of freedom is given as the number of unique values in the solution $\hat{\boldsymbol{\theta}}$. For the nearly-isotonic regression (4.24), Tibshirani et al. (2011) showed that the degrees of freedom has the same representation. In fact, we have a similar result for generalized estimators (4.25), which is a corollary of Theorem 4.1.

Corollary 4.17. Let $G = (V, E)$ be a directed graph, and $\{w_{i,j}\}_{(i,j) \in E}$ be non-negative weights on edges. Then, an unbiased estimate of the degrees of freedom of (4.25) is given by (4.9).

We should note that, by a similar argument as Proposition 4.15, the constant partition of (4.25) consists of (weakly) connected components in G.

Higher order regularization

The graph cut function is of order two in the sense that the value $f(A)$ is determined by weights over all singletons and pairs in A. We can consider LERs for three or more higher order functions. Hypergraph cut functions are examples of higher order adjacency functions. Let $\mathcal{H} = (V, H, w)$ be a hypergraph, where hyperedge set H is a collection of subsets of V, and $w : H \to \mathbb{R}_+$ is a non-negative weight. Then, a hypergraph cut function $f : 2^V \to \mathbb{R}$ is defined as

$$f(A) = \sum_{h \in H} w_h 1_{\{h \cap A \neq \emptyset,\ h \cap (V-A) \neq \emptyset\}}.$$

The Lovász extension is written as

$$\hat{f}(\boldsymbol{\theta}) = \sum_{h \in H} w_h \max_{i,j \in h} |\theta_i - \theta_j|.$$

Hein et al. (2013) proposed the total variation regularization on hypergraphs, which is a LERE for the hypergraph cut function. In addition, Takeuchi et al. (2015) proposed another LER related to hypergraphs.

4.5.2 SNR

Lasso

As mentioned above, the lasso is given as SNRE for the cardinality function $f(A) = |A|$. For full-rank design matrices, Zou et al. (2007) showed that the degrees of freedom of the lasso is given by the number of non-zero components. Theorem 4.11 provides essentially the same estimate as this result. In fact, for almost all $\boldsymbol{y} \in \mathbb{R}^n$, the non-zero components in $\hat{\boldsymbol{\theta}}_{\text{lasso}}$ are partitioned into singleton. This is because the base polyhedron of the cardinality function degenerates to a single point, and hence the normal vectors are not partition-wise constant. For the same reason, Theorem 4.11 recovers an existing result for general design matrices (Tibshirani and Taylor 2012).

Group lasso and variants

The ℓ_1/ℓ_2-group lasso (Yuan and Lin 2006) is a norm based regularization method that uses group structures to generate sparsity patterns. The degrees of freedom of ℓ_1/ℓ_2-group lasso was considered in (Kato 2009, Vaiter et al. 2012).

Here, we consider another variant of the group lasso, i.e., ℓ_1/ℓ_∞-group lasso. Let $\mathcal{G} = \{A_1, \ldots, A_k\}$ be a collection of subsets of V with $\bigcup_i G_i = V$. The weighted ℓ_1/ℓ_∞-group norm is defined as

$$\Omega_{\mathcal{G},\infty}(\boldsymbol{\theta}) := \sum_{i=1}^{k} w_i \|\boldsymbol{\theta}_{A_i}\|_\infty,$$

where $\|\boldsymbol{\theta}_{A_i}\|_\infty := \max_{i \in A_i} |\theta_i|$, and $w_i > 0$, $i \in \{1, \ldots, k\}$, are positive weights. Bach (2010) pointed out that this is a submodular norm associated with the overlapping count function defined as

$$f(A) = \sum_{i=1}^{k} w_i 1_{\{A \cap A_i \neq \emptyset\}}.$$

This relationship holds true regardless of whether $\mathcal{G}$ is a partition of V, i.e., elements in $\mathcal{G}$ can have overlaps. The degrees of freedom of ℓ_1/ℓ_∞-group lasso is given by Theorem 4.11.

OSCAR and SLOPE

The ordered weighted ℓ_1 norm (OWL) is defined as

$$\Omega_{\text{OWL},w}(\boldsymbol{\theta}) := \sum_{i=1}^{p} w_i |\theta_{\tau(i)}|,$$

where $w_1 > \cdots > w_p > 0$ is a decreasing sequence and $\tau : V \to V$ is a permutation that sorts $|\theta_i|$s in descending order. We see that this is the submodular norm that corresponds to

$$f_{\text{OWL},w}(A) := \begin{cases} \sum_{i=0}^{|A|} w_i, & A \neq \emptyset \\ 0, & A = \emptyset \end{cases},$$

which is a composition of the cardinality function $A \mapsto |A|$ and a concave function. The OSCAR penalty (Bondell and Reich 2008) and the SLOPE penalty (Bogdan et al. 2015) are examples of OWL penalties. The OSCAR penalty is defined as $\Omega_{\text{OSCAR}}(\boldsymbol{\theta}) = \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\theta_i|, |\theta|_j\}$, which is obtained as an OWL penalty with $w_i = \lambda_1 + \lambda_2 \left[\binom{p}{2} - \binom{i-1}{2}\right]$ (Obozinski and Bach 2016). The SLOPE penalty is another example with $w_i = \Phi^{-1}(1 - qi/2p)$, where Φ is the distribution function of the standard normal distribution, and $q \in (0, 1)$. Recent studies showed that the SLOPE estimator has nice adaptation properties to the unknown sparsity pattern (Su and Candés 2016, Bellec et al. 2018).

4.6 Numerical examples

This section contains numerical examples that illustrate the performance of the unbiased degrees of freedom estimators derived in this chapter. We focus on the SLOPE estimators. As mentioned in Section 4.5.2, the SLOPE is an example of SNREs, and an unbiased estimator of the degrees of freedom is given by Theorem 4.11.

First, we compared the following three estimators for the degrees of freedom:

(i) The estimator $\hat{\text{df}}_{\text{Exact}}$ defined as (4.19). This is the "exact" unbiased estimator derived as a consequence of our main theorem.

(ii) An "inexact" estimator $\hat{\text{df}}_{\text{Inexact}}$ defined as (4.10). This estimator ignores the rank degeneration in (4.19). Hence, we always have $\hat{\text{df}}_{\text{Exact}} \leq \hat{\text{df}}_{\text{Inexact}}$, and the equality holds whenever $\boldsymbol{X}$ has full column rank. Calculating $\hat{\text{df}}_{\text{Inexact}}$ requires only $\mathrm{O}(p \log p)$ comparisons of coordinate values in $\hat{\boldsymbol{\theta}}_\lambda$.

(iii) The unbiased estimator for the lasso estimator $\hat{\text{df}}_{\text{Lasso}}$ derived in (Tibshirani and Taylor 2012,Theorem 2), where $\hat{\text{df}}_{\text{Lasso}}$ is given as the dimension of the linear subspace $\text{span}\{\boldsymbol{x}_i : \hat{\theta}_i \neq 0\}$. Since this estimator ignores equalities among non-zero coordinates of $\hat{\boldsymbol{\theta}}$, we always have $\hat{\text{df}}_{\text{Exact}} \leq \hat{\text{df}}_{\text{Lasso}}$.

The data used in the simulation was obtained as follows. We fixed $n = 40$. We generated and fixed $n \times p$ design matrices $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)^\top$ with i.i.d. Gaussian rows $\boldsymbol{x}^i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{p,\rho})$. Here, $\boldsymbol{\Sigma}_{p,\rho}$ is a $p \times p$ matrix with diagonal elements equal to 1 and off-diagonal elements equal to $\rho \in \{0, 0.8\}$. For each $p \in \{20, 40, 80\}$, the true regression coefficient is set to $\boldsymbol{\theta} = (1, 1, 1, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$. Then, the target variables are generated according to the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.

To visualize the performance of the above three estimators, we also estimated the "true" degrees of freedom using Monte-Carlo simulations over the definition (4.5). We drawn

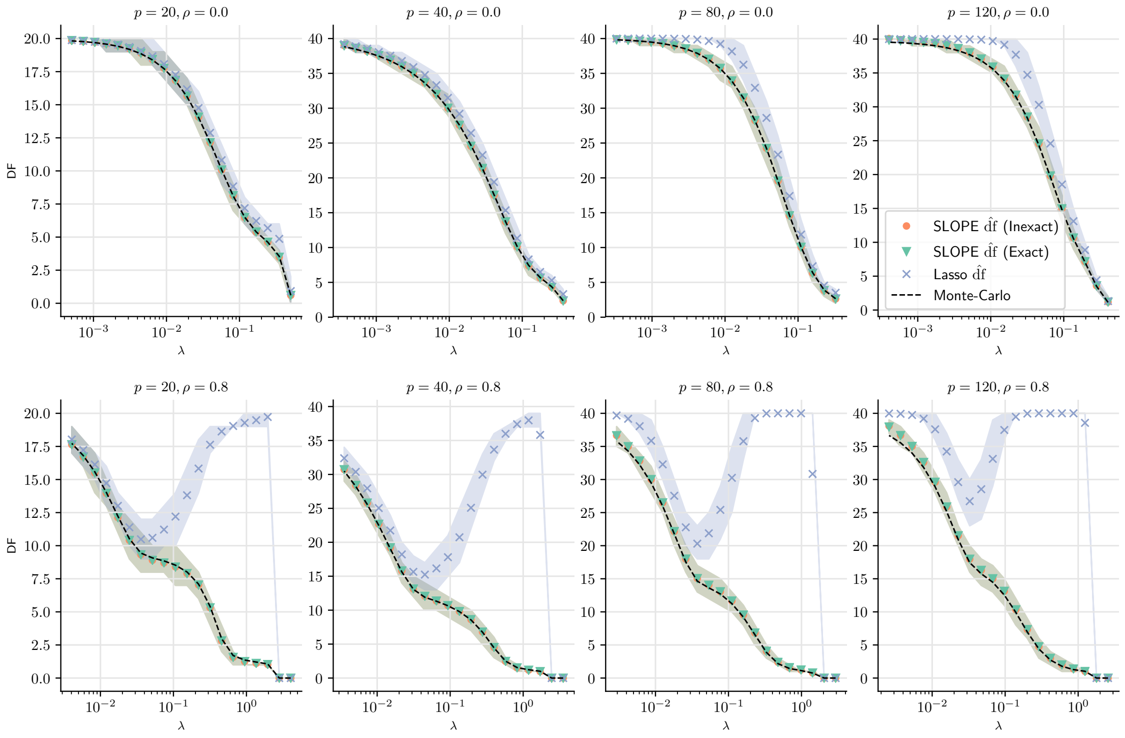

Fig. 4.2: Degrees of freedom (DF) and its estimators for SLOPE. The true DFs (dashed black lines) are calculated by Monte-Carlo simulations over the definition (4.5). The unbiased estimators (4.19) (green triangles) and upper bounds (4.10) (orange circles) are averaged over 2000 realizations of noises. The regions between 25% and 75% quantiles are also filled.

$M = 2000$ independent copies of the observation $\boldsymbol{y}^{(m)} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\xi}^{(m)}$, $m \in \{1, \ldots, M\}$, and calculated the following quantity:

$$\tilde{\mathrm{df}}_{\mathrm{MC}} := \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{\xi}_i^{(m)} [\boldsymbol{X}\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{y}^{(m)})]_i .$$

Here, we calculated the SLOPE estimator $\hat{\boldsymbol{\theta}}_\lambda$ by the FISTA algorithm explained in (Bogdan et al. 2015). For the stopping criterion, we stopped the algorithm if the duality gap becomes less than 10^{-8} or the number of iterations exceeds 3000. Note that we cannot calculate $\tilde{\mathrm{df}}_{\mathrm{MC}}$ only from the data because it requires information about the true noise $\boldsymbol{\xi}$.

We computed three estimators $\hat{\mathrm{df}}_{\mathrm{Exact}}$, $\hat{\mathrm{df}}_{\mathrm{Inexact}}$, and $\hat{\mathrm{df}}_{\mathrm{Lasso}}$ over 2000 realizations of the target variables. Since $\hat{\boldsymbol{\theta}}_\lambda$ is a vector of floating point numbers and calculated by an iterative algorithm, we used tolerance number $\epsilon = 10^{-8}$ to judge the equality $|\hat{\theta}_i| = |\hat{\theta}_j|$. Fig. 4.2 shows the results. The unbiased estimators $\hat{\mathrm{df}}_{\mathrm{Exact}}$ are reasonably close to the Monte-Carlo estimates $\tilde{\mathrm{df}}_{\mathrm{MC}}$. The inexact estimator $\hat{\mathrm{df}}_{\mathrm{Inexact}}$ is not guaranteed to be unbiased. However, in our simulations, their values were almost same as the exact estimators. On the other hand, the misspecified estimator $\hat{\mathrm{df}}_{\mathrm{Lasso}}$ clearly overestimates the degrees of freedom especially in high-correlation settings (i.e., $\rho = 0.8$).

Next, we consider the performance of the SURE-tuned SLOPE estimator. We chose the tuning parameter of the SLOPE estimator by the C_p-type criterion $\hat{\lambda} \in \operatorname{argmin}_{\lambda \geq 0} C_p(\lambda)$,

Table 4.1: Prediction errors of the C_p-type criterion (C_p), the 5-fold cross-validation (CV), and the oracle estimator (Oracle). For each (n, p), the prediction error $\|\boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2/n$ is averaged over 100 realizations of the noise variables. The numbers in parenthesis are the standard deviations.

		C_p	CV	Oracle
$n = 40$	$p = 20$	0.308 (0.128)	0.335 (0.139)	0.263 (0.117)
	$p = 40$	0.397 (0.140)	0.424 (0.168)	0.343 (0.120)
	$p = 80$	0.514 (0.179)	0.637 (0.227)	0.456 (0.159)
	$p = 120$	0.588 (0.189)	0.706 (0.245)	0.524 (0.175)
$n = 200$	$p = 20$	0.062 (0.027)	0.064 (0.026)	0.050 (0.022)
	$p = 40$	0.090 (0.034)	0.094 (0.032)	0.076 (0.027)
	$p = 80$	0.108 (0.045)	0.111 (0.042)	0.095 (0.038)
	$p = 120$	0.137 (0.046)	0.137 (0.048)	0.119 (0.037)
$n = 400$	$p = 20$	0.032 (0.018)	0.033 (0.016)	0.027 (0.015)
	$p = 40$	0.050 (0.018)	0.052 (0.019)	0.043 (0.016)
	$p = 80$	0.060 (0.024)	0.060 (0.022)	0.052 (0.020)
	$p = 120$	0.071 (0.026)	0.073 (0.027)	0.064 (0.021)

where

$$C_p(\lambda) := \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_\lambda\| + \frac{2\sigma^2}{n}\hat{\mathrm{df}}(\hat{\boldsymbol{\theta}}_\lambda).$$

Table 4.1 compares the prediction errors $\|\boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2/n$ of the slope estimators with different parameter selection rules, i.e., the C_p-type criterion (C_p) and the 5-folds cross-validation (CV). “Oracle” stands for the oracle choice of parameters $\lambda_{\mathrm{oracle}} \in \mathrm{argmin}_{\lambda \geq 0}\|\boldsymbol{X}\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2/n$ that cannot be observed. The design matrix $\boldsymbol{X}$ is obtained similarly to the previous examples with $\rho = 0$. The prediction errors are averaged over 100 realizations of the noise variables. We can see that the C_p-type criterion performs better than the CV for every pair of (n, p). We should note that calculating the C_p-type criterion requires the noise variance σ^2, while the CV works without knowing it.

4.7 Discussion on variance estimation

In Section 4.6, we investigated empirical performances of SURE-tuned estimators. For general settings, providing a unified theoretical guarantee for SURE-based parameter selection remains as an open question (see (Tibshirani and Rosset 2019) for a recent development for this topic). However, during the review process of corresponding publication of this chapter (Minami 2020), one of the anonymous reviewers suggested that an unbiased estimator for *variance* of SURE can be computed from the data. It may provide a data-dependent reliability of SURE-based parameter selection. The following proposition is a consequence of Theorem 2.1 of (Bellec and Zhang 2018).

Proposition 4.18. Let $\hat{\boldsymbol{\theta}}$ be an LERE (or SNRE), and $\hat{\mathrm{df}}$ be the unbiased estimator of the degrees of freedom derived in Theorem 4.7 (or Theorem 4.11). Let $\hat{R}_{\mathrm{SURE}} := \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\|_2^2 + 2\sigma^2\hat{\mathrm{df}} - n\sigma^2$. Then,

$$\mathbb{E}_{\boldsymbol{\theta}_0}[(\hat{R}_{\mathrm{SURE}} - \|X(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_2^2)^2] = \mathbb{E}_{\boldsymbol{\theta}_0}[4\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\|_2^2 + 4\sigma^4\hat{\mathrm{df}} - 2n\sigma^4].$$

In fact, we can check that a similar variance evaluation holds for any anti-projection estimator.

4.8 Proofs: Structural results for submodular polyhedra

The remaining three sections of this chapter provide missing proofs in the previous sections. In this section, we give some auxiliary results regarding structures of submodular polyhedra.

4.8.1 Facial structure of $B(f)$

We review some existing results for the characterization of faces of the base polyhedron $B(f)$.

Let F be any face of a polyhedron C determined by a linear inequality system $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$. The relative interior $\mathrm{relint}(F)$ is a set of points that share the same patterns of the equality constraints. In the case of the base polyhedron, such patterns correspond to sublattices of 2^V. In fact, for any $\boldsymbol{x} \in B(f)$, the tight sets $\mathcal{D}(\boldsymbol{x}) = \{A \subseteq V : \mathbf{1}_A^\top \boldsymbol{x} = f(A)\}$ becomes a sublattice of 2^V with $\emptyset, V \in \mathcal{D}(\boldsymbol{x})$ (see Lemma 2.5). We use $\mathbb{D} = \{\mathcal{D}(\boldsymbol{x}) : \boldsymbol{x} \in B(f)\}$ to denote a collection of all possible tight sets. Then, there is a one-to-one correspondence between the faces of $B(f)$ and the elements in $\mathbb{D}$.

Lemma 4.19 ((Fujishige 2005), Section 3.3 (d))**.**

(i) Let F be any face of the base polyhedron $B(f)$. There exists a distributive lattice $\mathcal{D} \in \mathbb{D}$ such that

$$F = F(\mathcal{D}) := \{\boldsymbol{x} \in \mathbb{R}^p : \forall A \in \mathcal{D}, \mathbf{1}_A^\top \boldsymbol{x} = f(A) \text{ and } \forall A \notin \mathcal{D}, \mathbf{1}_A^\top \boldsymbol{x} \leq f(A)\}$$

and

$$\mathrm{relint}(F) = \{\boldsymbol{x} \in \mathbb{R}^p : \forall A \in \mathcal{D}, \mathbf{1}_A^\top \boldsymbol{x} = f(A) \text{ and } \forall A \notin \mathcal{D}, \mathbf{1}_A^\top \boldsymbol{x} < f(A)\}.$$

In particular, $F = F(\mathcal{D}(\boldsymbol{s}))$ holds for all $\boldsymbol{x} \in \mathrm{relint}(F)$.

(ii) Let $\mathcal{D}_1, \mathcal{D}_2$ be two sublattices in $\mathbb{D}$. Then, $\mathcal{D}_1 \subseteq \mathcal{D}_2$ holds if and only if $F(\mathcal{D}_2) \subseteq F(\mathcal{D}_1)$.

(iii) Suppose that F is written as $F(\mathcal{D})$ for some $\mathcal{D} \in \mathbb{D}$. Then, we have $\dim(F) = p - |\Pi(\mathcal{D})|$.

Recall that $L(\Pi)$ is the linear subspace of partition-wise constant vectors on Π. Consider the orthogonal projection of any vector $\boldsymbol{z}$ in $F(\mathcal{D})$ onto the linear subspace $L(\Pi(\mathcal{D}))$. The following lemma shows that there are finitely many projected vectors defined in this way.

Lemma 4.20. Let F be any face of $B(f)$, and $\mathcal{D}$ be the corresponding distributive lattice in Lemma 4.19. Let $\Pi = \Pi(\mathcal{D})$ be the partition defined by Birkhoff's representation theorem. Then, $\boldsymbol{P}_{L(\Pi)}(\boldsymbol{z})$ does not depend on the choice of $\boldsymbol{z} \in F$.

Proof. Let $\boldsymbol{z} \in F$ be any point in the face $F = F(\mathcal{D})$. Let $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_k = V$ be any maximal chain of $\mathcal{D}$. Noting that $\Pi = \{S_i - S_{i-1} : i \in \{1, \ldots k\}\}$, the orthogonal projection map onto $L(\Pi)$ can be written as

$$\boldsymbol{P}_{L(\Pi)}\boldsymbol{z} = \sum_{i=1}^{k} \frac{(\mathbf{1}_{S_i} - \mathbf{1}_{S_{i-1}})^\top \boldsymbol{z}}{|S_i| - |S_{i-1}|}(\mathbf{1}_{S_i} - \mathbf{1}_{S_{i-1}}). \tag{4.26}$$

From Lemma 4.19 (i), we have $(\mathbf{1}_{S_i} - \mathbf{1}_{S_{i-1}})^\top \boldsymbol{z} = f(S_i) - f(S_{i-1})$ for all $i \in \{1, \dots, k\}$. Substituting this into (4.26), we have

$$\boldsymbol{P}_{L(\Pi)}\boldsymbol{z} = \sum_{i=1}^{k} \frac{f(S_i) - f(S_{i-1})}{|S_i| - |S_{i-1}|}(\mathbf{1}_{S_i} - \mathbf{1}_{S_{i-1}}), \tag{4.27}$$

which is independent of the choice of $\boldsymbol{z} \in F$. □

4.8.2 Structure of $|P|(f)$

We study the structure of faces and normal cones of the symmetric submodular polyhedron $|P|(f)$. We begin with a discussion on the facial structure of the submodular polyhedron $P(f)$. Since we have already considered the base polyhedron, we are mainly interested in the faces that are not contained in $B(f)$.

Let $\boldsymbol{s} \in P(f)$ be a subbase, and let $\hat{A}$ be the unique maximal element in $\mathcal{D}(\boldsymbol{s})$. If $\boldsymbol{s} \notin B(f)$, sublattice $\mathcal{D}(\boldsymbol{s})$ does not contain V. In this case, $\hat{A}$ is a proper subset of V. Below, we show the following fact: by splitting the coordinate into $\hat{A}$ and $V - \hat{A}$, the face containing $\boldsymbol{s}$ can be written as a product of two polyhedra in $\mathbb{R}^{\hat{A}}$ and $\mathbb{R}^{V-\hat{A}}$.

For a submodular function $f : 2^V \to \mathbb{R}$, we define its reduction $f^{\hat{A}} : 2^{\hat{A}} \to \mathbb{R}$ as $f^{\hat{A}}(B) := f(B)$ for all $B \subseteq \hat{A}$. We also define the contraction $f_{\hat{A}} : 2^{V-\hat{A}} \to \mathbb{R}$ of f as $f_{\hat{A}}(B) = f(B \cup \hat{A}) - f(\hat{A})$ for all $B \subseteq V - \hat{A}$. Clearly, if $\boldsymbol{s} \in P(f)$, the subvector $\boldsymbol{s}_{\hat{A}}$ is contained in $B(f^{\hat{A}})$. We also have that $\boldsymbol{s}_{V-\hat{A}} \in P(f_{\hat{A}})^\circ$. In fact, for any nonempty $B \subseteq V - \hat{A}$, we have $(\mathbf{1}_{B\cup\hat{A}})^\top \boldsymbol{s} < f(B \cup \hat{A})$ by definition of $\hat{A}$, which implies

$$\mathbf{1}_B^\top \boldsymbol{s} = (\mathbf{1}_{B\cup\hat{A}} - \mathbf{1}_{\hat{A}})^\top \boldsymbol{s} < f(B \cup \hat{A}) - f(\hat{A}) = f_{\hat{A}}(B).$$

Based on the above discussion, we have a representation of normal cones of $P(f)$. The following lemma will be used in the proof of Lemma 4.12.

Lemma 4.21. Let $f : 2^V \to \mathbb{R}$ be a submodular function. Let $\boldsymbol{s} \in P(f)$ be a subbase and let $\hat{A}$ be the unique maximal element in $\mathcal{D}(\boldsymbol{s})$. Then, we have $\boldsymbol{s}_{\hat{A}} \in B(f^{\hat{A}})$ and $\boldsymbol{s}_{V-\hat{A}} \in P(f_{\hat{A}})^\circ$. Furthermore, let $\hat{\Pi}$ be the partition of $\hat{A}$ that is obtained from Birkhoff's representation theorem. Then, the normal cone of $P(f)$ at $\boldsymbol{s}$ is given by

$$N_{P(f)}(\boldsymbol{s}) = N_{B(f^{\hat{A}})}(\boldsymbol{s}_{\hat{A}}) \times N_{P(f_{\hat{A}})}(\boldsymbol{s}_{V-\hat{A}}) = N_{B(f^{\hat{A}})}(\boldsymbol{s}_{\hat{A}}) \times \{\mathbf{0}_{V-\hat{A}}\},$$

which is contained in a linear subspace $\mathrm{span}\{\mathbf{1}_S : S \in \hat{\Pi}\}$.

Proof. This lemma follows from the following facts: (a) a normal vector of a product polyhedron is given by a direct product of normal vectors (see Rockafeller and Wets 1998,Proposition 6.41), and (b) $N_{P(f_{V-\hat{A}})}(\boldsymbol{s}_{V-\hat{A}}) = \{\mathbf{0}_{V-\hat{A}}\}$. The last assertion for the partition-wise constant property can be checked with a similar argument as in Section 4.4.2. □

Here, we assume that f is monotone non-decreasing. The following lemma provides the structure of the symmetric submodular polyhedron.

Lemma 4.22. Let $f : 2^V \to \mathbb{R}$ be a non-decreasing submodular function and F be an arbitrary face of $|P|(f)$.

(i) Let $\boldsymbol{s}$ be an arbitrary point in relint(F). The unique maximal element $\hat{A}$ of $\mathcal{D}(|\boldsymbol{s}|)$ does not depend on the choice of $\boldsymbol{s} \in$ relint(F). Let $F^{\hat{A}}$ be the face of the base polyhedron of reduction $B(f^{\hat{A}})$ that contains $|\boldsymbol{s}|$ in its relative interior, which is also independent of the choice of $\boldsymbol{s} \in \mathrm{relint}(F)$. Then, the original face F can be represented as

$$F = \boldsymbol{\gamma}(\boldsymbol{s}) \odot [F^{\hat{A}} \times |P|(f_{\hat{A}})],$$

where $|P|(f_{\hat{A}}) \subseteq \mathbb{R}^{V-\hat{A}}$ is the symmetric submodular polyhedron for the contraction $f_{\hat{A}}$.

(ii) Denote the orthogonal projection matrix onto $L_0(V - \hat{A}, \Pi(\mathcal{D}(|\boldsymbol{s}|)), \boldsymbol{\gamma}(\boldsymbol{s}))$ as $\boldsymbol{P}$. Then, a vector $\boldsymbol{P}\boldsymbol{s}'$ does not depend on the choice of $\boldsymbol{s}' \in F$.

4.9 Proofs for Section 4.4

In this section, we provide proofs for our main results. The most important parts of our proofs are described in Section 4.9.2, in which we prove the local invariance results of the constant partitions.

Proof of projection lemmas. Here, we prove Lemma 4.3 that gives the anti-projection representation of submodular regularization estimators. Since the proof is the same for the first and the second assertion, we will use common symbols $\mathcal{P}$ for submodular polyhedra and $\hat{\boldsymbol{\theta}}$ for estimators, namely, $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}_{\mathrm{LERE}}$ if $\mathcal{P} = B(f)$ and $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}_{\mathrm{SNRE}}$ if $\mathcal{P} = |P|(f)$.

Define a polyhedron C by $C := \{\boldsymbol{z} \in \mathbb{R}^n : \lambda^{-1}\boldsymbol{X}^\top \boldsymbol{z} \in \mathcal{P}\}$. From the first-order optimality condition, we have

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \in \lambda \partial \Omega_{\mathcal{P}}(\hat{\boldsymbol{\theta}}) = \lambda \operatorname*{argmax}_{\boldsymbol{z} \in \mathcal{P}} \boldsymbol{z}^\top \hat{\boldsymbol{\theta}}. \tag{4.28}$$

By Lemma 2.3, this is equivalent to $\hat{\boldsymbol{\theta}} \in N_{\mathcal{P}}(\lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}))$, which proves the latter assertion.

Next, we want to show that $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \mathrm{Proj}_C(\boldsymbol{y})$. Since (4.28) particularly says that $\lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \in \mathcal{P}$, we have $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} \in C$. Then, for any $\boldsymbol{w} \in C$, we have

$$\langle \boldsymbol{X}\hat{\boldsymbol{\theta}}, \boldsymbol{w} \rangle - \langle \boldsymbol{X}\hat{\boldsymbol{\theta}}, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} \rangle = \langle \hat{\boldsymbol{\theta}}, \boldsymbol{X}^\top \boldsymbol{w} \rangle - \max_{\boldsymbol{z} \in \lambda\mathcal{P}} \langle \hat{\boldsymbol{\theta}}, \boldsymbol{z} \rangle \leq 0,$$

which implies that $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ is a normal vector of C at $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}$, hence the result. □

Proof of Lemma 4.12. Recall that we want to show that

$$\hat{\boldsymbol{\theta}} \in L := L_0(Z_0(\boldsymbol{y}), \Pi_{\mathrm{bound},0}(\boldsymbol{y})\gamma_{\mathrm{bound},0}(\boldsymbol{y})),$$

where $\hat{\boldsymbol{\theta}}$ is an arbitrary SNRE solution and L is a subspace determined by the sparse boundary lattice (see Section 4.4.2 for a precise definition). For notation simplicity, we will write $Z := Z_0(\boldsymbol{y})$, $\Pi := \Pi_{\mathrm{bound},0}(\boldsymbol{y})$, and $\boldsymbol{\gamma} := \boldsymbol{\gamma}_{\mathrm{bound},0}(\boldsymbol{y})$. Since $\hat{\boldsymbol{\theta}}$ is a normal vector of $|P|(f)$ at $\hat{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$, it is sufficient to show that $N_{|P(f)|}(\hat{\boldsymbol{s}})$ is included in $L = \boldsymbol{\gamma} \odot \mathrm{span}\{\boldsymbol{1}_A : A \in \Pi\}$.

Let $\hat{A} = V - Z$ be the maximal element of $\mathcal{D}(|\hat{\boldsymbol{s}}|)$. By Lemma 4.21, $|\hat{\boldsymbol{s}}| = \boldsymbol{\gamma}^\top \hat{\boldsymbol{s}}$ is contained in $B(f^{\hat{A}}) \times P(f_{\hat{A}})^\circ$. The normal cone of $P(f)$ is given as $N_{P(f)}(|\hat{\boldsymbol{s}}|) = N_{B(f^{\hat{A}})} \times \{\boldsymbol{0}_{V-\hat{A}}\}$, which is clearly included in L. However, in general, the normal cone of $|P|(f)$ at $\hat{\boldsymbol{s}}$ does not coincide with $\boldsymbol{\gamma} \odot N_{P(f)}(|\hat{\boldsymbol{s}}|)$. These two sets can be different if there exists an index $i \in \hat{A}$ such that $\hat{s}_i = 0$. Below, we divide the case into (i) $I_0 := \{i \in \hat{A} : \hat{s}_i = 0\} = \emptyset$ and (ii) $I_0 \neq \emptyset$.

First, we assume $I_0 = \emptyset$. In this case, there exists a unique sign vector $\boldsymbol{\gamma}' \in \{-1, 1\}^{\hat{A}}$ such that $\hat{\boldsymbol{s}}_{\hat{A}} \in \boldsymbol{\gamma}' \odot B(f^{\hat{A}})$. In fact, such vector is obtained as $\boldsymbol{\gamma}' = \boldsymbol{\gamma}_{\hat{A}}$. Then, we have the equality $N_{|P|(f)}(\hat{\boldsymbol{s}}) = \boldsymbol{\gamma} \odot N_{P(f)}(|\hat{\boldsymbol{s}}|)$, and hence $\hat{\boldsymbol{\theta}} \in L$.

Next, we consider the case where $I_0 \neq \emptyset$. In this case, there are $2^{|I_0|}$ sign vectors $\boldsymbol{\gamma}'$ satisfying $\hat{\boldsymbol{s}}_{\hat{A}} \in \boldsymbol{\gamma}' \odot B(f^{\hat{A}})$. Hence, we have a set inclusion $N_{|P|(f)}(\hat{\boldsymbol{s}}) \supseteq \gamma \odot N_{P(f)}(|\hat{\boldsymbol{s}}|)$, while the opposite inclusion ($\subseteq$) does not hold in general. However, we can show that the smallest linear subspace containing $N_{|P|(f)}(\hat{\boldsymbol{s}})$ is $L = \boldsymbol{\gamma} \odot \mathrm{span}\{\mathbf{1}_A : A \in \Pi\}$. Fix any index $i \in I_0$ and a set $B \in \mathcal{D}(|\hat{\boldsymbol{s}}|)$ containing i. Then, there exist vectors $\boldsymbol{a}_+, \boldsymbol{a}_- \in \{-1, 0, 1\}^p$ satisfying

$$(a_+)_i = 1, \quad (a_-)_i = -1,$$
$$(a_+)_j = (a_-)_j = \begin{cases} \mathrm{sign}(\hat{s}_j), & j \in B - I_0 \\ -1 \text{ or } 1, & j \in I_0 - \{i\} \\ 0, & j \notin B \end{cases},$$

and $(\boldsymbol{a}_+)^\top \hat{\boldsymbol{s}} = (\boldsymbol{a}_-)^\top \hat{\boldsymbol{s}} = f(B)$. We can see that the sum of these vectors $\boldsymbol{a}_+ + \boldsymbol{a}_-$ is a normal vector of $|P|(f)$ with support $B - \{i\}$. On the other hand, the set $B - \{i\}$ is tight at $|\hat{\boldsymbol{s}}|$ (i.e., $B - \{i\} \in \mathcal{D}(|\hat{\boldsymbol{s}}|)$) because

$$(\mathbf{1}_{B-\{i\}})^\top \hat{\boldsymbol{s}} \le f(B - \{i\}) \le f(B) = \mathbf{1}_B^\top \hat{\boldsymbol{s}} = (\mathbf{1}_{B-\{i\}})^\top \hat{\boldsymbol{s}}. \tag{4.29}$$

Note that the second inequality in (4.29) follows by monotonicity of f. Repeating the above discussion, we can conclude that $\mathcal{D}(|\hat{\boldsymbol{s}}|)$ contains all subsets of B obtained as $B - \mathcal{I}$ for some $\mathcal{I} \subseteq I$. Therefore, we have

$$N_{|P|(f)}(\hat{\boldsymbol{s}}) \subseteq \boldsymbol{\gamma} \odot \mathrm{span}\{\mathbf{1}_B : B \in \mathcal{D}(|\hat{\boldsymbol{s}}|)\} = \boldsymbol{\gamma} \odot \mathrm{span}\{\mathbf{1}_A : A \in \Pi\} = L,$$

which is the desired result. □

Remark 4.23. If f is strictly increasing (i.e., $A \subset B \Rightarrow f(A) < f(B)$), then $I_0 = \emptyset$ always holds in the above proof. In fact, assuming $I_0 \neq \emptyset$, we have (4.29) that contradicts the strict monotonicity. For example, the submodular functions associated with the lasso and the SLOPE are strictly increasing, while that of the ℓ_1/ℓ_∞-group lasso is not.

4.9.1 Local invariance of boundary lattices

We prove local invariance of the boundary lattice (Lemma 4.9) and the sparse boundary lattice (Lemma 4.13). Both lemmas are derived mainly from the following property of inverse images of polyhedra.

Lemma 4.24. Let $C \subseteq \mathbb{R}^p$ be a polyhedron and $T : \mathbb{R}^n \to \mathbb{R}^p$ be a linear map.

(i) The inverse image $T^{-1}(C)$ is a polyhedron in $\mathbb{R}^n$.

(ii) Let F be a nonempty face of $T^{-1}(C)$. Then, there exists a face G of C such that $T(\mathrm{relint}(F)) \subseteq \mathrm{relint}(G)$.

Proof. The first assertion is a well-known result (Rockafellar 1970,Theorem 19.3).

We will prove the second assertion. Since $T(\mathrm{relint}(F)) = \mathrm{relint}(T(F))$ (Rockafellar 1970,Theorem 6.6), $T(\mathrm{relint}(F))$ is relatively open in $\mathbb{R}^p$. Note that a set is relatively open (in the convex analysis sense) if it is open with respect to the relative topology induced by its affine hull. Using the fact that any relatively open set in a polyhedron C is contained in a single face G (Rockafellar 1970,Theorem 18.2), we have the desired result. □

According to the above lemma, we can show that the face of the submodular polyhedron that contains $\hat{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ is locally invariant, which implies local invariance of the boundary lattices. The followings are detailed discussions.

Proof of Lemma 4.9. By Lemma 4.3, we have $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\theta}} = \mathrm{Proj}_C(y)$, where $C = (X^\top)^{-1}(\lambda B(f))$. There is a measure zero set $\mathcal{M}_1 \subset \mathbb{R}^n$ such that $\boldsymbol{y} \mapsto \mathrm{Proj}_C(\boldsymbol{y})$ is a locally affine on $\mathbb{R}^n - \mathcal{M}_1$ (Tibshirani and Taylor 2012,Lemma 2). In particular, there exist a neighborhood U of $\boldsymbol{y}$ and a face F of C such that $\mathrm{Proj}_C(\boldsymbol{y}') \in \mathrm{relint}(F)$ for all $\boldsymbol{y}' \in U$. By Lemma 4.24, there is a unique face G of $B(f)$ that contains $\hat{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{X}^\top\mathrm{Proj}_C(\boldsymbol{y})$ in its relative interior. Thus, the boundary lattice $\mathcal{D}(\hat{\boldsymbol{s}})$ is invariant on U by Lemma 4.19. □

Proof of Lemma 4.13. Except for a measure zero set $\mathcal{M}_3$, $\hat{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ is contained in the relative interior of a single face of $|P|(f)$, or in the p-dimensional interior of $|P|(f)$ itself. Here, for the sake of simplicity, we assume that f is strictly increasing. Then, there exists a neighborhood U where $\mathcal{D}(|\hat{\boldsymbol{s}}|)$ and $\mathrm{sign}(\hat{\boldsymbol{s}})_{\hat{A}}$ are invariant. □

4.9.2 Local invariance of constant partitions

Here, we prove the local invariance results for the constant partition (Lemma 4.10) and the sparse constant partition (Lemma 4.14).

Proof of Lemma 4.10. Step 1. First, we establish a locally affine representation of $\boldsymbol{X}\hat{\boldsymbol{\theta}}$. For any $\boldsymbol{y} \in \mathbb{R}^n$, let $\Pi_{\mathrm{bound}} = \Pi_{\mathrm{bound}}(\boldsymbol{y})$ be the partition associated with the boundary lattice. Also, let $\hat{\boldsymbol{\theta}}$ be an arbitrary LERE solution, and denote $\Pi_{\mathrm{const}} = \Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$. We define two orthogonal projection matrices $\boldsymbol{P}_1 := \boldsymbol{P}_{L(\hat{\Pi}_{\mathrm{bound}})}$ and $\boldsymbol{P}_2 := \boldsymbol{P}_{L(\Pi_{\mathrm{const}})}$.

Below, let ℓ denote either 1 or 2. Combining $\lambda\hat{\boldsymbol{s}} = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ and $\boldsymbol{P}_\ell\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$, we have

$$\boldsymbol{P}_\ell\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{P}_\ell\hat{\boldsymbol{\theta}} = \boldsymbol{P}_\ell\boldsymbol{X}^\top\boldsymbol{y} - \lambda\boldsymbol{P}_\ell\hat{\boldsymbol{s}}.$$

By multiplying on the left by $(\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+$, we have

$$\text{L.H.S.} = (\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{P}_\ell\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+\boldsymbol{X}\boldsymbol{P}_\ell\hat{\boldsymbol{\theta}} = \boldsymbol{X}\boldsymbol{P}_\ell\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\theta}},$$

and

$$\text{R.H.S.} = (\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\boldsymbol{X}^\top\boldsymbol{y} - \lambda(\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\hat{\boldsymbol{s}} = (\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\boldsymbol{X}^\top\{\boldsymbol{y} - \lambda(\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\hat{\boldsymbol{s}}\}$$

Combining these, we have

$$\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+\{\boldsymbol{y} - \lambda(\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\hat{\boldsymbol{s}}\}.$$

Here, we choose an extremal point $\hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}(\boldsymbol{y})$ of $B(f)$ such that $\boldsymbol{P}_\ell\hat{\boldsymbol{z}} = \boldsymbol{P}_\ell\hat{\boldsymbol{s}}$. This can be done by the following construction. Let F be the minimal face that contains $\hat{\boldsymbol{s}}$ in its relative interior. We define $\hat{\boldsymbol{z}}$ as an extremal point in F. If there are more than two extremal points in F, we can choose one by any well-ordering defined on the set of extremal points in $B(f)$. Then, we immediately have $\boldsymbol{P}_1\hat{\boldsymbol{s}} = \boldsymbol{P}_1\hat{\boldsymbol{z}}$ from Lemma 4.20. Next, we will show that $\boldsymbol{P}_2\hat{\boldsymbol{s}} = \boldsymbol{P}_2\hat{\boldsymbol{z}}$. Let $F(\hat{\boldsymbol{\theta}})$ be a face defined as $\mathrm{argmax}_{\boldsymbol{z}\in B(f)}\, \boldsymbol{z}^\top\hat{\boldsymbol{\theta}}$. From the first order optimality condition, we have $\hat{\boldsymbol{s}} \in F(\hat{\boldsymbol{\theta}})$, and hence $F \subseteq F(\hat{\boldsymbol{\theta}})$. Let $A_1, \ldots, A_k$ be elements of Π_{const} that are sorted in the ascending order determined by values in $\hat{\boldsymbol{\theta}}$. Lemma 4.19-2 implies that the lattice that corresponds to $F(\hat{\boldsymbol{\theta}})$ is a sublattice of that of F (say $\mathcal{D}$). Therefore, there is a chain $\emptyset = S_0 \subset \cdots \subset S_k = V$ in $\mathcal{D}$ such that $A_i = S_i - S_{i-1}$,

$i \in \{1, \dots, k\}$. In particular, we can check that $\mathbf{1}_{A_i}^\top \hat{\boldsymbol{s}} = \mathbf{1}_{A_i}^\top \hat{\boldsymbol{z}} = f(S_i) - f(S_{i-1})$. By a similar calculation as the proof of Lemma 4.20, we conclude

$$\boldsymbol{P}_2 \hat{\boldsymbol{s}} = \boldsymbol{P}_2 \hat{\boldsymbol{z}} = \sum_{i=1}^{k} \frac{f(A_1 \cup \cdots \cup A_i) - f(A_1 \cup \cdots \cup A_{i-1})}{|A_i|} \mathbf{1}_{A_i}.$$

Consequently, we have

$$\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+ \{\boldsymbol{y} - \lambda(\boldsymbol{P}_\ell \boldsymbol{X}^\top)^+ \boldsymbol{P}_\ell \hat{\boldsymbol{z}}\}. \tag{4.30}$$

Furthermore, we can represent any solution $\hat{\boldsymbol{\theta}}$ in (4.3) as

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)^+ \{\boldsymbol{y} - \lambda(\boldsymbol{P}_\ell \boldsymbol{X}^\top)^+ \boldsymbol{P}_\ell \hat{\boldsymbol{z}}\} + \boldsymbol{b}_\ell, \tag{4.31}$$

where $\boldsymbol{b}_\ell$ is a vector in $\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_\ell)$.
Step 2. Next, we will prove the local invariance of the constant partition. More precisely, we want to construct a sufficiently small neighborhood U so that every $\boldsymbol{y}' \in U$ has a LERE solution $\hat{\boldsymbol{\eta}}$ with the same constant partition as $\hat{\boldsymbol{\theta}}$. Let $\hat{\boldsymbol{\eta}} : U \to \mathbb{R}^p$ be a map of the form $\hat{\boldsymbol{\eta}}(\boldsymbol{y}') = (\boldsymbol{X}\boldsymbol{P}_2)^+ \{\boldsymbol{y}' - \lambda(\boldsymbol{P}_2 \boldsymbol{X}^\top)^+ \boldsymbol{P}_2 \hat{\boldsymbol{z}}\} + \boldsymbol{b}'(\boldsymbol{y}')$. Here, $\boldsymbol{b}'(\boldsymbol{y}')$ is contained in $\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$ for all $\boldsymbol{y}' \in U$. We will write $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(\boldsymbol{y}')$ and $\boldsymbol{b}' = \boldsymbol{b}'(\boldsymbol{y}')$ by omitting the dependence on $\boldsymbol{y}'$. We will prove a more detailed version of Lemma 4.10 as follows.

Lemma 4.25. There is a measure-zero set $\mathcal{M}_2 \subset \mathbb{R}^n$ with the following property. For any $\boldsymbol{y} \notin \mathcal{M}_2$, fix an arbitrary LERE solution $\hat{\boldsymbol{\theta}}$. Then, there are a neighborhood U of $\boldsymbol{y}$ and a map $\boldsymbol{b}' : U \to \mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$ satisfying the following conditions:

(i) For any $\boldsymbol{y}' \in U$, the constant partition of $\hat{\boldsymbol{\eta}}$ equals to that of $\hat{\boldsymbol{\theta}}$. Moreover, for any A_1, A_2 in $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$, $\hat{\boldsymbol{\theta}}_{A_1} > \hat{\boldsymbol{\theta}}_{A_2}$ implies $\hat{\boldsymbol{\eta}}_{A_1} > \hat{\boldsymbol{\eta}}_{A_2}$.
(ii) For any $\boldsymbol{y}' \in U$, $\hat{\boldsymbol{\eta}}$ is an optimal solution in (4.3).

First, we show that $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\eta}})$ is a cover of $\Pi_{\mathrm{const}}(\hat{\boldsymbol{\theta}})$. Note that we can rewrite this as $\boldsymbol{P}_2 \hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}$, or equivalently

$$(\boldsymbol{I}_p - \boldsymbol{P}_2)[(\boldsymbol{X}\boldsymbol{P}_2)^+ \{\boldsymbol{y}' - \lambda(\boldsymbol{P}_2 \boldsymbol{X}^\top)^+ \boldsymbol{P}_2 \hat{\boldsymbol{z}}\} + \boldsymbol{b}'] = \mathbf{0}. \tag{4.32}$$

Interpreting (4.32) as a linear equation of $\boldsymbol{b}_2' \in \mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$, we consider the solvability. Define a set $\mathcal{M}_2 \subset \mathbb{R}^n$ as

$$\begin{aligned}\mathcal{M}_2 := \bigcup_{\mathcal{D}, \Pi, \boldsymbol{z}} \Big\{ \boldsymbol{w} \in \mathbb{R}^n \ : \ & \boldsymbol{P}_{\{(\boldsymbol{I}_p - \boldsymbol{P}_{L(\Pi)}) \mathrm{null}(\boldsymbol{X}\boldsymbol{P}_{L(\Pi)})\}^\perp} \times \\ & (\boldsymbol{I}_p - \boldsymbol{P}_{L(\Pi)})(\boldsymbol{X}\boldsymbol{P}_{L(\Pi)})^+ \{\boldsymbol{w} - \lambda(\boldsymbol{P}_{L(\Pi)} \boldsymbol{X}^\top)^+ \boldsymbol{P}_{L(\Pi)} \boldsymbol{z}\} = \mathbf{0} \Big\}.\end{aligned} \tag{4.33}$$

Here, the union in the right-hand side is taken over the triple $(\mathcal{D}, \Pi, \boldsymbol{z})$ specified as follows:

- $\mathcal{D} \in \mathbb{D}$ is a distributive lattice that determines a face of the base polyhedron.
- Π is a cover of partition $\Pi(\mathcal{D})$ such that the matrix $\boldsymbol{P}_{\{(\boldsymbol{I}_p - \boldsymbol{P}_{L(\Pi)}) \mathrm{null}(\boldsymbol{X}\boldsymbol{P}_{L(\Pi)})\}^\perp} (\boldsymbol{I}_p - \boldsymbol{P}_{L(\Pi)})(\boldsymbol{X}\boldsymbol{P}_{L(\Pi)})^+$ is not identical to the zero matrix $\boldsymbol{O}_{p \times n}$.
- $\boldsymbol{z}$ is a vertex of the base polyhedron.

It is clear that there are finitely many possible patterns of such triples. Thus, $\mathcal{M}_2$ becomes a measure zero set because it is a finite union of linear subspaces whose codimension is more than 1.

Now, we assume that $\boldsymbol{y} \in \mathbb{R}^n - \mathcal{M}_2$. Combining $\boldsymbol{P}_2\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ with (4.31), we have

$$(\boldsymbol{I}_p - \boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+\{\boldsymbol{y} - \lambda(\boldsymbol{P}_2\boldsymbol{X}^\top)^+\boldsymbol{P}_2\hat{\boldsymbol{z}}\} = -(\boldsymbol{I}_p - \boldsymbol{P}_2)\boldsymbol{b}_2.$$

Noting that $(\boldsymbol{I}_p - \boldsymbol{P}_2)\boldsymbol{b}_2 \in (\boldsymbol{I}_p - \boldsymbol{P}_2)\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$, we have

$$\boldsymbol{P}_{\{(\boldsymbol{I}_p - \boldsymbol{P}_2)\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)\}^\perp}(\boldsymbol{I}_p - \boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+\{\boldsymbol{y} - \lambda(\boldsymbol{P}_2\boldsymbol{X}^\top)^+\boldsymbol{P}_2\hat{\boldsymbol{z}}\} = \boldsymbol{0}.$$

However, since $\boldsymbol{y} \notin \mathcal{M}$, this implies that the matrix $\boldsymbol{P}_{\{(\boldsymbol{I}_p - \boldsymbol{P}_{L(\Pi)})\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_{L(\Pi)})\}^\perp}(\boldsymbol{I}_p - \boldsymbol{P}_{L(\Pi)})(\boldsymbol{X}\boldsymbol{P}_{L(\Pi)})^+$ must be $\boldsymbol{O}_{p,n}$. Thus, we conclude that $\mathrm{col}\{(\boldsymbol{I}_p - \boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+\} \subseteq (\boldsymbol{I}_p - \boldsymbol{P}_2)\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$. This set inclusion guarantees that, for any $\boldsymbol{y}' \in \mathbb{R}^n$, there exists $\boldsymbol{b}' \in \mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$ that solves (4.32).

Second, we want to find a sufficiently small neighborhood $\bar{U}$ so that the order preservation condition holds:

$$\hat{\boldsymbol{\theta}}_{A_1} > \hat{\boldsymbol{\theta}}_{A_2} \Rightarrow \hat{\boldsymbol{\eta}}_{A_1} > \hat{\boldsymbol{\eta}}_{A_2} \quad \text{for all } A_1, A_2 \in \Pi_{\mathrm{const}}. \tag{4.34}$$

Let $a_1 > \cdots > a_k$ be the distinct values in $\hat{\boldsymbol{\theta}}$ in the descending order. Let ε denote the smallest jump $\max_{1 \le j \le k-1} a_j - a_{j+1}$. If we prove that $\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\eta}}\|_\infty \le \|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\eta}}\|_2 < \varepsilon/2$, then the assertion (4.34) follows. By the triangle inequality, we have

$$\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\eta}}\|_2 \le \|(\boldsymbol{X}\boldsymbol{P}_2)^+(\boldsymbol{y} - \boldsymbol{y}')\|_2 + \|\boldsymbol{b}_2 - \boldsymbol{b}'\|_2.$$

By the continuity of $\boldsymbol{y} \mapsto (\boldsymbol{X}\boldsymbol{P}_2)^+\boldsymbol{y}$, we can bound the first term in the right-hand side by $\varepsilon/4$ if we choose a neighborhood U_1 small enough. Next, we provide a bound on the second term. Note that $\boldsymbol{I}_p - \boldsymbol{P}_2$ is a symmetric matrix, and that $\boldsymbol{z} \mapsto (\boldsymbol{I}_p - \boldsymbol{P}_2)\boldsymbol{z}$ is bijective if it is regarded as a map from $\mathrm{row}(\boldsymbol{I}_p - \boldsymbol{P}_2)$ to $\mathrm{col}(\boldsymbol{I}_p - \boldsymbol{P}_2)$. Denote by $(\boldsymbol{I}_p - \boldsymbol{P}_2)^{-1}_{\mathrm{row}}$ the (restricted) inverse. By the bounded inverse theorem, there exists $M > 0$ such that $\|(\boldsymbol{I}_p - \boldsymbol{P}_2)^{-1}_{\mathrm{row}}\|_{\mathrm{op}} \le M$. Hence, we have

$$\|\boldsymbol{b}_2 - \boldsymbol{b}'\|_2 \le M\|(\boldsymbol{I}_p - \boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+(\boldsymbol{y} - \boldsymbol{y}')\|_2.$$

By continuity, we can choose a small neighborhood U_2 in which $\|\boldsymbol{b}_2 - \boldsymbol{b}'\|_2 < \varepsilon/4$. Therefore, the order preserving condition (4.34) is satisfied whenever $\boldsymbol{y}' \in \bar{U} := U_1 \cap U_2$, which proves (i) of Lemma 4.25.

Next, we prove (ii). We will verify the optimality condition

$$\boldsymbol{X}^\top(\boldsymbol{y}' - \boldsymbol{X}\hat{\boldsymbol{\eta}}) \in \lambda \operatorname*{argmax}_{\boldsymbol{z} \in B(f)} \boldsymbol{z}^\top \hat{\boldsymbol{\eta}}. \tag{4.35}$$

From (i) and the basic property of the Lovász extension (Section 2.2.2), we can see that $\hat{\boldsymbol{\eta}}^\top \boldsymbol{P}_2\hat{\boldsymbol{z}} = \hat{f}(\hat{\boldsymbol{\eta}}) = \max_{\boldsymbol{z} \in B(f)} \boldsymbol{z}^\top \hat{\boldsymbol{\eta}}$. Then, we have

$$\begin{aligned}
\hat{\boldsymbol{\eta}}^\top \boldsymbol{X}^\top(\boldsymbol{y}' - \boldsymbol{X}\hat{\boldsymbol{\eta}}) &= \hat{\boldsymbol{\eta}}^\top(\boldsymbol{X}\boldsymbol{P}_2)^\top(\boldsymbol{y}' - \boldsymbol{X}\boldsymbol{P}_2\hat{\boldsymbol{\eta}}) \\
&= \hat{\boldsymbol{\eta}}^\top\{(\boldsymbol{X}\boldsymbol{P}_2)^\top(\boldsymbol{X}\boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+\boldsymbol{y}' - (\boldsymbol{X}\boldsymbol{P}_2)^\top(\boldsymbol{X}\boldsymbol{P}_2)\hat{\boldsymbol{\eta}}\} \\
&= \lambda\hat{\boldsymbol{\eta}}^\top(\boldsymbol{X}\boldsymbol{P}_2)^+(\boldsymbol{X}\boldsymbol{P}_2)\boldsymbol{P}_2\hat{\boldsymbol{z}} \\
&\overset{(\star)}{=} \lambda\hat{\boldsymbol{\eta}}^\top\boldsymbol{P}_2\hat{\boldsymbol{z}} = \lambda \max_{\boldsymbol{z} \in B(f)} \boldsymbol{z}^\top\hat{\boldsymbol{\eta}},
\end{aligned}$$

which proves the maximality in (4.35). Here, we used the fact that $\boldsymbol{P}_2\hat{\boldsymbol{z}} = \boldsymbol{P}_2\hat{\boldsymbol{s}} \in \mathrm{col}(\boldsymbol{P}_2\boldsymbol{X}^\top)$ to prove equality $(\star)$. It remains to show that $\boldsymbol{X}^\top(\boldsymbol{y}' - \boldsymbol{X}\hat{\boldsymbol{\eta}}) \in \lambda B(f)$. Let

$\hat{\boldsymbol{\theta}}'$ be any optimal solution for $\boldsymbol{y}'$, and let $\hat{\boldsymbol{s}}' := \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}') \in B(f)$. By the local invariance of the boundary lattice (i.e., Lemma 4.9), there is a small neighborhood U_3 such that $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{s}}'$ are both contained in the relative interior of a single face. Then, by a similar derivation as (4.30), we can see that $\boldsymbol{X}^\top(\boldsymbol{y}' - \boldsymbol{X}\hat{\boldsymbol{\eta}}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}') \in B(f)$. This holds true whenever $\boldsymbol{y}' \in U := U_1 \cap U_2 \cap U_3$. Consequently, we have proved Lemma 4.25 and thus Lemma 4.10. □

Proof of Lemma 4.14. To avoid redundancy, we prove Lemma 4.14 by appropriately modifying the lemma for LERE. Let $\hat{\boldsymbol{\theta}}$ be an arbitrary solution of SNRE. Let $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ be projection matrices onto the linear spaces $\mathcal{L}_1$ and $\mathcal{L}_2$ defined in (4.20) and (4.21), respectively. It is clear from the definition that $\boldsymbol{P}_2\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$. From Lemma 4.12, we also have $\boldsymbol{P}_1\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$.

According to an argument similar to Step 1 in the proof of Lemma 4.10, we have the following representation:

$$\boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+\{\boldsymbol{y} - \lambda(\boldsymbol{P}_\ell\boldsymbol{X}^\top)^+\boldsymbol{P}_\ell\hat{\boldsymbol{z}}\}, \quad \ell \in \{1, 2\}. \tag{4.36}$$

Here, $\hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}(\boldsymbol{y})$ is a point of $|P|(f)$ defined as follows. By Lemma 4.22 (i), the minimal face containing $\hat{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ is given by $F = \hat{\gamma}(\boldsymbol{y}) \odot \{F^{\hat{A}} \times |P|(f_{\hat{A}})\}$, where $F^{\hat{A}}$ is the minimal face of $B(f^{\hat{A}})$ containing subvector $\hat{\boldsymbol{s}}_{\hat{A}}$. Then, we define $\hat{\boldsymbol{z}} \in |P|(f)$ by letting $\hat{\boldsymbol{z}}_{\hat{A}}$ be an extremal point in $F^{\hat{A}}$ and $\hat{\boldsymbol{z}}_{V-\hat{A}} = \boldsymbol{0}$. We can see that $\boldsymbol{P}_\ell\hat{\boldsymbol{s}} = \boldsymbol{P}_\ell\hat{\boldsymbol{z}}$ by Lemma 4.22 (ii).

We also have

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}\boldsymbol{P}_2)^+\{\boldsymbol{y} - \lambda(\boldsymbol{P}_2\boldsymbol{X}^\top)^+\boldsymbol{P}_2\hat{\boldsymbol{z}}\} + \boldsymbol{b}_2, \tag{4.37}$$

where $\boldsymbol{b}_2 \in \mathrm{null}(\boldsymbol{X}P_2)$ depends on the choice of the SNRE solution. For solutions of the form (4.37), we prove the optimality condition and the local invariance of the sparse constant partition.

Define a set $\mathcal{M}_4 \subset \mathbb{R}^n$ as

$$\mathcal{M}_4 := \bigcup_{\mathcal{D}, \boldsymbol{\gamma}, \Pi, B, \boldsymbol{z}} \mathcal{M}(\mathcal{D}, \boldsymbol{\gamma}, \Pi, B, \boldsymbol{z}), \tag{4.38}$$

where we write

$$\begin{aligned} &\mathcal{M}(\mathcal{D}, \boldsymbol{\gamma}, \Pi, B, \boldsymbol{z}) \\ &:= \left\{\boldsymbol{w} \in \mathbb{R}^n : \boldsymbol{P}_{\{(\boldsymbol{I}_p - \boldsymbol{P}_\mathcal{L})\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_\mathcal{L})\}^\perp}(\boldsymbol{I}_p - \boldsymbol{P}_\mathcal{L})(\boldsymbol{X}\boldsymbol{P}_\mathcal{L})\{\boldsymbol{w} - \lambda(\boldsymbol{P}_\mathcal{L}\boldsymbol{X}^\top)^+\boldsymbol{P}_\mathcal{L}\boldsymbol{z}\} = \boldsymbol{0}\right\}, \end{aligned} \tag{4.39}$$

$$\mathcal{L} = \mathcal{L}(\mathcal{D}, \boldsymbol{\gamma}, \Pi, B) := L_0(V - (A(\mathcal{D}) \cap B), \Pi, \boldsymbol{\gamma}). \tag{4.40}$$

Here, the union on the right-hand side of (4.38) is taken over quintuple $(\mathcal{D}, \boldsymbol{\gamma}, \Pi, B, \boldsymbol{z})$ specified as follows:

- $\mathcal{D}$ is a sublattice of 2^V that is not identical to singleton $\{\emptyset\}$. Let $A(\mathcal{D})$ be its (nonempty) maximal element.
- $\boldsymbol{\gamma} \in \{-1, 1\}^p$ is a sign vector.
- (Π, B) is a pair of a partition Π and a set $B \subseteq A(\mathcal{D})$. Denote by $\Pi(\mathcal{D})$ a partition of $A(\mathcal{D})$ determined by Birkhoff's representation theorem. Π is a cover of $\Pi(\mathcal{D})$ such that the projection matrix onto the linear space

$$\{(\boldsymbol{I}_p - \boldsymbol{P}_{L_0(V - A(\mathcal{D}) \cap B, \Pi, \boldsymbol{\gamma})})\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_{L_0(V - A(\mathcal{D}) \cap B, \Pi, \boldsymbol{\gamma})})\}^\perp$$

  is not identical to the zero matrix.

- $\boldsymbol{z}$ is an extremal point of $|P|(f)$.

From the finiteness character of the quintuple, we can see that $\mathcal{M}_4$ is a measure zero set.

We comment on the above definition. The exception set $\mathcal{M}_4$ is defined in order to avoid unfavorable choices of $\hat{\boldsymbol{\theta}}$. Given a sublattice $\mathcal{D} = \hat{\mathcal{D}}_0(\boldsymbol{y})$, $\Pi(\mathcal{D})$ is the finest partition in which the non-zero components in $|\hat{\boldsymbol{\theta}}|$ are constant. However, a coarse partition $\Pi = \Pi_{\mathrm{const},0}(\hat{\boldsymbol{\theta}})$ can arise from particular choices of $\hat{\boldsymbol{\theta}}$. In addition, the index set of non-zero components $V - Z(\hat{\boldsymbol{\theta}})$ can be smaller than the maximal choice $A(\mathcal{D})$. By excluding $\mathcal{M}_4$, we ensure that no inconvenience is caused by such solution choices.

Now, we assume that $\boldsymbol{y} \in \mathbb{R}^n - \mathcal{M}_4$. For $\boldsymbol{y}' \in \mathbb{R}^n$, define

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{X}\boldsymbol{P}_2)^+\{\boldsymbol{y}' - \lambda(\boldsymbol{P}_2\boldsymbol{X}^\top)^+\boldsymbol{P}_2\hat{\boldsymbol{z}}\} + \boldsymbol{b}'.$$

A similar argument to the case of the LER yields

$$\boldsymbol{P}_{\{(\boldsymbol{I}_p - \boldsymbol{P}_2)\mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)\}^\perp}(\boldsymbol{I}_p - \boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+ = \boldsymbol{O}_{p,n}.$$

Hence, for any $\boldsymbol{y}'$, we can choose $\boldsymbol{b}' \in \mathrm{null}(\boldsymbol{X}\boldsymbol{P}_2)$ so that $\boldsymbol{P}_2\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}$. By the continuity argument similar to Step 2 in the proof of Lemma 4.10, we can also choose a neighborhood U of $\boldsymbol{y}$ so that the sparse constant partition of $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(\boldsymbol{y}')$ is invariant on it. On the other hand, we can prove that the optimality condition for SNRE

$$\boldsymbol{X}^\top(\boldsymbol{y}' - \boldsymbol{X}\hat{\boldsymbol{\eta}}) \in \lambda \operatorname*{argmax}_{\boldsymbol{z} \in |P|(f)} \hat{\boldsymbol{\eta}}^\top \boldsymbol{z}$$

holds for all $\boldsymbol{y}' \in U$. We have thus proved the desired result. □

Proof of Theorem 4.7 and Theorem 4.11. Here, we prove our main results for the degrees of freedom. To avoid redundancy, we provide a proof only for Theorem 4.7. Theorem 4.11 can be proved in the same way.

Let $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$ be a measure zero set obtained by the union of the exception sets in Lemma 4.9 and Lemma 4.12. Suppose that $\boldsymbol{y} \in \mathbb{R}^n - \mathcal{M}$, and define $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ as in the proof of Lemma 4.10. Then, there exists a neiborhood U of $\boldsymbol{y}$ such that

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\theta}}(\boldsymbol{y}') &= (\boldsymbol{X}\boldsymbol{P}_1)(\boldsymbol{X}\boldsymbol{P}_1)^+\{\boldsymbol{y}' - \lambda(\boldsymbol{P}_1\boldsymbol{X}^\top)^+\boldsymbol{P}_1\hat{\boldsymbol{z}}\}\\ &= (\boldsymbol{X}\boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+\{\boldsymbol{y}' - \lambda(\boldsymbol{P}_2\boldsymbol{X}^\top)^+\boldsymbol{P}_2\hat{\boldsymbol{z}}\}\end{aligned}$$

holds for any $\boldsymbol{y}' \in U$. Since the two affine functions are equal in the open set U, we have $(\boldsymbol{X}\boldsymbol{P}_1)(\boldsymbol{X}\boldsymbol{P}_1)^+ = (\boldsymbol{X}\boldsymbol{P}_2)(\boldsymbol{X}\boldsymbol{P}_2)^+$. Thus, $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ is differentiable at $\boldsymbol{y}$, and the divergence is given as

$$(\nabla \cdot \boldsymbol{X}\hat{\boldsymbol{\theta}})(\boldsymbol{y}) = \mathrm{tr}((\boldsymbol{X}\boldsymbol{P}_\ell)(\boldsymbol{X}\boldsymbol{P}_\ell)^+) = \dim(\boldsymbol{X}\mathrm{col}(\boldsymbol{P}_\ell)).$$

By Stein's lemma, we have the desired result.

□

4.10 Proofs for Section 4.5

Proof of Proposition 4.15. Since the solution of full-rank LERE is unique, we only have to consider the representation based on the boundary lattice $\mathcal{D}_{\mathrm{bound}}(\boldsymbol{y})$.

Let $\boldsymbol{s} \in B(f)$ be a point in the base polyhedron. We can define a directed graph $\mathcal{G}(\boldsymbol{s})$ with vertex set V as follows. Given $\boldsymbol{s} \in B(f)$ and $i \in V$, define the dependence function $\operatorname{dep}(\boldsymbol{s}, i) \subseteq V$ as the unique minimum element of the distributed lattice $\mathcal{D}(\boldsymbol{s}, i) = \{A \subset V : i \in A, \mathbf{1}_A^\top \boldsymbol{s} = f(A)\}$. Intuitively, $\operatorname{dep}(\boldsymbol{s}, i)$ expresses the exchangeability of the directions that $\boldsymbol{s}$ can proceed on the base polyhedron. We define a set of directed edges $E(\boldsymbol{s})$ by $E(\boldsymbol{s}) = \{(i, j) : j \in \operatorname{dep}(\boldsymbol{s}, i)\}$. The resulting graph $\mathcal{G}(\boldsymbol{s}) = (V, E(\boldsymbol{s}))$ is called the exchangeability graph.

Now, we are interested in a cut function of the undirected graph G. From the fundamental result on the duality between cut functions and flows (see (2.65) in (Fujishige 2005)), every point $\boldsymbol{s} \in B(f_{\mathrm{cut}})$ is a boundary of a feasible flow on G. Hence, there exists an edge (i, j) in the exchangeability graph $\mathcal{G}(\boldsymbol{s})$ only if it is a self-loop ($i = j$) or an undirected edge (i, j) is contained in the original graph G. On the other hand, by Lemma 3.41 in (Fujishige 2005), the partition $\Pi(\mathcal{D}(\boldsymbol{s}))$ is given by the set of connected components in $\mathcal{G}(\boldsymbol{s})$. Therefore, every element in the partition determined from the boundary lattice becomes a connected component in the original graph G. □

Proof of Proposition 4.16. Fix an observation vector $\boldsymbol{y} \in \mathbb{R}^n$. If $\lambda > 0$ is taken large enough, the orthogonal projection of $\boldsymbol{y}$ onto $\lambda B(f)$ is equivalent to that onto the conical hull of $B(f)$, i.e., there exists $\lambda_+ \in (0, \infty)$ such that

$$\operatorname{Proj}_{\lambda B(f)}(\boldsymbol{y}) = \operatorname{Proj}_{\operatorname{cone}(B(f))}(\boldsymbol{y}) \tag{4.41}$$

holds for all $\lambda \geq \lambda_+$.

Note that the polar cone of $K := \operatorname{cone}(B(f))$ is given as

$$\begin{aligned} K^\circ &:= \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^\top \boldsymbol{\theta} \leq 0, \ \forall \boldsymbol{\theta} \in \operatorname{cone}(B(f))\} \\ &= \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^\top \boldsymbol{\theta} \leq 0, \ \forall \boldsymbol{\theta} \in B(f)\} \\ &= \{\boldsymbol{z} \in \mathbb{R}^n : \hat{f}(\boldsymbol{z}) \leq 0\}. \end{aligned}$$

For the cut function of DAG $G = (V, E)$, K° coincides with the set of all vectors satisfying $\theta_i \leq \theta_j$ for all $(i, j) \in E$. Using the basic fact that any vector $\boldsymbol{y} \in \mathbb{R}^n$ can be decomposed as $\boldsymbol{y} = \operatorname{Proj}_K(\boldsymbol{y}) + \operatorname{Proj}_{K^\circ}(\boldsymbol{y})$, we have the decomposition $\boldsymbol{y} = \operatorname{Proj}_{\lambda B(f)}(\boldsymbol{y}) + \operatorname{Proj}_{K^\circ}(\boldsymbol{y})$. Combining with (4.41), we conclude that the LERE solution can be written as

$$\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{y} - \operatorname{Proj}_{\lambda B(f)}(\boldsymbol{y}) = \operatorname{Proj}_{K^\circ}(\boldsymbol{y}).$$

Indeed, the right-hand side is the isotonic regression estimator on G. □

Chapter 5

Estimating Piecewise Monotone Signals

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 mainly focus 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. Under a weak assumption, the estimator achieve a nearly minimax convergence rate over certain classes of piecewise monotone signals. We also 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.

This chapter is based on Minami (2019).

5.1 Overview

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 $\theta_1^* \leq \theta_2^* \leq \cdots \leq \theta_n^*$, 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:

$$\text{minimize } \|y - \theta\|_2 \quad \text{subject to } \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n. \tag{5.1}$$

In other words, the isotonic regression is the least squares estimator $\hat{\theta} = \hat{\theta}_{K_n^{\uparrow}}$ over a closed convex cone $K_n^{\uparrow} := \{\theta \in \mathbb{R}^n : \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n\}$. 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 (2017) and references therein.

In this chapter, 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

$$y_i = \theta_i^* + \xi_i, \quad i = 1, 2, \ldots, n, \tag{5.2}$$

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.

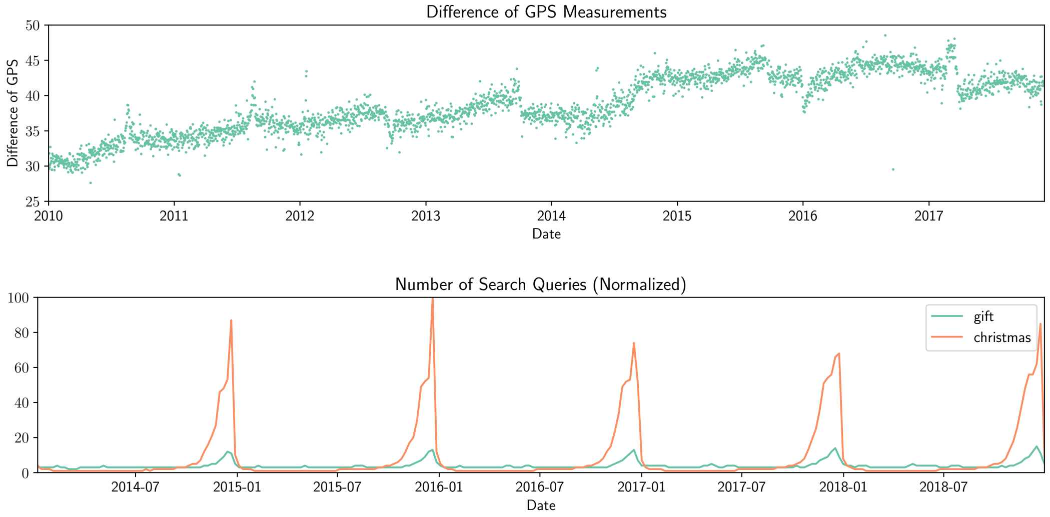

Fig. 5.1: **Examples of piecewise monotone signals in real-world data. Top**: The difference of the east-west component of GPS measurements between Victoria (British Columbia, Canada) and Seattle (United States). The trend factor seems to be approximated by a piecewise monotone signal. A possible reason for this behavior is the seismological phenomenon reported in Roggers and Dragert (2003). See Section 5.7.3 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`).

Definition 5.1. 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:

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

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 5.1, a monotone vector in $K_n^{\uparrow}$ is m-piecewise monotone for any $m \geq 1$. 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 $\hat{\theta}_{\text{id}} = y$.

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 5.1 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 5.1 for examples). Third, in the ranking systems in online

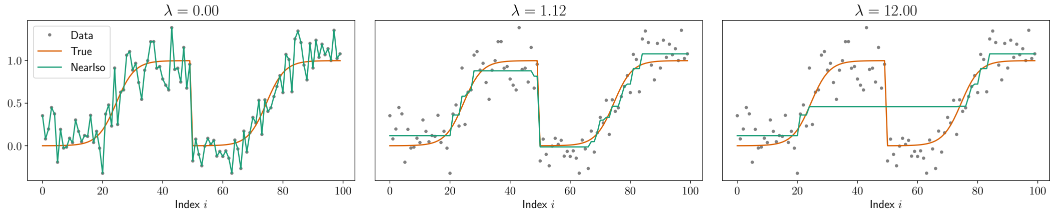

Fig. 5.2: **Examples of the nearly-isotonic regression estimators with different choices of tuning parameters.** The nearly-isotonic regression interpolates between the identity estimator $\hat{\theta}_{\mathrm{id}} = y$ and the isotonic regression $\hat{\theta}_{K_n^{\uparrow}}$.

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 chapter, 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 $\lambda \geq 0$, the nearly-isotonic regression estimator $\hat{\theta}_\lambda$ is defined as

$$\hat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})_+ \right\}, \tag{5.3}$$

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 (5.1).

Our goal in this chapter 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 5.2 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 the isotonic regression can be much worse than that of the identity estimator, which coincides with σ^2/n (see Section 5.3.2). Indeed, we can choose a 2-piecewise monotone vector $\theta^* \in K_{n/2}^{\uparrow} \times K_{n/2}^{\uparrow}$ 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.

5.1.1 Summary of theoretical results

In this chapter, we investigate the adaptation property of the nearly-isotonic regression estimators defined in (5.3).

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

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

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 $\mathcal{V}(\theta^*) \leq \mathcal{V}$ (Zhang 2002). Hence, a natural question is whether a similar rate can be achieved in piecewise monotone regression.

In Section 5.3.1, 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 5.3.1). 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 5.5, 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 5.4, we provide the following risk bound for the nearly-isotonic regression estimator (5.3). A precise statement is given in Corollary 5.16.

Claim 5.2. Let θ^* be a piecewise monotone vector on a partition $\Pi = (A_1, A_2, \ldots, A_m)$. 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 $\mathcal{V}(\theta^*_{A_j}) \leq \mathcal{V}/m$.

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

Then, the estimator (5.3) with optimally tuned parameter λ satisfies the following risk bound:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq C\left\{\left(\frac{\sigma^2\mathcal{V}\log en}{n}\right)^{2/3} + \frac{\sigma^2 m}{n}\log\frac{en}{m}\right\}. \tag{5.4}$$

The above claim is obtained as a corollary of a more general risk bound in Section 5.4. 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 (5.4) 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 (5.3). We will provide a precise definition of the growth condition in Section 5.4.3.

5.1.2 Organization

The rest of this chapter is organized as follows. In Section 5.2, 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 5.3. In Section 5.4, 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 5.2 in the above is provided in Section 5.4.3. In Section 5.5, 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. In Section 5.6, we review the algorithms for the nearly-isotonic regression and related estimators and discuss their computational complexities. Furthermore, we present some numerical examples in Section 5.7. After that, we have also included all proofs of the theoretical results.

5.1.3 Notation

Throughout this chapter, 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 nondecreasing, then $\mathcal{V}(\theta) = \theta_n - \theta_1$ and $\mathcal{V}_-(\theta) = 0$. In this chapter, 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 $A = \{\tau, \tau+1, \dots, \tau+J-1\}$ is a connected subset of $[n]$, we denote by θ_A a sub-vector $(\theta_\tau, \theta_{\tau+1}, \dots, \theta_{\tau+J-1})^\top \in \mathbb{R}^J$. We also denote by $\mathcal{V}^A(\theta_A)$ the total variation of θ_A.

5.2 Related work

There are two classes of estimators that are closely related to the nearly-isotonic regression (5.3): 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. 2017), the multi-dimensional convex regression (Han and Wellner 2016), and the constrained form trend filtering estimator (Guntuboyina et al. 2017). 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. 2017). In particular, for the one-dimensional isotonic regression, Chatterjee et al. (2015) and Bellec (2018) proved the following oracle inequality

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{K_n^{\uparrow}}-\theta^*\|_2^2 \le \inf_{\theta\in K_n^{\uparrow}}\left\{\frac{1}{n}\|\theta-\theta^*\|_2^2+\frac{\sigma^2 k(\theta)}{n}\log\frac{en}{k(\theta)}\right\}, \tag{5.5}$$

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

$$\hat{\theta}_{\mathrm{fused},\lambda}=\operatorname*{argmin}_{\theta\in\mathbb{R}^n}\left\{\frac{1}{2}\|y-\theta\|_2^2+\lambda\sum_{i=1}^{n-1}|\theta_i-\theta_{i+1}|\right\}, \tag{5.6}$$

where $\lambda \ge 0$ is the tuning parameter. The fused lasso poses the penalty whenever $\theta_i \neq \theta_{i+1}$, whereas the penalty of the nearly-isotonic regression (5.3) 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. (2017). In particular, Guntuboyina et al. (2017) showed an oracle inequality of the following form:

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathrm{fused},\lambda^*}-\theta^*\|_2^2 \le \inf_{\theta\in\mathbb{R}^n}\left\{\frac{1}{n}\|\theta-\theta^*\|_2^2+C\frac{\sigma^2 k(\theta)}{n}\log\frac{en}{k(\theta)}+C\Delta_{\mathrm{fused}}(\theta)\right\}, \tag{5.7}$$

One can control the quantity $\Delta_{\mathrm{fused}}(\theta)$ by assuming a mild regularity condition on θ^* so that the inequality (5.7) recovers the minimax rate for the piecewise constant vectors (see e.g., Gao et al. (2017)). However, even if θ^* is a monotone vector, (5.7) does not recover the rate of the isotonic regression (5.5) because $\Delta_{\mathrm{fused}}(\theta)$ becomes zero if and only if θ is just a constant vector.

Our risk bound for the nearly-isotonic regression in Section 5.4.2 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:

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

Like in the case of the fused lasso (5.7), this inequality provides a meaningful risk bound even if we cannot approximate θ^* by a monotone vector. Furthermore, $\Delta_{\mathrm{neariso}}(\theta)$ becomes zero for any monotone vector $\theta \in K_n^{\uparrow}$. Hence, our result can exactly recover the rate achieved by the isotonic regression (5.5).

5.3 Lower bounds

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

5.3.1 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 5.3. 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 $\theta_i^* := \theta_{A_i}^*$ 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

$$\inf_{\hat{\theta}_i} \sup_{\substack{\theta_i^* \in K_{A_i}^{\uparrow}: \\ \mathcal{V}(\theta_i^*) \le \mathcal{V}_i}} \frac{1}{n_i} \mathbb{E}_{\theta_i^*} \|\hat{\theta}_i - \theta_i^*\|_2^2 \ge C_1 \left(\frac{\sigma^2 \mathcal{V}_i}{n_i} \right)^{2/3} \quad \text{for all } i = 1, \dots, m. \tag{5.8}$$

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

$$C_1 \sum_{i=1}^{m^*} \frac{n_i}{n} \left(\frac{\sigma^2 \mathcal{V}_i}{n_i} \right)^{2/3}. \tag{5.9}$$

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 (5.9) 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 5.3.

Proposition 5.4. 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 5.3. Then, for any estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$, we have the following lower bound:

$$\sup_{\theta^* \in \Theta} \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|_2^2 \geq C \max\left\{ \left(\frac{\sigma^2 \mathcal{V}}{n}\right)^{2/3}, \quad \frac{\sigma^2 m}{n} \log \frac{\mathrm{e}n}{m} \right\}, \tag{5.10}$$

where $C > 0$ is a universal constant.

It remains to verify that the lower bound (5.10) is tight. Thus, in Section 5.5, we will construct an estimator that adaptively achieves a similar rate.

5.3.2 Lower bound for projection estimators

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 (5.9). 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

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

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 5.5. 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 $\mathrm{dist}(\theta^*, K_{2n}^{\uparrow}) = \sqrt{nM^2/2}$ diverges as $M \to \infty$. Figure 5.2 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.

5.4 Risk bounds for nearly-isotonic regression

In this section, we develop the risk bound for the nearly-isotonic regression estimator (5.3). Proofs of all the theorems and propositions in this section are presented in Appendix 5.9.

5.4.1 Risk bounds for constrained estimators

Before considering the original version of the nearly-isotonic regression (5.3), we consider the performance of the *constrained form nearly-isotonic regression* $\hat{\theta}_{\mathcal{V}}$ defined by the following constrained optimization problem:

$$\text{minimize } \|y-\theta\|_2^2 \quad \text{subject to } \sum_{i=1}^{n-1}(\theta_i-\theta_{i+1})_+ \le \mathcal{V}, \tag{5.11}$$

where $\mathcal{V} \ge 0$ is the tuning parameter. By the fundamental duality theorem in convex optimization, there exists a Lagrange multiplier $\lambda_{\mathcal{V}} \ge 0$ such that the regularization type formulation (5.3) admits the same solution $\hat{\theta}_{\lambda_{\mathcal{V}}} = \hat{\theta}_{\mathcal{V}}$. Hence, the solution path of penalized estimators $\{\hat{\theta}_\lambda : \lambda \ge 0\}$ and that of constrained estimators $\{\hat{\theta}_{\mathcal{V}} : \mathcal{V} \ge 0\}$ are equivalent. However, the properties of estimators with fixed values of $\lambda \ge 0$ and $\mathcal{V} \ge 0$ can be different in the following sense:

- From a computational perspective, calculating the constrained estimator (5.11) for a given $\mathcal{V} \ge 0$ is more difficult than the regularization estimator (5.3). For the regularization estimator (5.3), we can use the Modified Pool Adjacent Violators Algorithm (Modified PAVA) proposed by Tibshirani et al. (2011), which outputs the solution path for every $\lambda \ge 0$. In particular, given $\lambda \ge 0$, 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 (5.11) that run as fast as the algorithms for the penalized problem (5.3). We present detailed explanations for the algorithms in Section 5.6.
- 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 (5.3) or (5.11) cannot be directly applied to the other.

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

Before proceeding, we introduce some notations. Suppose that $\theta \in \mathbb{R}^n$ is piecewise constant on a connected partition $\Pi_{\text{const}} = \{A_1, \dots, A_k\}$ of $[n]$. We denote by $k(\theta) := |\Pi_{\text{const}}|$ 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) $A_i = \{\tau_i, \tau_i + 1, \dots, \tau_{i+1} - 1\}$ for $i = 1, \dots, k$ 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 $(i = 1, \dots, k+1)$ as

$$\begin{aligned} &w_1 = w_{k+1} = 0 \quad \text{and} \\ &w_i = \begin{cases} 1 & (t_{i-1} > t_i) \\ 0 & (t_{i-1} < t_i) \end{cases} \quad \text{for } i = 2, \dots, k. \end{aligned} \tag{5.12}$$

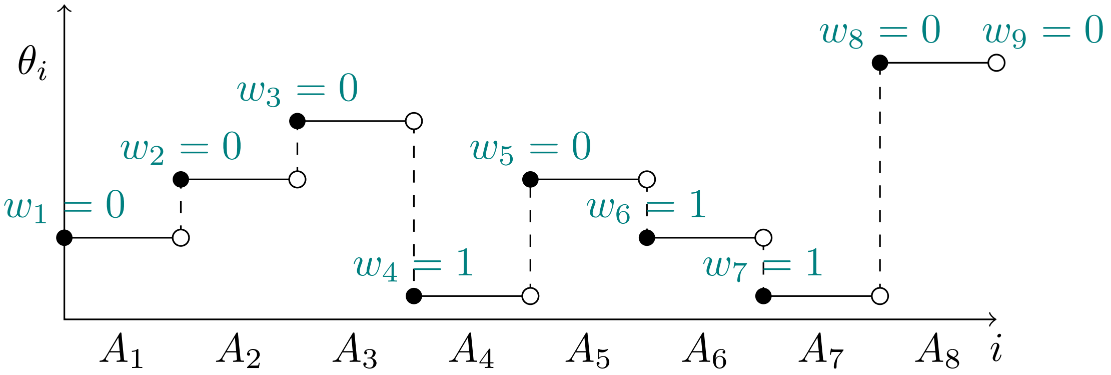

Fig. 5.3: **Illustration of the knot signs defined in** (5.12)**.** In this example, θ is assumed to be k-piecewise constant with $k = 8$. The corresponding signs are given as $(w_1, w_2, \ldots, w_8, w_9) = (0, 0, 0, 1, 0, 1, 1, 0, 0)$. Moreover, if we assume $|A_1| = |A_2| = \cdots = |A_8|$, the quantity $M(\theta)$ defined in (5.13) is given as $M(\theta) = \frac{1}{|A_4|} + \frac{1}{|A_5|} + \frac{1}{|A_6|} + \frac{1}{|A_8|} = \frac{4k}{n}$.

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 5.3 for the graphical illustration. Then, we define $M(\theta)$ as

$$M(\theta) := \sum_{j=2}^{k} \max\left\{\frac{1}{|A_j|}, \frac{k}{n}\right\} 1_{\{w_{j-1} \neq w_j\}}. \tag{5.13}$$

$M(\theta)$ 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 5.6. For any $\theta^* \in \mathbb{R}^n$, the constrained nearly-isotonic regression (5.11) satisfies the following oracle inequality:

$$\begin{aligned} &\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \\ &\leq \inf_{\substack{\theta \in \mathbb{R}^n: \\ \mathcal{V}_-(\theta) = \mathcal{V}}} \left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)}\right\}. \end{aligned} \tag{5.14}$$

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

$$\begin{aligned} &\frac{1}{n}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \\ &\leq \inf_{\substack{\theta \in \mathbb{R}^n: \\ \mathcal{V}_-(\theta) = \mathcal{V}}} \left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)}\right\} \\ &+ \frac{4\sigma^2\log\eta^{-1}}{n} \end{aligned} \tag{5.15}$$

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 5.6.

Corollary 5.7. Suppose $\theta^* \in \mathbb{R}^n$. Choose $\mathcal{V}^* \geq 0$ that minimizes the upper bound in (5.14) (thus, $\mathcal{V}^*$ depends on the true parameter θ^*). Then, we have

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}^*} - \theta^*\|_2^2\\ &\leq \inf_{\theta \in \mathbb{R}^n}\left\{\frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)}\right\}. \qquad (5.16)\end{aligned}$$

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

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2 \leq C\sigma^2\left\{\frac{k(\theta^*)}{n}\log\frac{en}{k(\theta^*)} + \frac{M(\theta^*)}{k(\theta^*)}\log\frac{en}{k(\theta^*)}\right\}. \qquad (5.17)$$

Remark 5.8. We briefly comment on the proof of Theorem 5.6 and Corollary 5.7. 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. (2017). See Appendix 5.9 for detailed proofs.

By restricting the region over which the infimum in (5.16) 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 (5.16), we have to control the quantity $M(\theta)$ defined in (5.13). 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

$$\min\{|A_i| : 1 \leq i \leq k, w_i \neq w_{i+1}\} \geq \frac{cn}{k}, \qquad (5.18)$$

where the partition $\Pi_{\text{const}} = \{A_1, A_2, \ldots, A_k\}$ and the signs w_i $(i = 1, \ldots, k+1)$ are defined as in (5.13). Intuitively, a signal $\theta \in \mathbb{R}^n$ is well approximated by another signal that satisfies the minimal length condition if θ has "moderate growth" around the order-violating jumps. For further discussion on such growth conditions, see Section 5.4.3.

Based on the minimal length condition, we have the following result from Theorem 5.6 .

Corollary 5.9. Suppose that $\theta^* \in \mathbb{R}^n$ satisfies the minimal length condition (5.18) for a constant $c > 0$. Assume that θ^* is $k(\theta^*)$-piecewise constant and $m(\theta^*)$-piecewise monotone. Then, the constrained nearly-isotonic regression (5.11) satisfies

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_{\mathcal{V}} - \theta^*\|_2^2\\ &\leq (\mathcal{V}_-(\theta^*) - \mathcal{V})^2 + C\sigma^2\left(\frac{k(\theta^*)}{n} + \frac{2c^{-1}(m(\theta^*) - 1)}{n}\right)\log\frac{en}{k(\theta^*)}. \qquad (5.19)\end{aligned}$$

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 5.10. If θ is k-piecewise constant and m-piecewise monotone, it is always true that $k \geq 2(m-1)$. Hence, the inequality (5.19) 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 5.11. We comment on the minimal length condition and the relation to estimation of piecewise constant vectors. The minimal length condition for the total variation regularization estimator is considered by Guntuboyina et al. (2017). In the problem of estimating k-piecewise constant vectors, it is shown that the minimax rate is $\frac{k}{n}\log\frac{en}{k}$ (see, e.g., Gao et al. 2017). For the fused lasso, Fan and Guan (2017) 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. (2017), Remark 2.5). On the other hand, it is proved that there are other classes of estimators that do not suffer from the minimal length condition (Gao et al. 2017, Fan and Guan 2017).

5.4.2 Risk bounds for penalized estimators

In this section, we consider the risk bounds for the nearly-isotonic regression (5.3) in the original penalized form by Tibshirani et al. (2011).

Theorem 5.12. For any $\lambda \geq 0$, let $\hat{\theta}_\lambda$ denote the nearly-isotonic regression estimator defined in (5.3). 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:

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq \frac{1}{n}\|\theta - \theta^*\|_2^2 + C\sigma^2\frac{k(\theta)}{n}\log\frac{en}{k(\theta)} + C\sigma^2\frac{M(\theta)}{k(\theta)}\log\frac{en}{k(\theta)}\\ + 3(\lambda - \lambda^*)^2 M(\theta),\end{aligned} \tag{5.20}$$

where $M(\theta)$ and $k(\theta)$ are defined similarly as in Theorem 5.6. Furthermore, for any $\eta \in (0,1)$, the inequality

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

holds with probability $1 - \eta$.

We comment on some direct consequences of Theorem 5.12. In this theorem, $\lambda^*(\theta)$ is defined as a function of θ. To understand the risk bound (5.20), 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 (5.16):

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

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 (5.18) on θ^*, we obtain a simplified bound of the form (5.17).

We move on to discuss a precise expression of $\lambda^*(\theta)$ in Theorem 5.12. The next proposition provides an upper bound for $\lambda^*(\theta)$.

Proposition 5.13. Suppose $\theta \in \mathbb{R}^n$. Let $\Pi_{\mathrm{const}}(\theta) := \{A_1, A_2, \ldots, A_k\}$ be the constant partition of θ, and $w_1, w_2, \ldots, w_{k+1}$ be the associated signs defined in (5.12). Then, there is a universal constant $C > 0$ such that $\lambda^*(\theta)$ in Theorem 5.12 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 5.13 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 (5.6), 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. 2017). For a detailed comparison of these results, see Remark 2.7 by Guntuboyina et al. (2017) and references therein.

Remark 5.14. 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 5.13 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)

$$\mathrm{SURE}(\lambda) = \frac{1}{n}\|y - \hat{\theta}_\lambda\|_2^2 + \frac{2\hat{\sigma}^2}{n}\hat{\mathrm{df}}(\hat{\theta}_\lambda) + (\text{constant}). \tag{5.21}$$

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 (5.2) is interpreted as a discrete observation of a continuous signal. Specifically, suppose that the data is generated according to the following nonparametric regression model:
$$y_i = f^*(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{5.22}$$
where $x_1 < x_2 < \dots < x_n$ 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
$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \sum_{i=1}^{n-1} \frac{(\theta_i - \theta_{i+1})_+}{x_{i+1} - x_i} \right\}. \tag{5.23}$$
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 $f^*(x^{\text{new}})$ by $\hat{f}_\lambda(x^{\text{new}})$. Hence, we can naturally apply the cross-validation in this situation.

5.4.3 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 5.15. 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 5.4 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 (5.18) 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 5.16. 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 $j = 1, \dots, m$.
(c) $\theta^*_{A_j}$ satisfies the moderate growth condition for each $j = 1, 2, \dots, m$.

Then, the estimator (5.3) with optimally tuned parameter λ satisfies the following risk bound:
$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq C \max\left\{ \left(\frac{\sigma^2 \mathcal{V} \log \frac{en}{m}}{n} \right)^{2/3}, \ \frac{\sigma^2 m}{n} \log \frac{en}{m} \right\}. \tag{5.24}$$

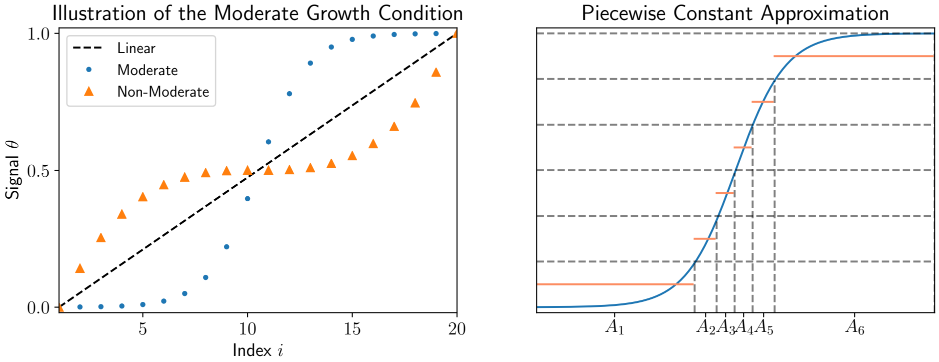

Fig. 5.4: **Illustration of the moderate growth condition. Left:** The plotted three signals are monotone vectors in $K_n^{\uparrow}$ with $n = 20$ and $\mathcal{V}(\theta) = 1$. The dotted line represents the linear signal $\theta_i^{\text{linear}} = i/n$ $(i = 1, 2, \ldots, n)$. The blue circles depict an example of a signal that satisfies the moderate growth condition. That is, it is not larger than the linear signal θ_i^{linear} for $1 \leq i \leq 10$, and not less than θ_i^{linear} for $10 \leq i \leq 20$. 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 5.9.5 for a detailed explanation.

The risk bound (5.24) achieves the minimax rate over $\Theta_n(m, \mathcal{V})$ in Proposition 5.4 up to a multiplicative factor of $\log^{2/3} \frac{en}{m}$. We should note that the restrictive assumption (a) in Corollary 5.16 is employed merely for the sake of simplicity of the proof. We may relax this assumption as

$$\min_{1 \leq i \leq m} |A_i| \geq \frac{c'n}{m}$$

for some $c' > 0$.

5.5 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 5.4 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 $K_\Pi^{\uparrow}$. 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:

$$\hat{\Pi} \in \operatorname*{argmin}_{\Pi} \left\{ \|y - \hat{\theta}_{\Pi}\|_2^2 + \mathrm{pen}(\Pi) \right\}. \tag{5.25}$$

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

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

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

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

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

$$\min_{\Pi} \min_{\mathbf{V} \in \mathscr{V}(|\Pi|)} \left\{ \|y - \hat{\theta}_{\Pi,\mathbf{V}}\|_2^2 + \mathrm{pen}(\Pi, \mathbf{V}) \right\}. \tag{5.26}$$

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

Theorem 5.17. 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

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

Let $(\hat{\Pi}, \hat{\mathbf{V}})$ be the minimizer in (5.26).

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

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

$$\begin{aligned} &\frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}} - \theta^*\|_2^2 \\ &\leq 2C_{\mathrm{pen}} \left\{ \sum_{i=1}^{m} \frac{|A_i|}{n} \left(\frac{\sigma^2(\mathcal{V}^{A_i}(\theta^*_{A_i}) + 1)}{|A_i|} \right)^{2/3} + \frac{\sigma^2 m}{n} \log \frac{en}{m} \right\} + \frac{256\sigma^2}{n}. \end{aligned} \tag{5.27}$$

We emphasize that Theorem 5.17 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 5.4. However, the penalized model selection estimator used in Theorem 5.17 is not practical. One reason is that the constant C_{pen} in the definition of the penalty term is too large for a practical purpose. Another reason is the computational issue. The estimator (5.26) is obtained through the minimization over exponentially many possible partitions Π.

The dependence on the total variation of each segment in (5.27) 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., (5.25)) and remove the dependence on the sieve resolution remains an open problem.

5.6 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 5.18).

5.6.1 Penalized estimators

Here, we introduce two algorithms to solve the penalized form nearly-isotonic regression (5.3). In Section 5.6.1, 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 10, 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.

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 (5.3). Here, we present a variant of the modified PAVA for the following weighted version of the estimator:

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

where $c_i > 0$ $(i = 1, 2, \dots, n-1)$ 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 (5.23).

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 edge weights, this

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

```
Input: $y \in \mathbb{R}^n$
Output: Set of finitely many breakpoints $\Lambda = \{\lambda_0, \lambda_1, \dots, \lambda_N\}$, solution path
        $\{\hat{\theta}_\lambda\}_{\lambda \in \Lambda}$
$\lambda_0 \leftarrow 0$, $\hat{\theta}_{\lambda_0} \leftarrow y$
Let $\Pi_0$ be the constant partition of $\hat{\theta}_{\lambda_0}$. Below, the solution $\hat{\theta}_{\lambda_i}$ is kept to be
  constant on $\Pi_i$.
for $i = 1, 2, \dots$ do
    Let $k = |\Pi_{i-1}|$. Let $A_j = \{\tau_j, \tau_j + 1, \dots, \tau_{j+1} - 1\}$ be the $j$-th element in the
      partition $\Pi_{i-1}$, and $t_j$ be the value of $\hat{\theta}_{\lambda_{i-1}}$ on $A_j$ $(j = 1, 2, \dots, k)$.
    Set $s_0 = s_k = 0$ and $c_0 = 0$. Compute $s_j = 1_{\{t_j > t_{j+1}\}}$ for $j = 1, 2, \dots, k - 1$.
    Compute the slopes $m_j$ $(j = 1, 2, \dots, k)$ by
        $$m_j = \frac{c_{\tau_j - 1} s_{j-1} - c_{\tau_{j+1} - 1} s_j}{|A_j|}.$$
    Compute $\delta$ 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 $t_j + m_j \delta$
      $(j = 1, 2, \dots, k)$.
    Set $\Pi_i$ to be the constant partition of $\hat{\theta}_{\lambda_i}$.
end
```

need not be true. Instead, we have a sufficient condition for the validity of Algorithm 1 as follows:

Proposition 5.18. Algorithm 1 outputs the exact solution path if the edge weights satisfy the following condition.

$$\frac{c_{j+1}}{c_j} \le \frac{j+1}{j} \quad \text{for all } j = 1, 2, \dots, n-2.$$

For instance, we can apply Algorithm 1 to calculate the solution path of (5.23) if the design points $x_1 < x_2 < \dots < x_n$ satisfies

$$x_{j+2} - x_{j+1} \ge \frac{j}{j+1}(x_{j+1} - x_j)$$

for all $j = 1, 2, \dots, n-2$. For a detailed discussion for this condition, see Remark 5.45 in Appendix 5.9.6.

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

$$\hat{\theta}_{G,\lambda} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \mathcal{V}_G(\theta) \right\} \tag{5.28}$$

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

$$\mathcal{V}_G(\theta) := \sum_{(i,j) \in E} c_{(i,j)} (\theta_i - \theta_j)_+ . \tag{5.29}$$

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, (5.28) can be regarded as a generalization of the nearly-isotonic regression to general directed graphs.

The problem of the form (5.28) has been well studied in the optimization literature. In particular, we can see that solving (5.28) 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, (5.28) 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. (2011a) 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 (5.28). In the inner loop, the algorithm recursively solves max-flow problems by defining smaller networks (Algorithm 3). See Figure 5.5 for examples of networks used in the first two recursions in the algorithm.

Algorithm 2: Divide-and-Conquer algorithm for the generalized nearly-isotonic regression 5.28

```
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 (5.28)
Construct a flow network $\mathcal{N}$ by adding a source node $s$ and a sink node $t$ to the
  graph $G$.
Compute $\hat{\theta}_\lambda = \mathrm{Prox}_{\lambda F_{\mathcal{N}}}(y)$ according to Algorithm 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:

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; y) + \lambda \mathcal{V}_G(\theta) \right\}, \tag{5.30}$$

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.

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 5.5-(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 5.5-(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 5.5-(c)).
```

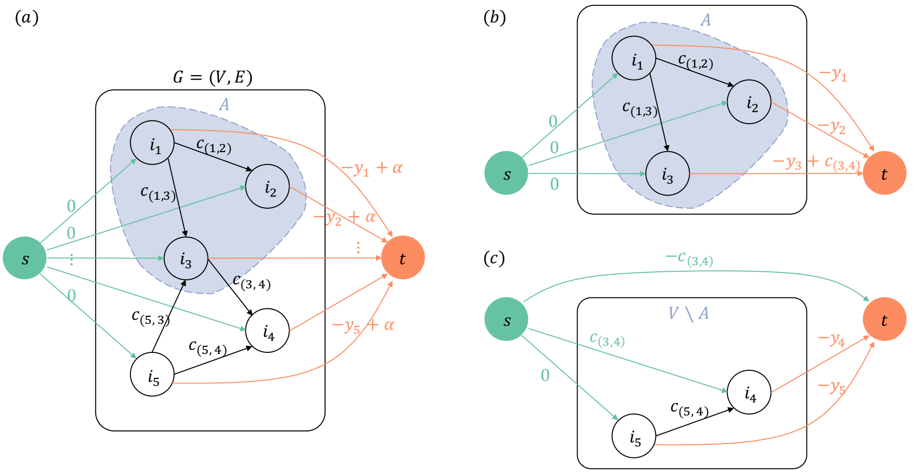

Fig. 5.5: **Flow networks in Algorithm 3.** In this example, we assume $\lambda = 1$. (a) A network that corresponds to the minimization problem in line 3. (b) A network that corresponds to the function $B \mapsto \lambda F_{\mathcal{N}|A}(B) - y(B)$ in line 5. (c) A network that corresponds to the function $B \mapsto \lambda F_{\mathcal{N}^A}(B) - y(B)$ in line 6.

Since we already have the proximal operator (5.28) of the penalty term $\lambda \mathcal{V}_G$, we can also obtain algorithms that output approximate minimizers of the above problem. 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 $\mathrm{O}(\epsilon^{-2})$ calls of the minimization algorithm for (5.28).

5.6.2 Constrained estimators

Consider the following generalized version of the constrained form of nearly-isotonic regression (5.11):

$$\text{minimize } \|y-\theta\|_2^2 \quad \text{subject to} \sum_{(i,j)\in E} c_{(i,j)}(\theta_i-\theta_j)_+ \le \mathcal{V}, \tag{5.31}$$

Unlike the penalized estimators, it is difficult to find an exact solution of (5.31). Since problem (5.31) 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 the cutting plane method.

Proposition 5.19. 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 $\mathrm{O}((n+|E|)n^2\log^{\mathrm{O}(1)}\frac{n}{\epsilon|E|})$.

5.7 Simulations

We provide some numerical examples for piecewise monotone regression problems.

5.7.1 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 (5.3) 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:

$$\hat{\theta}_{\text{boundary},\lambda,\mu} = \operatorname*{argmin}_{\theta\in\mathbb{R}^n}\left\{\frac{1}{2}\|y-\theta\|_2^2 + \lambda\sum_{i=1}^{n}(\theta_i-\theta_{i+1})_+ + \mu(\theta_n-\theta_1)\right\}, \tag{5.32}$$

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 (5.3) 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. (2019), Wu et al. (2015) and Luss and Rosset (2017).

5.7.2 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 (5.22), 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:

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

It is worth noting that the former sigmoidal function f_{sigmoid} satisfies the moderate growth condition (see Definition 5.15), whereas the latter cubic function f_{cube} does not. Hence, for the case of piecewise sigmoidal functions $f_{\text{sigmoid}}^{(m)}$, the minimax rate of $\mathrm{O}(n^{-2/3})$ is achieved by both the nearly-isotonic regression and the fused lasso (see Corollary 5.16 above and Corollary 2.8 by Guntuboyina et al. (2017)).

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 (5.3).
- `NearisoBC`: The nearly-isotonic regression with boundary correction (5.32)
- `Fused`: The fused lasso (5.6).
- `PO`: The projection estimator with the partition oracle, i.e., the projection estimator onto $K_\Pi^\uparrow$ provided with the true partition Π.

For `Neariso` and `Fused`, the tuning parameter λ is selected by generalized C_p criteria (i.e., minimizing SURE (5.21)). 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 5.6 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 5.6 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.

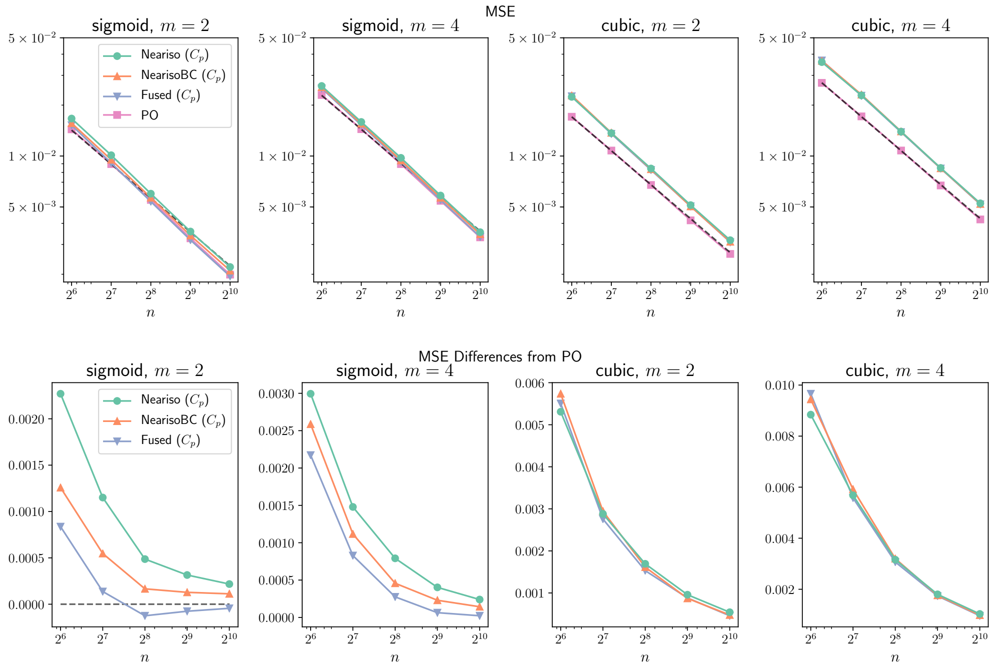

Fig. 5.6: **The risks of nearly-isotonic type estimators on simulated data**. The upper line shows log-log plots of the MSEs versus n. The lower line shows the difference of the MSEs between regularization type estimators (i.e., `Neariso` `NearisoBC` and `Fused`) and the projection estimator with the oracle partition choice (`PO`).

5.7.3 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 5.7 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 (5.30) with $\mathcal{L}(\theta; y) = \sum_{i=1}^{n} \ell_\delta(\theta_i - y_i)$. Here, ℓ_δ is the Huber loss:

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

which is commonly used in the robust regression literature.

We applied the nearly-isotonic regression (5.3) 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 5.7 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 in Figure 5.7 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 5.7 shows that the robust version of the nearly-isotonic regression outputs similar clusters of change points as in the second panel.

5.7.4 Supplemental experiments on two-dimensional grids

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 (5.28). 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 $\theta_{ij} \leq \theta_{kl}$ whenever $i \leq k$ and $j \leq l$. 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

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

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 5.8 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

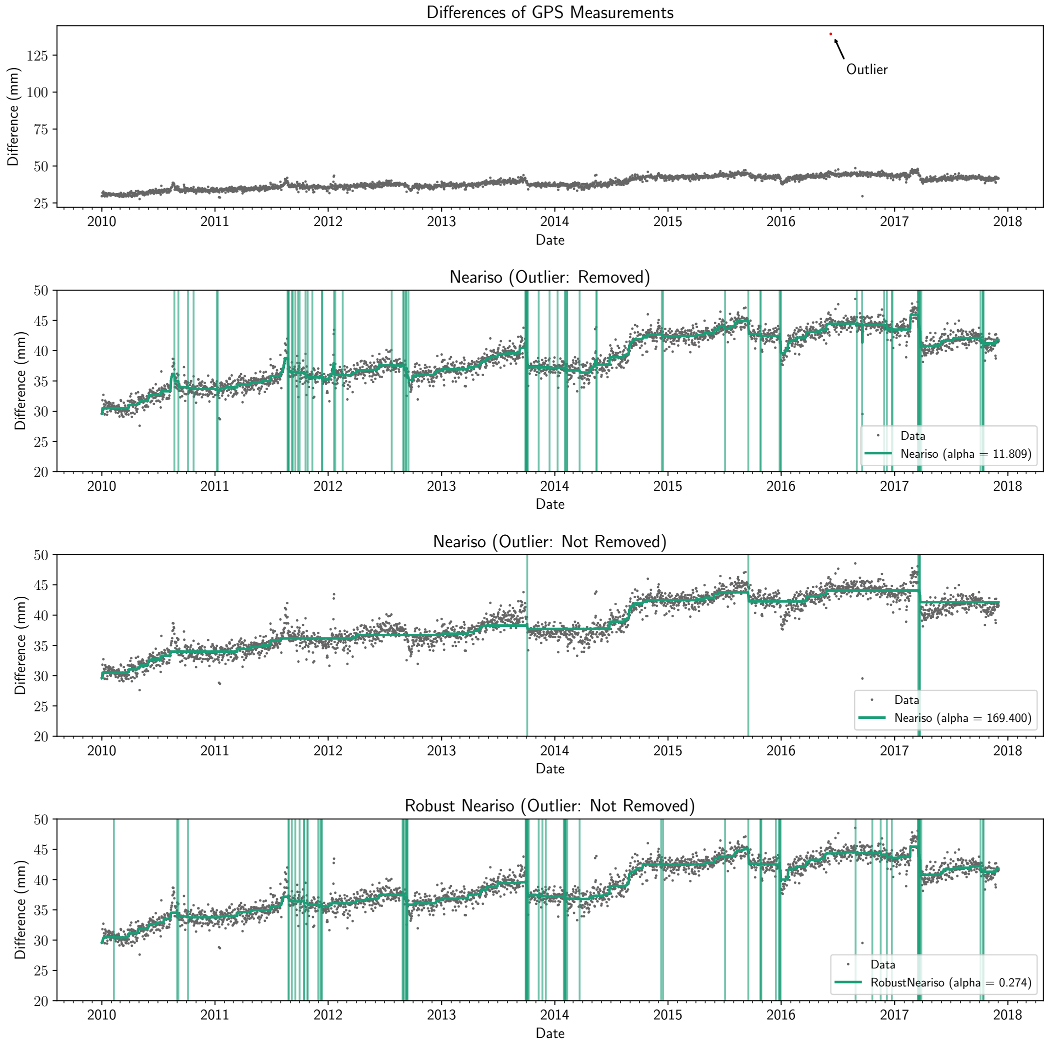

Fig. 5.7: **Nearly-isotonic type estimators applied to GPS data**. See the text for details.

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 5.5. `Neariso2` is the special case of the generalized nearly-isotonic regression (5.28) 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 5.8 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

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

or

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

where we write $x_i = \frac{i-1}{k-1}$ for $i = 1, 2, \dots, k$. With the former choice, θ^* becomes an m^2-piecewise monotone matrix. With the latter choice, θ^* becomes an m-piecewise monotone

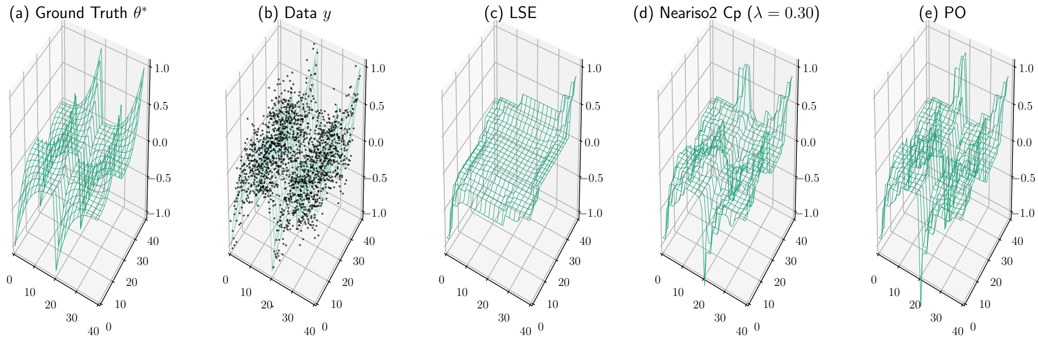

Fig. 5.8: **Examples of estimators for piecewise monotone matrices.** The true parameter θ^* is a 32×32 matrix that is monotone on each 16×16 segment. The bivariate isotonic regression (`LSE`) does not capture the piecewise monotone structure. The solution of the nearly-isotonic regression (`Neariso2`) seems to be close to the partition oracle (`PO`).

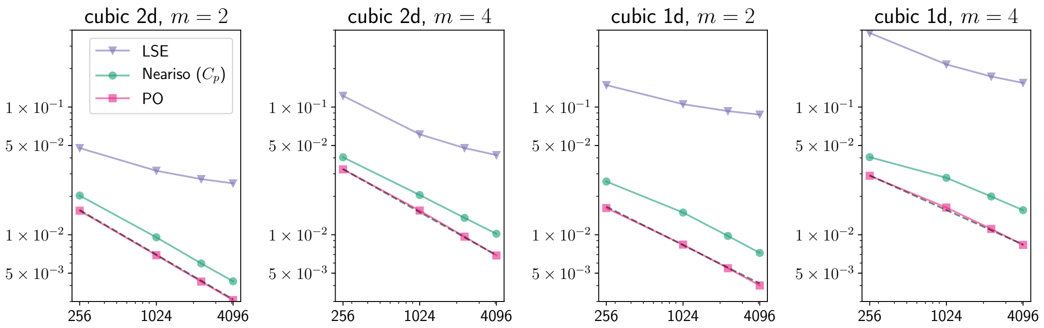

Fig. 5.9: **The risks in piecewise monotone matrix estimation**. See the text for details.

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 5.9 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 5.9. 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).

5.8 Proofs for Section 5.3

The remaining four sections in this chapter provide missing proofs in the previous sections.

5.8.1 Proof of Proposition 5.4

Let Θ be either $\tilde{\Theta}_n(m, \mathcal{V})$ or $\Theta_n(m, \mathcal{V})$, which are defined in Definition 5.3. The minimax lower bound (5.10) is proved by combining the following two lower bounds:

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

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

(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 $|\{i : \theta_i \neq \theta_{i+1}\}| \leq m - 1$. 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 $1_{\{i \geq j\}}$. Then, $\mathcal{C}(m)$ contains the set $\{\theta = X\beta : \|\beta\|_\infty \leq m\}$, and the lower bound for the minimax risk over $\mathcal{C}(m)$ follows from the well-known results for ℓ_∞ balls (e.g., Raskutti et al. (2011)). In particular, for any $m \geq 3$, the following lower bound is presented in Gao et al. (2017):

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

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 $|\Pi| \leq m$ so that each $A \in \Pi$ does not contain such large jumps, which implies that $\theta \in \Theta_n(m, \mathcal{V})$.

5.8.2 Proof of Proposition 5.5

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 5.20 (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

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

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

$$\frac{1}{n} \max\left\{ t_{\theta^*}^2 - C_1 t_{\theta^*}^{3/2}, 0 \right\} \leq \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta}_K - \theta^*\|_2^2 \leq \frac{C_2}{n} \max\left\{ t_{\theta^*}^2, \sigma^2 \right\}. \tag{5.33}$$

To prove Proposition 5.5, we use the lower bound in (5.33). 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 $t \geq t_0$ so that $t^2 - Ct^{3/2} \geq n\sigma^2$. Then, for any θ^* such that $\text{dist}(\theta^*, K) \geq t$, we have

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

Remark 5.21. 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.

5.9 Proofs for Section 5.4

5.9.1 Preliminaries

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

Definition 5.22. 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

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

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

(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 5.22 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 5.25 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 5.23. 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 $\hat{\theta}_K = \operatorname{argmin}_{\theta\in K}\|y-\theta\|_2$. 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

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

Lemma 5.24 (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

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

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 $\lambda \geq 0$. We define the proximal denoising estimator $\hat{\theta}_\lambda$ as

$$\hat{\theta}_\lambda := \operatorname*{argmin}_{\theta\in\mathbb{R}^n}\left\{\frac{1}{2}\|y-\theta\|_2^2 + \sigma\lambda f(\theta)\right\}. \tag{5.34}$$

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. (2017) 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 5.25. 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 (5.34). Then, we have

$$\frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq \inf_{\theta\in\mathbb{R}^n}\left\{\frac{1}{n}\|\theta-\theta^*\|_2^2 + \frac{\sigma^2}{n}\mathbf{D}(\lambda\partial f(\theta))\right\}. \tag{5.35}$$

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

$$\frac{1}{n}\|\hat{\theta}_\lambda - \theta^*\|_2^2 \leq \inf_{\theta\in\mathbb{R}^n}\left\{\frac{1}{n}\|\theta-\theta^*\|_2^2 + \frac{2\sigma^2}{n}\mathbf{D}(\lambda\partial f(\theta^*))\right\} + \frac{16\sigma^2\log(\eta^{-1})}{n} \tag{5.36}$$

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

Proof. Below, we write $\hat{\theta} := \hat{\theta}_\lambda$. To prove (5.35), 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 $\theta \neq \hat{\theta}$ because otherwise the inequality is trivial.

From the first order optimality condition of the convex minimization problem (5.34), 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

$$\|\hat{\theta} - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2 \le 2\sigma\lambda(f(\theta) - f(\hat{\theta})) - 2\sigma\langle z, \theta - \hat{\theta} \rangle - \|\theta - \hat{\theta}\|_2^2. \tag{5.37}$$

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

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

Hence, the right-hand side of (5.37) is bounded from above by

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

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

$$\|\hat{\theta} - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2 \le \sigma^2 \inf_{v \in \partial f(\theta)} \|\lambda v - z\|_2^2 = \sigma^2 \mathrm{dist}^2(z, \lambda \partial f(\theta)). \tag{5.38}$$

By taking the expectation of both sides, (5.35) is proved.

To prove the high-probability bound (5.36), 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

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

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

$$|\mathrm{dist}(z_1, \lambda \partial f(\theta)) - \mathrm{dist}(z_2, \lambda \partial f(\theta))| \le \|(z_1 - P(z_1)) - (z_2 - P(z_2))\|_2 \le 2\|z_1 - z_2\|_2,$$

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

5.9.2 Risk bounds for constrained estimators (Proof of Theorem 5.6)

In this subsection, we provide the proof of Theorem 5.6 as an application of Lemma 5.24. To this end, we have to evaluate the statistical dimension of the tangent cone of a convex set

$$K_-(\mathcal{V}) := \{\theta \in \mathbb{R}^n : \mathcal{V}_-(\theta) \leq \mathcal{V}\} = \left\{\theta \in \mathbb{R}^n : \sum_{i=1}^{n-1}(\theta_i - \theta_{i+1})_+ \leq \mathcal{V}\right\}. \tag{5.39}$$

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 $K(\mathcal{V}) = \{\theta \in \mathbb{R}^n : \mathcal{V}(\theta) \leq \mathcal{V}\}$ in Guntuboyina et al. (2017). Our goal is to show the following upper bound for the statistical dimension:

Proposition 5.26. Suppose that θ is a vector with $\mathcal{V}_-(\theta) = \mathcal{V}$. Then, there exists a universal constant $C > 0$ such that

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

where $M(\theta)$ is defined in (5.13).

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 5.24 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 (5.12). Roughly speaking, the tangent cone of $K_-(\mathcal{V})$ is characterized by the sign of jumps.

Lemma 5.27. 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 $\Pi_{\text{const}}(\theta)$ 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

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

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

$$T(\Pi, w) = \left\{v \in \mathbb{R}^n : \sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(v_{B_i}) \leq \sum_{i=2}^{k'} w_i(v_{\tau_i} - v_{\tau_i - 1})\right\}, \tag{5.40}$$

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:

[*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 $B \subseteq A$.

(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 $T(\theta)$, 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 (5.12). Fix any $z \in T(\Pi, w)$. We have to show that there exists a (sufficiently small) $t > 0$ such that $\theta + tz \in K_-(\mathcal{V})$. Here, we have

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

Recall that $w_2, \ldots, w_k$ 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 $t > 0$ so that

$$((\theta_{\tau_i-1} + tz_{\tau_i-1}) - (\theta_{\tau_i} + tz_{\tau_i}))_+ = w_i((\theta_{\tau_i-1} + tz_{\tau_i-1}) - (\theta_{\tau_i} + tz_{\tau_i}))$$

for every $i = 2, 3, \ldots, k$. Indeed, if we choose $t > 0$ so that

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

the signs of θ do not change by adding tz. Consequently, we have

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

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

From Proposition 5.23-(i), we can bound the statistical dimension by the Gaussian width as follows:

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

Here, $B_n := \{v \in \mathbb{R}^n : \|v\|_2 \leq 1\}$ 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. (2017), we obtain the following characterization of this set.

Lemma 5.28. Let θ be a vector in $\mathbb{R}^n$ such that $\mathcal{V}_-(\theta) = \mathcal{V}$. Let $\Pi = \{B_1, B_2, \ldots, B_{k'}\}$ be any connected refinement of $\Pi_{\text{const}}(\theta)$. Define the signs $w_2, w_3, \ldots, w_{k'}$ as in Lemma 5.27, and let $w_1 = w_{k'+1} = 0$. Then, for every $v \in T_{K_-(\mathcal{V})}(\theta)$ with $\|v\|_2 \leq 1$, there exists indices $\ell_1 \in B_1, \ell_2 \in B_2, \ldots, \ell_{k'} \in B_{k'}$ such that

$$\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) \leq \left(\sum_{i=1}^{k'} \frac{1}{|B_i|} 1_{\{w_i \neq w_{i+1}\}} \right)^{\frac{1}{2}}, \tag{5.41}$$

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

$$\Gamma_i(v, \ell_i) := \mathcal{V}_-^{B_i}(v_{B_i}) - w_i(v_{\tau_i} - v_{\ell_i}) - w_{i+1}(v_{\ell_i} - v_{\tau_{i+1}-1}) \quad \text{for } i = 1, 2, \ldots, k'. \tag{5.42}$$

Proof. Fix $v \in T_{K_-(\mathcal{V})}(\theta) \cap B_n$. By Lemma 5.27, we have

$$\sum_{i=1}^{k'} \mathcal{V}_-^{B_i}(v_{B_i}) \leq \sum_{i=2}^{k'} w_i(v_{\tau_i} - v_{\tau_i - 1}) = \sum_{i=1}^{k'+1} w_i(v_{\tau_i} - v_{\tau_i - 1}). \tag{5.43}$$

Let $\ell_1 \in B_1, \ell_2 \in B_2, \ldots, \ell_{k'} \in B_{k'}$ be indices which will be specified later. Defining $\Gamma_i(v, \ell_i)$ as in (5.42), we can rewrite (5.43) as

$$\begin{aligned}
\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) &\leq \sum_{i=1}^{k'} w_i(v_{\ell_i} - v_{\tau_i}) + \sum_{i=1}^{k'} w_{i+1}(v_{\tau_{i+1}-1} - v_{\ell_i}) + \sum_{i=1}^{k'+1} w_i(v_{\tau_i} - v_{\tau_i - 1}) \\
&= \sum_{i=1}^{k'} (w_i - w_{i+1}) v_{\ell_i} \\
&\leq \sum_{i=1}^{k'} 1_{\{w_i \neq w_{i+1}\}} v_{\ell_i} \qquad (5.44)
\end{aligned}$$

Now, let t_i denote the ℓ_2 norm of v_{B_i} for $i = 1, 2, \ldots, k'$. By the assumption, $\sum_{i=1}^{k'} t_i^2 = \|v\|_2 \leq 1$. Then, for any $i \in \{1, 2, \ldots, k'\}$, there exists $\ell_i \in B_i$ such that $v_{\ell_i} \leq t_i / \sqrt{|B_i|}$. For these choices of ℓ_i, the right-hand side of (5.44) is bounded from above by

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

which proves the desired result. □

Remark 5.29. 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

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

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

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

In particular, we obtain $\Gamma_i(v, \ell_i) \geq 0$. If θ is monotone non-decreasing (i.e., $w_0 = w_1 = \cdots = w_{k+1} = 0$), then the right-hand side of (5.41) equals to 0, and so $\Gamma_i(v, \ell_i) = 0$.

Step 2: Quantizing the tangent cone Now, let $\Pi = \{B_1, B_2, \ldots, B_{k'}\}$ be a connected refinement of $\Pi_{\mathrm{const}}(\theta)$. Lemma 5.28 implies that $T_{K_-(\mathcal{V})}(\theta) \cap B_n$ is contained in the set such that $\sum_{i=1}^{k'} \|v_{B_i}\|_2^2 \leq 1$ and $\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) \leq \gamma$ 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) $\mathbf{t} = (t_1, t_2, \ldots, t_{k'})$ and $\mathbf{q} = (q_1, q_2, \ldots, q_{k'})$ are vectors that consist of non-negative numbers, and

(b) $\mathbf{l} = (\ell_1, \ell_2, \ldots, \ell_{k'})$ is a set of indices such that $\ell_i \in B_i$ for $i = 1, 2, \ldots, k'$.

For such triple, we define a set

$$T(\mathbf{t}, \mathbf{q}, \mathbf{l}) = \left\{ v \in \mathbb{R}^n : \|v_{B_i}\|_2^2 \leq t_i \quad \text{and} \quad \Gamma_i(v, \ell_i) \leq q_i \gamma \quad \text{for } i = 1, 2, \ldots, k' \right\}, \tag{5.45}$$

where γ is taken as the right-hand side of (5.41):

$$\gamma := \gamma(\theta, \Pi) = \left(\sum_{i=1}^{k'} \frac{1}{|B_i|} 1_{\{w_i \neq w_{i+1}\}} \right)^{\frac{1}{2}}. \tag{5.46}$$

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 5.30. Suppose that $\Pi = (B_1, B_2, \ldots, B_{k'})$ is a connected refinement of $\Pi_{\mathrm{const}}(\theta)$. Define the signs $w_1, w_2, \ldots, w_{k'}$ as in Lemma 5.28. Let $\mathcal{Q}$ be a set of allocation vectors satisfying the following condition; there exists an integer vector $\mathbf{m} = (m_1, m_2, \ldots, m_{k'}) \in \mathbb{N}$ such that $1 \leq m_i \leq k'$ $(i = 1, 2, \ldots, k')$ and $\sum_{i=1}^{k'} m_i \leq 2k'$, and the allocation vector $q = (q_1, q_2, \ldots, q_{k'}) \in \mathcal{Q}$ can be written as

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

Let $\mathcal{L}$ be a set of indices $\mathbf{l} = (\ell_1, \ell_2, \ldots, \ell_{k'})$ such that $\ell_i \in B_i$ for all $i = 1, 2, \ldots, k'$. 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 (5.45). Then, we have

$$T_{K_-(\mathcal{V})}(\theta) \cap B_n \subseteq \bigcup_{\substack{\mathbf{t}, \mathbf{q} \in \mathcal{Q}, \\ \mathbf{l} \in \mathcal{L}}} T(\mathbf{t}, \mathbf{q}, \mathbf{l}). \tag{5.47}$$

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

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

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

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

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

Next, by Lemma 5.28, there exist $\mathbf{l} = (\ell_1, \dots, \ell_{k'}) \in \mathcal{L}$ such that $\sum_{i=1}^{k'} \Gamma_i(v, \ell_i) \leq \gamma$. Hence, for any i, there exists an integer $1 \leq l_i \leq k'$ such that

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

Suppose $\gamma > 0$. Summing over $i = 1, 2, \dots, k'$, we have $\sum_{i=1}^{k'} l_i \leq 2k'$ 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 5.31. Let $\mathcal{Q}$ and $\mathcal{L}$ are the sets defined in Lemma 5.30. Then, we have:

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

Proof. For the first part, we have

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

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

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

□

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

$$\tilde{W}(\theta) := w(T_{K_-(\mathcal{V})}(\theta) \cap B_n) = \mathbb{E}\left[\sup_{v \in T_{K_-(\mathcal{V})}(\theta) \cap B_n} \langle v, Z \rangle \right], \tag{5.48}$$

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 5.28. Using the decomposition in Lemma 5.30, we have

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

Besides, leveraging a general result for Gaussian suprema (5.49 below), we have

$$\tilde{W}(\theta) \leq \max_{\mathbf{t},\mathbf{q}\in\mathcal{Q},\ \mathbf{l}\in\mathcal{L}} \mathbb{E}\left[\sup_{v\in T(\mathbf{t},\mathbf{q},\mathbf{l})} \langle v, Z\rangle\right] + 3\sqrt{k'\log\frac{\mathrm{e}n}{k'}} + \sqrt{\frac{\pi}{2}}. \tag{5.49}$$

Here, we used Proposition 5.31 to bound the cardinality of the set $\mathcal{Q}^2 \times \mathcal{L}$. More precisely, we used the following evaluation:

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

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 $v_{B_1}, v_{B_2}, \dots, v_{B_{k'}}$, this quantity is bounded from above as $\tilde{W}(\mathbf{t},\mathbf{q},\mathbf{l}) \leq \sum_{i=1}^{k'} \tilde{W}_i(t_i, q_i, \ell_i)$, where

$$\tilde{W}_i(t_i,q_i,\ell_i) := \mathbb{E}_{Z_i\sim N(0,I_{|B_i|})}\left[\sup_{v_{B_i}\in T_i(t_i,q_i,\ell_i)} \langle v_{B_i}, Z_i\rangle\right]. \tag{5.50}$$

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

We now consider the quantity (5.50). In the set $T_i(m_i,q_i,\ell_i)$ over which the supremum taken, the lower total variation of v_{B_i} is bounded from above as

$$\mathcal{V}_-^{B_i}(v_{B_i}) \leq w_i(v_{\tau_i} - v_\ell) + w_{i+1}(v_{\ell_i} - v_{\tau_{i+1}-1}) + q_i\gamma. \tag{5.51}$$

As mentioned in Remark 5.29, the reverse inequality

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

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 (5.51) is that v_{B_i} is approximated by two nearly monotone pieces. This suggests that the complexity of $T_i(m_i,q_i,\ell_i)$ can be evaluated by that of the class of monotone functions.

Below, we provide the upper bound of the Gaussian width of the form (5.50). First, the following lemma treats a special case where ℓ_i is taken as the rightmost point in B_i.

Lemma 5.32. For every $n\geq 1$, $t>0$, $w\in\{0,1\}$ and $\gamma\geq 0$, we have

$$\mathbb{E}\Biggl[\sup\Biggl\{\langle v, Z\rangle\ : v\in\mathbb{R}^n, \|v\|_2\leq t, \text{ and } \sum_{i=1}^{n-1}(v_i - v_{i+1})_+ \leq w(v_1 - v_n) + \gamma\Biggr\}\Biggr] \leq (t + 2\gamma\sqrt{n-1})\sqrt{\log(\mathrm{e}n)}. \tag{5.52}$$

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

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

We also define another monotone vector v^- as

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

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

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

and

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

Denote by $\tilde{W}$ the left-hand side in (5.52) with $t = 1$. The argument in the previous paragraph implies that

$$\begin{aligned}\tilde{W} &\le \mathbb{E}\left[\sup_{v^+ \in K_n^\uparrow:\ \|v^+\|_2 \le \gamma\sqrt{n-1}} \langle v^+, Z\rangle\right] + \mathbb{E}\left[\sup_{v^- \in K_n^\uparrow:\ \|v^-\|_2 \le 1+\gamma\sqrt{n-1}} \langle v^-, Z\rangle\right] \\ &\le (1 + 2\gamma\sqrt{n-1}) \cdot \mathbb{E}\left[\sup_{v \in K_n^\uparrow:\ \|v\|_2 \le 1} \langle v, Z\rangle\right]. \qquad (5.53)\end{aligned}$$

The expectation in the last line is bounded as

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

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 (5.53), 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 5.33. 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

$$\tilde{W} \le \begin{cases} (t + 2\gamma\sqrt{\ell-1})\sqrt{\log(\mathrm{e}\ell)} + (t + 2\gamma\sqrt{n-\ell})\sqrt{\log(\mathrm{e}(n-\ell+1))} & \text{if } 1 < \ell < n \\ (t + 2\gamma\sqrt{n-1})\sqrt{\log(\mathrm{e}n)} & \text{if } \ell = 1 \text{ or } n. \end{cases} \tag{5.54}$$

In particular, we deduce a simpler bound

$$\tilde{W} \le 2(t + 2\gamma\sqrt{n-1})\sqrt{\log(\mathrm{e}n)}. \tag{5.55}$$

Proof. Let (A_1, A_2) be a pair of sub-vectors of $[n]$ defined as $A_1 = \{1, 2, \ldots, \ell\}$ and $A_2 = \{\ell, \ell+1, \ldots, n\}$. 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 5.32.

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 5.32. □

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

Proposition 5.34. Fix $\theta \in \mathbb{R}^n$. Let $\Pi = (B_1, B_2, \ldots, B_{k'})$ be any connected refinement of $\Pi_{\text{const}}(\theta)$, and $w_1, w_2, \ldots, w_{k'}$ be the signs associated with Π as in Lemma 5.28. Define $\gamma \ge 0$ as (5.46). Then, the quantity $\tilde{W}(\theta)$ defined in (5.50) is bounded from above by

$$\tilde{W}(\theta) \le \max_{\mathbf{t}, \mathbf{q} \in \mathcal{Q}} \left\{\sum_{i=1}^{k'} 2(\sqrt{t_i} + 2q_i\gamma\sqrt{|B_i| - 1})\sqrt{\log(\mathrm{e}|B_i|)} + 3\sqrt{k' \log\frac{\mathrm{e}n}{k'}} + \sqrt{\frac{\pi}{2}}\right\}. \tag{5.56}$$

Proof. This is a direct consequence of (5.49) and (5.55). □

Step 4: Applying Lemma 5.24 We now are ready to complete the proof of Theorem 5.6.

Recall that our goal is to obtain an upper bound for $\tilde{W}(\theta)$ which is defined in (5.50). 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 (5.56). In fact, from an argument parallel to that in Guntuboyina et al. (2017), there exists a refinement $\Pi = (B_1, B_2, \ldots, B_{k'})$ such that

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

and $k(\theta) \leq k' \leq 2k(\theta)$. We also define the signs $w_1, w_2, \ldots, w_{k'}$ in a similar way as Lemma 5.27, 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 (5.56) as the following two steps. First, from the Cauchy–Schwarz inequality and the fact that $\mathbf{t} \in \mathcal{Q}$, we have

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

Second, by the above construction of Π, we have

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

Therefore, the right-hand side in (5.56) can be bounded from above by

$$10\sqrt{k(\theta)\log\frac{\mathrm{e}n}{k(\theta)}} + 20\gamma\sqrt{\frac{n}{k(\theta)}\log\frac{\mathrm{e}n}{k(\theta)}}. \tag{5.57}$$

Here, to hide the constant term $\sqrt{\pi/2}$, we have also used the fact that $\sqrt{m\log(\mathrm{e}n/m)} \geq 1$ for every integer $1 \leq m \leq n$.

Let $w_1^0, w_2^0, \ldots, w_{k(\theta)+1}^0$ be the signs associated with the constant partition $\Pi_{\mathrm{const}}(\theta) = (A_1, A_2, \ldots, A_{k(\theta)})$ (recall the definition (5.12)). Then, we can choose the values of w_i so that the following inequality holds:

$$\begin{aligned}
\gamma^2 &= \sum_{i=1}^{k'} |B_i|^{-1} 1_{\{w_i \neq w_{i+1}\}} \leq \sum_{j=1}^{k(\theta)} \left[\min\left\{|A_j|,\ \left\lfloor\frac{2n}{k(\theta)}\right\rfloor\right\}\right]^{-1} 1_{\{w_j^0 \neq w_{j+1}^0\}} \\
&\leq \sum_{i=1}^{k(\theta)} \left[\min\left\{|A_i|,\ \frac{n}{k(\theta)}\right\}\right]^{-1} 1_{\{w_i^0 \neq w_{i+1}^0\}} \\
&= M(\theta).
\end{aligned} \tag{5.58}$$

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 (5.58), (5.57) and Proposition 5.23, the statistical dimension of $T_{K_-(\mathcal{V})}(\theta)$ is bounded from above as

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

where we also used the elementary fact that $(a+b)^2 \leq 2(a^2+b^2)$. Consequently, applying Lemma 5.24, we have desired result.

5.9.3 Proof of Corollary 5.9

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 (5.12)) 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 (5.14), we have

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

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

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

From the minimal length condition (5.18) and the definition of $M(\theta)$, we also have

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

Combining the above inequalities, we have the desired result.

5.9.4 Risk bounds for penalized estimators (Proof of Theorem 5.12)

We prove Theorem 5.12 as an application of Lemma 5.25. 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. (2017). 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 5.23-(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:

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

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:

$$\mathbf{D}(\lambda\partial\mathcal{V}_-(\theta)) \leq \mathbf{D}(\mathrm{cone}(\partial\mathcal{V}_-(\theta))) + \Delta(\theta,\lambda), \tag{5.59}$$

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 chapter.

Proposition 5.35. 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 5.26.

Proof. Let us write $T := T_{K_-(\mathcal{V}(\theta))}(\theta)$. In the light of Proposition 5.23-(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 in Guntuboyina et al. (2017)). 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 (5.59). 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 5.22-(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 (5.59). To this end, we use the following result proved by Guntuboyina et al. (2017).

Lemma 5.36 (Guntuboyina et al. (2017), 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

$$v_0 := \operatorname*{argmin}_{v \in \mathrm{aff}(\partial f(\theta))} \|v\|_2, \tag{5.60}$$

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

$$\mathbf{D}(\lambda \partial f(\theta)) \leq 4 + \left(\sqrt{\mathbf{D}(\mathrm{cone}(\partial f(\theta)))} + \frac{4\|v^*\|_2}{\|v_0\|_2} + 2 + (\lambda - \lambda^*)\|v^*\|_2 \right)^2 . \tag{5.61}$$

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

$$\underset{v \in \mathrm{aff}(\partial f(\theta))}{\mathrm{argmin}} \ \|v\|_2 \in \partial f(\theta) \tag{5.62}$$

for every $\theta \in \mathbb{R}^n$. In other words, we can choose $v_0 \equiv v^*$ in (5.61) if f is weakly decomposable. Under the assumption that f is weakly decomposable, the inequality (5.61) can be simplified as follows:

Corollary 5.37. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and weakly decomposable. Under the same notation as in Lemma 5.36, we have

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

Now, we apply Lemma 5.36 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 (5.61). The proof is postponed to Appendix 5.9.6.

Proposition 5.38. (i) $\theta \mapsto \mathcal{V}_-(\theta)$ is weakly decomposable.

(ii) For any $\theta \in \mathbb{R}^n$, let us define v_0 as (5.60). Then, we have

$$\|v_0\|_2^2 = \sum_{i=1}^{k} \frac{1}{|A_i|} 1_{w_i \neq w_{i+1}}. \tag{5.63}$$

From Proposition 5.38 and Corollary 5.37, $\mathbf{D}(\lambda \partial \mathcal{V}_-(\theta))$ 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 $\lambda \geq \lambda^*$. Here, $C' > 0$ is a universal constant. Combining this bound with Lemma 5.25, 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 5.39. Suppose that $\theta \in \mathbb{R}^n$ and $\mathcal{V}_-(\theta) > 0$. For any $z \in \mathbb{R}^n$, define $\lambda(z)$ as

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

Then, we have

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

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 5.9.6, $\mathcal{V}_-$ 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

$$\lambda(z)\mathcal{V}_-(\theta) = \langle\theta, P_C(z)\rangle = \langle\theta, z - P_T(z)\rangle,$$

where $T := T_{K_-(\mathcal{V}_-(\theta)}(\theta)$ is the polar cone of C (see Proposition 5.35). Taking the expectation of both sides with respect to $z \sim N(0, I_n)$, we have

$$\mathcal{V}_-(\theta)\mathbb{E}[\lambda(z)] \le \|\theta\|_2\mathbb{E}\|P_T(z)\|_2 \le \|\theta\|_2(\mathbb{E}\|P_T(z)\|_2^2)^{1/2} = \|\theta\|_2(\delta(T))^{1/2},$$

which implies the desired result. Here, we used the fact that $\delta(T) = \mathbb{E}_{Z\sim N(0,I_n)}\|P_T(Z)\|_2^2$ (see Proposition 3.1 in Amelunxen et al. (2014)).

To prove the other inequality, we use the characterization of $\mathrm{aff}(\partial\mathcal{V}_-(\theta))$ given in (5.69) in Appendix 5.9.6 below. In particular, if we take v^* as in (5.72), we have

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

and

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

and hence the result follows. □

5.9.5 Proof of Corollary 5.16

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, which can be regarded as a special case of Lemma 2 in Bellec and Tsybakov (2015).

Lemma 5.40. 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

$$k = \max\left\{3, \left\lceil\left(\frac{\mathcal{V}^2 n}{\sigma^2\log(en)}\right)^{1/3}\right\rceil\right\}. \tag{5.64}$$

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

(ii) We have

$$\frac{1}{n}\|\theta - \theta'\|_2^2 \le \frac{1}{4}\max\left\{\left(\frac{\sigma^2\mathcal{V}\log(en)}{n}\right)^{2/3}, \frac{3\sigma^2\log(en)}{n}\right\} \tag{5.65}$$

and

$$\frac{\sigma^2 k}{n}\log\frac{en}{k} \le 2\max\left\{\left(\frac{\sigma^2\mathcal{V}\log(en)}{n}\right)^{2/3},\ \frac{3\sigma^2\log(en)}{n}\right\}. \tag{5.66}$$

(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 (5.64). 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 5.4 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 5.16. Applying Lemma 5.40 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 (5.18) 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 5.12 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 5.13.

5.9.6 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 5.9.4) of $\mathcal{V}_-$.

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

$$\mathcal{V}_-(\theta) = \max_{v\in\mathcal{B}}\langle v, \theta\rangle, \tag{5.67}$$

where $\mathcal{B} := \{v \in \mathbb{R}^n : \forall\theta \in \mathbb{R}^n,\ \langle v, \theta\rangle \le \mathcal{V}_-(\theta)\}$. 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 (5.29). For any subset $A \subseteq [n]$, we also define $\kappa_G(A)$ by the total weights of outgoing edges:

$$\kappa_G(A) := \sum_{(i,j)\in E:\ i\in A,\ j\notin A} c_{(i,j)}. \tag{5.68}$$

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

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

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

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

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 (5.29) is actually the Lovász extension of the cut function (5.68).

Proposition 5.41. 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). □

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

Proposition 5.42. 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 $(A_1, A_2, \dots, A_k) \subseteq [n]$ such that

$$\operatorname{aff}(\partial f(\theta)) = \left\{ v \in \mathbb{R}^n : \sum_{j \in S_i} v_j = F(S_i) \text{ for all } i = 1, 2, \dots, k \right\}, \tag{5.69}$$

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

(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:

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

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)). □

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 (5.68), 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 5.43. 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

$$\frac{|C|}{|A|}(F(B \cup A) - F(B)) \leq F(B \cup C) - F(B). \tag{5.70}$$

Recall the definition of weak decomposability (5.62). The following proposition provides a sufficient condition for the weak decomposability of the Lovász extension.

Proposition 5.44. Let $F : 2^{[n]} \to \mathbb{R}$ be a submodular function satisfying the AC condition in Definition 5.43. 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 $A_1, A_2, \dots, A_k$ be a partition of $[n]$ such that $\operatorname{aff}(\partial f(\theta))$ is represented as (5.69). For $i = 1, 2, \dots, k$, 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

$$v^* := \sum_{i=1}^{k} \frac{F(S_i) - F(S_{i-1})}{|A_i|} 1_{A_i}. \tag{5.71}$$

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 $v^* \in \operatorname{aff}(\partial f(\theta))$. Moreover, v^* is also contained in the normal cone of $\operatorname{aff}(\partial f(\theta))$. 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 $v^* \in \operatorname{argmin}_{v \in \partial f(\theta)} \|v\|_2^2$.

Now, our goal is to prove $v^* \in \partial f(\theta)$ under the AC condition. If $k = n$, then it is clear from (5.69) that $\partial f(\theta) = \{v^*\}$. Below, we assume that $k < n$. Since $v^* \in \operatorname{aff}(\partial f(\theta))$, it suffices to show that $\sum_{i \in S} v_i^* \leq F(S)$ 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 $|A_i| \geq 2$. 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 (5.69), which defines a part of the relative boundary of $\partial f(\theta)$. Therefore, we have to show that $\sum_{i \in S} v_i^* \leq F(S)$ 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

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

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

Remark 5.45. The AC condition was originally introduced in Bach (2011). In that paper, the author consider the proximal denoising estimators (5.34) 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 (5.34) 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. 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 5.44, the penalty of the one-dimensional fused lasso is weakly decomposable. This provides an alternative proof for Lemma 2.7 in Guntuboyina et al. (2017). 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 (5.3) 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 5.38.

Proposition 5.46. 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 (5.63).

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 5.1. Here, C represents any proper subset of A, and "None" means that A contains 1 or n.

Table 5.1: 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	$\geq 1_{\{C \neq \emptyset\}}$

In each case, we can easily check that the inequality (5.70) is satisfied. Hence, F satisfies the AC condition.

The second statement is a consequence of Proposition 5.44.

The last statement follows from fact that the minimizer of $\|v\|_2^2$ in $\partial f(\theta)$ coincides with that in $\operatorname{aff}(\partial f(\theta))$, which is given as (5.71). 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

$$v^* = \sum_{i=1}^{k} \frac{F(S_i) - F(S_{i-1})}{|A_i|} 1_{A_i} = \sum_{i=1}^{k} \frac{1_{w_i \neq w_{i+1}}}{|A_i|} 1_{A_i} \tag{5.72}$$

which proves the desired result. □

5.10 Proofs for Section 5.5

The goal of this section is to prove Theorem 5.17. 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 5.10.1. 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 5.10.2.

5.10.1 Proof overview

Let $(\hat{\Pi}, \hat{\mathbf{V}})$ be the selected pair in (5.26). Fix any connected partition Π and $\mathbf{V} \in \mathscr{V}(|\Pi|)$. By the definition of the estimator, we have

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

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

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

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

$$\|\theta^* - \hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}}\|_2^2 \le \|\theta^* - \theta^*_{\Pi',\mathbf{V}'}\|_2^2 - \mathrm{pen}(\hat{\Pi}, \hat{\mathbf{V}}) + \mathrm{pen}(\Pi', \mathbf{V}') + 2\langle \hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}} - \theta^*_{\Pi',\mathbf{V}'},\ \xi\rangle. \quad (5.73)$$

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

$$\max_{\Pi} \sup_{\mathbf{V} \in \mathscr{V}(|\Pi|)} \sup_{\theta \in K_{\Pi}^{\uparrow}(\mathbf{V})} \frac{\langle \theta - \theta^*_{\Pi',\mathbf{V}'},\ \xi\rangle}{(\|\theta - \theta^*\|_2 + \|\theta' - \theta^*\|_2)^2 + \eta(\Pi, \mathbf{V}, z)} \le \frac{c}{4} \quad (5.74)$$

holds on some event Ω_z that occurs with probability at least $1 - \mathrm{e}^{-z}$. Combining this inequality with (5.73), we have on the same event

$$(1-c)\|\theta^* - \hat{\theta}_{\hat{\Pi},\hat{\mathbf{V}}}\|_2^2 \le (1+c)\|\theta^* - \theta^*_{\Pi',\mathbf{V}'}\|_2^2 - \mathrm{pen}(\hat{\Pi}, \hat{\mathbf{V}}) + \mathrm{pen}(\Pi', \mathbf{V}') + c\eta(\hat{\Pi}, \hat{\mathbf{V}}, z), \quad (5.75)$$

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

5.10.2 Controlling the normalized process

Now, our goal is to provide an inequality of the form (5.74). Below, we fix $\theta' := \theta^*_{\Pi',\mathbf{V}'}$.

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

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

Note that $Z_{\Pi,\mathbf{V}}$ is the supremum of a sample-continuous Gaussian process. By the concentration inequality for Gaussian processes (Lemma 2.9), we have

$$\Pr\left\{Z_{\Pi,\mathbf{V}} - \mathbb{E}[Z_{\Pi,\mathbf{V}}] \geq \sqrt{2v(x+z)}\right\} \leq \exp(-(x+z)) \tag{5.76}$$

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

$$v := \sup_{\theta \in K_{\Pi}^{\uparrow}(\mathbf{V})} [Z_{\Pi,\mathbf{V}}^2] \leq \frac{\sigma^2}{4\eta}$$

because $\omega(\theta) \geq \|\theta - \theta'\|_2^2 + \eta \geq 2\eta^{1/2}\|\theta - \theta'\|_2$, 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 $K_{\Pi}^{\uparrow}(\mathbf{V})$. Note that

$$\mathbb{E}[Z_{\Pi,\mathbf{V}}] \leq \underbrace{\mathbb{E}\left[\sup_{\theta \in K_{\Pi}^{\uparrow}(\mathbf{V})} \frac{\langle \theta - \theta_{\Pi,\mathbf{V}}^*,\ \xi \rangle}{\omega(\theta)}\right]}_{\text{(a)}} + \underbrace{\mathbb{E}\left[\frac{|\langle \theta_{\Pi,\mathbf{V}}^* - \theta',\ \xi \rangle|}{\inf_{\theta \in K_{\Pi}^{\uparrow}(\mathbf{V})} \omega(\theta)}\right]}_{\text{(b)}}. \tag{5.77}$$

The second term (b) in the right-hand side of (5.77) is bounded from above by $\sigma\eta^{-1/2}$. Indeed, since

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

we have

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

To bound the term (a) in (5.77), we use the following lemma:

Lemma 5.47. Let $\Pi = (A_1, A_2, \ldots, A_m)$ be any partition and $\mathbf{V} = (\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m)$. Fix any $\bar{\theta} \in K_{\Pi}^{\uparrow}(\mathbf{V})$. For any $t > 0$, we have

$$\mathbb{E}\left[\sup_{\theta \in K_{\Pi}^{\uparrow}(\mathbf{V}) : \|\theta - \bar{\theta}\|_2 \leq t} \langle \xi, \theta - \bar{\theta} \rangle\right] \leq C\sigma t^{1/2}\left(\sum_{i=1}^m |A_i|^{1/3}\mathcal{V}_i^{2/3}\right)^{3/4} + C\sigma t\sqrt{m \log \frac{en}{m}}, \tag{5.78}$$

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

$$\mathbb{E}\left[\sup_{\theta \in K_{\Pi}^{\uparrow}(\mathbf{V})} \frac{\langle \theta - \bar{\theta},\ \xi \rangle}{\|\theta - \bar{\theta}\|_2 + \eta}\right] \leq 4C\sigma\left\{\eta^{-3/4}\left(\sum_{i=1}^m |A_i|^{1/3}\mathcal{V}_i^{2/3}\right)^{3/4} + \eta^{-1/2}\sqrt{m \log \frac{en}{m}}\right\}, \tag{5.79}$$

where C is the same constant as in (5.78).

Proof. We will prove the first inequality (5.78). Let $W := W(\Pi, \mathbf{V})$ denote the left-hand side of (5.78). We consider a collection of finitely many sets $S(\mathbf{q})$ as follows: Let $\mathcal{Q} := \mathcal{Q}(m)$ be a collection of vectors $\mathbf{q} = (q_1, q_2, \ldots, q_m)$ that can be written as $\mathbf{q} = t^2\mathbf{a}/m$ for some integer vector $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ such that $1 \leq a_i \leq m$ and $\sum_{i=1}^m a_i \leq 2m$. Note that, by Proposition 5.31, the cardinality of $\mathcal{Q}$ is bounded by $(2\mathrm{e})^m$. For any $\mathbf{q} \in \mathcal{Q}$, define the set

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

Then, we can easily check that

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

From Lemma 5.48 below, there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left[\sup_{\theta \in S(\mathbf{q})} \langle \theta, \xi \rangle\right] \leq C\sigma \sum_{i=1}^{m} \left\{\sqrt{2} q_i^{1/4} |A_i|^{1/4} \mathcal{V}_i^{1/2} + q_i^{1/2} \sqrt{\log \mathrm{e}|A_i|}\right\}. \tag{5.80}$$

Here, by Hölder's inequality, we have

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

and by the Cauchy-–Schwarz inequality, we also have

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

Then, by Lemma 5.49 below, we have

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

for some $C' > 0$. Thus, (5.78) has been proved.

The second inequality (5.79) is a consequence of the peeling lemma (Lemma 2.10 below).

□

Combining (5.76), (5.77) and (5.79), we conclude that

$$\begin{aligned} Z_{\Pi,\mathbf{V}} &\leq 4C\sigma\eta^{-3/4} \left(\sum_{i=1}^{m} |A_i|^{1/3} \mathcal{V}_i^{2/3}\right)^{3/4} \\ &\quad + \sigma\eta^{-1/2} \left\{4C \sqrt{m \log \frac{\mathrm{e}n}{m}} + (2\pi)^{-1/2} + 2^{-1/2} \sqrt{x + z}\right\} \end{aligned} \tag{5.81}$$

holds with probability at least $1 - \exp(-(x + z))$, where C is the constant in (5.79). 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 (5.81) 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

$$\sum_{\Pi}\sum_{\mathbf{V}} \exp(-x(\Pi,\mathbf{V})) \le 1, \tag{5.82}$$

and hence we conclude that (5.74) holds with $c = 1/2$. Indeed, (5.82) 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

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

5.10.3 Proof of Theorem 5.17

Now, we are ready to complete the proof of Theorem 5.17. Define $\mathrm{pen}(\Pi, \mathbf{V})$ as

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

where C is the constant in (5.79). 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 (5.75) and (5.74) 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

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

5.11 Proofs: Auxiliary lemmas

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

Lemma 5.48 (Guntuboyina et al. (2017), 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

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

Lemma 5.49 (Guntuboyina et al. (2017), Lemma D.1)**.** Suppose $p, n \geq 1$ and let $\Theta_1, \ldots, \Theta_p$ 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 $\xi \sim N(0, \sigma^2 I)$, we have

$$\mathbb{E}\left[\max_{1 \leq i \leq p} \sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle\right] \leq \max_{1 \leq i \leq p} \mathbb{E}\left[\sup_{\theta \in \Theta_i} \langle \xi, \theta \rangle\right] + D\sigma \left(\sqrt{2 \log p} + \sqrt{\frac{\pi}{2}}\right). \tag{5.83}$$

Chapter 6

Conclusions

In this thesis, we studied statistical properties of regularization based estimators inducing structured low-dimensionality.

In Chapter 4, we derived the degrees of freedom of two classes of regularization estimators related to submodular functions: the LEREs and the SNREs. Regardless of the choices of the design matrix and the underlying submodular function, we can derive simple and unified representations of the degrees of freedom. As mentioned in Section 4.5, our result recovers existing results derived for some particular regularization estimators or projection estimators. In fact, submodular regularization estimators can be regarded as special cases of anti-projection estimators with respect to certain classes of polyhedra, and general results for anti-projection estimators are still valid for our cases. We stress that our result provides "solution-dependent" representations of the degrees of freedom that can be calculated in low computational complexity.

In Chapter 5, 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 estimator 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.

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. 2017). 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 (5.28) 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 (2017).

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 5.5 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 (5.25).

Acknowledgment

First of all, I greatly thank the supervisor Fumiyasu Komaki for his entire support. When I was worried about which Ph.D. course to take, he strongly recommended me the Department of Mathematical Informatics. Looking back now, it was the right way for me. During my Ph.D. period at this department, I had a really fruitful time and great opportunities.

I acknowledge the members of the Ph.D. committee, Yoshihiro Kanno, Tomonari Sei, Taiji Suzuki, Kenji Yamanishi, for their time and variable feedback.

I thank all the members of Mathematical Informatics 4th Laboratory. I am grateful to all staffs, Fumiyasu Komaki, Tomonari Sei, Hiromichi Nagao, Yoshito Hirata, Yoshihiro Hirose, Keisuke Yano, Takeru Matsuda and Sumito Kurata. Especially, I am deeply grateful to Keisuke Yano and Takeru Matsuda for discussing a lot of things about research since I was an undergraduate student. I would like to thank Hiromichi Nagao for suggesting the example of a seismological phenomenon mentioned in Chapter 5. I also thank past students in the laboratory for sharing time with me. I am grateful to Yumiko Fujii for her daily support for various office work.

During my Ph.D. period, I learned a lot from daily discussions at voluntary reading groups. Especially, I am grateful to (in alphabetical order) Han Bao, Kazuto Fukuchi, Satoshi Hara, Kohei Hayashi, Junya Honda, Masaaki Imaizumi, Ryo Karakida, Jumpei Komiyama, Takuya Konishi, Atsushi Nitanda, Hirofumi Ohta, Kenta Oono, Sho Sonoda, and Takeshi Teshima. Also, I had some chances of collaboration with people outside the university. I thank Yuichi Yoshida for discussing problems related to submodular functions. I thank Masakazu Ishihata for discussing combinatorial testing problems during my short stay at Hokkaido University. I thank Naonori Kato, Yosuke Murakami and Hiromi Arai for co-authoring a presentation at a domestic conference on law.

I am currently working as a researcher at Preferred Networks, Inc (PFN). I thank my colleagues at PFN for daily research discussions and encouraging me to write this thesis outside my work hours.

I thank the Japan Society for the Promotion of Science (JSPS) for a grant that is necessary to complete this research.

Finally, I thank my family, Miki Minami and Mitsuki Minami, for all their support to my daily life.

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiadó.

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.

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.

F. Bach. Structured sparsity-inducing norms through submodular functions. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 118–126, USA, 2010. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=2997189.2997203`.

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

F. Bach. *Learning with Submodular Functions: A Convex Optimization Perspective*. Now Publishers Inc., Hanover, MA, USA, 2013. ISBN 1601987560, 9781601987563.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

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.

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

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.

P. C. Bellec and C.-H. Zhang. Second order stein: SURE for SURE and other applications in high-dimensional inference. arXiv:1811.04121, 2018. URL `https://arxiv.org/abs/1811.04121`.

P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

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

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 2015.

H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

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

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

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

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

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

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.

X. Chen, Q. Lin, and B. Sen. On degrees of freedom of projection estimators with applications to multivariate nonparametric regression. *Journal of the American Statistical Association*, 2019. doi: 10.1080/01621459.2018.1537917.

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

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

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.

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

B. Efron. The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *Journal of the American Statistical Association*, 99:619–642, 2004.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Z. Fan and L. Guan. Approximate l_0-penalized estimation of piecewise-constant signals on graphs. arXiv:1703.01421, 2017.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Statistics*, 1(2):302–332, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. URL `https://www.jstatsoft.org/v033/i01`.

S. Fujishige. *Submodular Functions and Optimizations*, volume 58 of *Annals of Discrete Mathematics*. Elsevier, 2nd edition, 2005.

G. Gallo, M. Grigoriadis, and R. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55, 1989.

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

C. Giraud. *Introduction to High-Dimensional Statistics*. CRC Press, 2015.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

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

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *arxiv preprint arXiv:1709.05707*, 2017.

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

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. arXiv:1708.09468, 2017.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

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

M. Hein, S. Setzer, L. Jost, and S. S. Rangapuram. The total variation on hypergraphs - learning on hypergraphs revisited. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 2427–2435, USA, 2013. Curran Associates Inc.

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

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

S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, 2009. ACM.

S. Jegelka, H. Lin, and J. A. Bilmes. On fast approximate submodular minimization. In *Advances in Neural Information Processing Systems*, pages 460–468, 2011.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

K. Kato. On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis*, 100(7):1338–1352, 2009.

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.

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

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.-C. Li. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross validation: discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

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.

J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. *Journal of Machine Learning Research*, 14:2449–2485, 2013.

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, 2011a.

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

C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

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

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

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. URL `https://doi.org/10.7283/P2D08S`.

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

K. Minami. Estimating piecewise monotone signals, 2019. URL `https://arxiv.org/abs/1905.01840`. Submitted to Electronic Journal of Statistics.

K. Minami. Degrees of freedom in submodular regularization: A computational perspective of Stein's unbiased risk estimate. *Journal of Multivariate Analysis*, 175:104546, 2020.

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.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

G. Obozinski and F. Bach. A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. Technical report, 2016. URL `https://hal-enpc.archives-ouvertes.fr/hal-01412385`.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

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.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

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

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1988.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, New Jersey, 1970.

R. T. Rockafeller and R. J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, Heidelberg, 1998.

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.

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

A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

W. Su and E. J. Candés. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

K. Takeuchi, Y. Kawahara, and T. Iwata. Higher order fused regularization for supervised learning with grouped parameters. In A. Appice, P. P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, and A. Jorge, editors, *Machine Learning and Knowledge Discovery in Databases*, ECML PKDD 2015, pages 577–593, Cham, 2015. Springer.

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

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.

R. J. Tibshirani and S. Rosset. Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*, 114(526): 697–712, 2019.

R. J. Tibshirani and J. Taylor. The solution path of generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

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

S. Vaiter, C. Deledalle, G. Peyré, J. M. Fadili, and C. Dossal. The degrees of freedom of the group lasso, 2012. Technical report.

S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2009.

S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1):72–86, 2014.

S. van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV - 2015*, volume 2159 of *Lecture Notes in Mathematics.* Springer, 2015.

C. van Eeden. Maximum likelihood estimation of ordered probabilities. In *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings. Series A*, volume 59, pages 444–455, 1956.

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

M. J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

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

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

Y. Yang. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society: Series B*, 68(1):49–67, 2006.

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

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2014.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.