Electronic Journal of Statistics
Vol. 12 (2018) 4517–4570
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1519

# On the total variation regularized estimator over a class of tree graphs

**Francesco Ortelli and Sara van de Geer**

*Rämistrasse 101*
*8092 Zürich*
*e-mail:* francesco.ortelli@stat.math.ethz.ch; *e-mail:* geer@stat.math.ethz.ch

**Abstract:** We generalize to tree graphs obtained by connecting path graphs an oracle result obtained for the Fused Lasso over the path graph. Moreover we show that it is possible to substitute in the oracle inequality the minimum of the distances between jumps by their harmonic mean. In doing so we prove a lower bound on the compatibility constant for the total variation penalty. Our analysis leverages insights obtained for the path graph with one branch to understand the case of more general tree graphs. As a side result, we get insights into the irrepresentable condition for such tree graphs.

**Keywords and phrases:** Total variation regularization, lasso, fused Lasso, edge Lasso, path graph, branched path graph, tree, compatibility constant, oracle inequality, irrepresentable condition, harmonic mean.

Received June 2018.

## Contents

## 1. Introduction

The aim of this paper is to refine and extend to the more general case of a class of tree graphs the approach used by Dalalyan, Hebiri and Lederer (2017) to prove an oracle inequality for the Fused Lasso estimator, also known as total variation regularized estimator. As a side result, we will obtain some insight into the irrepresentable condition for signal pattern recovery over tree graphs in that class.

The main reference of this article is Dalalyan, Hebiri and Lederer (2017), who consider the path graph. We refine and generalize their approach (i.e. their Theorem 3, Proposition 2 and Proposition 3) to the case of more general tree graphs. The main refinements we prove are an oracle theory for the total variation regularized estimators over trees when the first coefficient is not penalized, a proof of an (in principle tight) lower bound for the compatibility constant and, as a consequence of this bound, the substitution in the oracle bound of the minimum of the distances between jumps by their harmonic mean. We elaborate the theory from the particular case of the path graph to the more general case of tree graphs which can be cut into path graphs. The tree graph with one branch is in this context the simplest instance of such more complex tree graphs, which allows us to develop insights into more general cases, while keeping the overview.

The paper is organized as follows: in Section 1 we expose the framework together with a review of the literature on the topic; in Section 2 we refine the proof of Theorem 3 of Dalalyan, Hebiri and Lederer (2017) and adapt it to the case where one coefficient of the Lasso is left unpenalized: this proof will be a working tool for establishing oracle inequalities for total variation penalized estimators; in Section 3 we introduce the notation needed for the rest of the article; in Section 4 we expose how to easily compute objects related to projections which are needed for finding explicit bounds on weighted compatibility constants and when the irrepresentable condition is satisfied; in Section 5 we present a tight lower bound for the (weighted) compatibility constant for the Fused Lasso and use it with the approach exposed in Section 2 to prove an oracle inequality; in Section 6 we generalize Section 5 to the case of the branched path graph; Section 7 presents further extensions to more general tree graphs; Section 8 handles the asymptotic signal pattern recovery properties of the total variation regularized estimator on the (branched) path graph and exposes an extension to more general tree graphs; Section 9 concludes the paper.

### *1.1. General framework*

We study total variation regularized estimators on graphs, their oracle properties and their asymptotic signal pattern recovery properties.

For a vector $v \in \mathbb{R}^n$ we write $\|v\|_1 = \sum_{i=1}^n |v_i|$ and $\|v\|_n^2 = \frac{1}{n}\sum_{i=1}^n v_i^2$.

Let $\mathcal{G} = (V, E)$ be a graph, where $V$ is the set of vertices and $E$ is the set of edges. Let $n := |V|$ be its number of vertices and $m := |E|$ its number of edges. Let the elements of $E$ be denoted by $e(i, j)$, where $i, j \in V$ are the vertices connected by an edge.

Let $D_{\mathcal{G}} \in \mathbb{R}^{m \times n}$ denote the **incidence matrix** of a graph $\mathcal{G}$, defined as

$$(D_e)_k = \begin{cases} -1, & \text{if } k = \min(i, j) \\ +1, & \text{if } k = \max(i, j) \\ 0, & \text{else,} \end{cases}$$

where $D_e \in \mathbb{R}^n$ is the row of $D_{\mathcal{G}}$ corresponding to the edge $e(i, j)$.

Let $f \in \mathbb{R}^n$ be a function defined at each vertex of the graph. The **total variation** of $f$ over the graph $\mathcal{G}$ is defined as

$$\mathrm{TV}_{\mathcal{G}}(f) := \|D_{\mathcal{G}} f\|_1 = \sum_{e(i,j)\in E} |f_j - f_i|.$$

Assume we observe the values of a signal $f^0 \in \mathbb{R}^n$ contaminated with some Gaussian noise $\epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{I}_n)$, i.e. $Y = f^0 + \epsilon$. The **total variation regularized estimator** $\widehat{f}$ of $f^0$ over the graph $\mathcal{G}$ is defined as

$$\widehat{f} := \arg\min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_n^2 + 2\lambda \|D_{\mathcal{G}} f\|_1 \right\},$$

where $\lambda > 0$ is a tuning parameter. This is a special case of the generalized Lasso with design matrix $\mathrm{I}_n$ and penalty matrix $D_{\mathcal{G}}$. Hereafter we suppress the subscript $\mathcal{G}$ in the notation of the incidence matrix of the graph $\mathcal{G}$.

In this article, we restrict our attention to tree graphs, i.e. connected graphs with $m = n - 1$. For a tree graph we have that $D \in \mathbb{R}^{(n-1)\times n}$ and $\mathrm{rank}(D) = n-1$. In order to manipulate the above problem to obtain an (almost) ordinary Lasso problem, we define $\widetilde{D}$, the **incidence matrix rooted at vertex** $i$, as

$$\widetilde{D} := \begin{bmatrix} A \\ D \end{bmatrix} \in \mathbb{R}^{n\times n},$$

where

$$A = (0, \dots, 0, \underbrace{1}_{i}, 0, \dots, 0) \in \mathbb{R}^n.$$

In the following, we are going to root the incidence matrix at the vertex $i = 1$, obtaining in this way a lower triangular matrix with ones on the diagonal, and minus ones as nonzero off-diagonal elements. The square matrix $\widetilde{D}$ is invertible and we denote its inverse by $X := \widetilde{D}^{-1}$.

We now perform a change of variables. Let $\beta := \widetilde{D} f$, then $f = X\beta$. The above problem can be rewritten as

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^n} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \sum_{i=2}^{n} |\beta_i| \right\},$$

i.e. an ordinary Lasso problem with $p = n$, where the first coefficient $\beta_1$ is not penalized. Note that, in order to perform this transformation, it is necessary that we restrict ourselves to tree graphs, since we want $\widetilde{D}$ to be invertible.

Let $X = (X_1, X_{-1})$, where $X_1 \in \mathbb{R}^n$ denotes the first column of $X$ and $X_{-1} \in \mathbb{R}^{n\times(n-1)}$ the remaining $n-1$ columns of $X$. Let $\beta_{-1} \in \mathbb{R}^{n-1}$ be the vector $\beta$ with the first entry removed. Thanks to some easy calculations and denoting by $\widetilde{Y}$ and $\widetilde{X}_{-1}$ the column centered versions of $Y$ and $X_{-1}$, it is possible to write

$$\widehat{\beta}_{-1} = \arg\min_{\beta_{-1}\in\mathbb{R}^{n-1}} \left\{ \|\widetilde{Y} - \widetilde{X}_{-1}\beta_{-1}\|_n^2 + 2\lambda\|\beta_{-1}\|_1 \right\}$$

and

$$\widehat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i - (X_{-1})_i \widehat{\beta}_{-1},$$

and both $\widehat{\beta}_{-1}$ and $\widehat{\beta}_1$ depend on $\lambda$.

Note that prediction properties of $\widehat{\beta}$, i.e. the properties of $X\widehat{\beta}$, will translate into properties of the estimator $\widehat{f}$, often also called Edge Lasso estimator.

**Remark.** In the construction of an invertible matrix starting from $D$, it would be possible to choose $A = (1, \dots, 1) =: 1_n \in \mathbb{R}^n$ as well. Indeed, when we

perform the change of variables from $f$ to $\beta$, $\hat{\beta}_{-1}$ estimates the differences of the signal across the edges of $\mathcal{G}$ and thus gives information about the relative location of the signal. However to be able to estimate the absolute location of the signal we either need an estimate of the absolute location of the signal at one point (choice $A = (0, \dots, 0, 1, 0, \dots, 0)$, $\hat{\beta}_1 = \hat{f}_i$, in particular we consider the case $i = 1$), or of the "mean" location of the signal (choice $A = (1, \dots, 1) = 1_n$, $\hat{\beta}_1 = \sum_{i=1}^n \hat{f}_i$).

### *1.2. The path graph and the path graph with one branch*

In this article we are interested, besides the more general case of tree graphs, in the particular cases of $D$ being the incidence matrix of either the path graph or the path graph with one branch. The choice of $A$ makes it easy to calculate the matrix $X$ and gives a nice interpretation of it.

Let $P_1$ be the **path matrix** of the graph $\mathcal{G}$ with reference root the vertex 1. The matrix $P_1$ is constructed as follows:

$$(P_1)_{ij} := \begin{cases} 1, & \text{if the vertex } j \text{ is on the path from vertex 1 to vertex } i, \\ 0, & \text{else.} \end{cases}$$

**Theorem 1.1** (Inversion of the rooted incidence matrix)**.** *For a tree graph, the rooted incidence matrix $\widetilde{D}$ is invertible and*

$$X = \widetilde{D}^{-1} = P_1.$$

*Proof of Theorem 1.1.* For a formal proof we refer to Jacobs et al. (2008) and to Bapat (2014). The intuition behind this theorem is to proceed as follows. We have to check that $\operatorname{rank}(\widetilde{D}) = n$. One can perform Gaussian elimination on the rooted incidence matrix. Keep the first row as it is and for row $i$ add up the rows indexed by the vertices belonging to the path going from vertex 1 to vertex $i$. In this way one can obtain an identity matrix and thus $\operatorname{rank}(\widetilde{D}) = n$. Similarly one can find the inverse, which obviously corresponds to $P_1$. □

**Example 1.2** (Incidence matrix and path matrix with reference vertex 1 for the path graph)**.** Let $\mathcal{G}$ be the path graph with $n = 6$ vertices. The incidence matrix is

$$D = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{5 \times 6}$$

and the path matrix with reference vertex 1 is

$$X = \begin{pmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & 1 & 1 & & & \\ 1 & 1 & 1 & 1 & & \\ 1 & 1 & 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{6\times 6}.$$

**Example 1.3** (Incidence matrix and path matrix with reference vertex 1 for the path graph with one branch)**.** Let $\mathcal{G}$ be the path graph with one branch. The graph has in total $n = n_1 + n_2$ vertices. The main branch consists in $n_1$ vertices, the side branch in $n_2$ vertices and is attached to the vertex number $b < n_1$ of the main branch. Take $n_1 = 4$, $n_2 = 2$ and $b = 2$. The incidence matrix is

$$D = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & -1 & & & 1 & \\ & & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{5\times 6}$$

and the path matrix with reference vertex 1 is

$$X = \begin{pmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & 1 & 1 & & & \\ 1 & 1 & 1 & 1 & & \\ 1 & 1 & & & 1 & \\ 1 & 1 & & & 1 & 1 \end{pmatrix} \in \mathbb{R}^{6\times 6}.$$

### *1.3. Review of the literature*

While to our knowledge there is no attempt in the literature to analyze the specific properties of the total variation regularized least squares estimator over general branched tree graphs, there is a lot of work in the field of the so called Fused Lasso estimator. An early analysis of the Fused Lasso estimator can be found in Mammen and van de Geer (1997). Some other early work is exposed in Tibshirani et al. (2005); Friedman et al. (2007); Tibshirani and Taylor (2011), where also computational aspects are considered.

In the literature we can find two main currents of research, the one focusing on the pattern recovery properties (which is going to be briefly exposed in Section 8) and the other on the analysis of the mean squared error to prove oracle inequalities.

### *1.3.1. Minimax rates*

In this subsection we expose some results on minimax rates, making use of the notation found in Sadhanala, Wang and Tibshirani (2016). In particular, let

$$\mathcal{T}(C) = \{f \in \mathbb{R}^n : \|Df\|_1 \leq C\}$$

be the class of (discrete) functions of bounded total variation on the path graph, where $D$ is its incidence matrix. Assume the linear model with $f^0 \in \mathcal{T}(C)$ for some $C > 0$ and with iid Gaussian noise with variance $\sigma \in (0, \infty)$ . It has been shown in Donoho and Johnstone (1998) that the minimax risk over the class of functions with bounded total variation $\mathcal{R}(\mathcal{T}(C))$ satisfies

$$\mathcal{R}(\mathcal{T}(C)) := \inf_{\widehat{f}} \sup_{f^0 \in \mathcal{T}(C)} \mathbb{E}[\|\widehat{f} - f^0\|_n^2] \asymp (C/n)^{2/3}.$$

Mammen and van de Geer (1997) prove that, if $\lambda \asymp n^{-2/3}C^{1/3}$, then the Fused Lasso estimator achieves the minimax rate within the class $\mathcal{T}(C)$. Sadhanala, Wang and Tibshirani (2016) also point out, that estimators which are linear in the observations can not achieve the minimax rate within the class of functions of bounded total variation, since they are not able to adapt to the spatially inhomogeneous smoothness of some elements of this class.

### *1.3.2. Oracle inequalities*

We expose some recent results, appeared in the papers by Hütter and Rigollet (2016); Dalalyan, Hebiri and Lederer (2017); Lin et al. (2017); Guntuboyina et al. (2017). In particular we give the rates of the remainder term in the (sharp) oracle inequalities holding with high probability exposed in these papers.

- **Hütter and Rigollet (2016)** obtain a quite general result, in the sense that it applies to any graph $\mathcal{G}$ with incidence matrix $D \in \mathbb{R}^{m \times n}$. In particular for the choice of the tuning parameter $\lambda = \sigma\rho\sqrt{2\log(em/\delta)}/n, \delta \in (0, \frac{1}{2})$, they obtain that the remainder term in their oracle inequality has the rate

  $$\mathcal{O}\left(\frac{|S|\rho^2}{n\kappa_D^2(S)}\log(em/\delta)\right),$$

  with probability at least $1 - 2\delta$, where, for a set $S \subseteq [m]$,

  $$\kappa_D(S) := \inf_{f \in \mathbb{R}^n} \frac{\sqrt{|S|}\|f\|_2}{\|(Df)_S\|_1}, S \neq \emptyset$$

  is called **compatibility factor** and $\rho$ is the largest $\ell^2$-norm of a column of the Moore-Penrose pseudoinverse $D^+ = (\delta_1^+, \ldots, \delta_m^+) \in \mathbb{R}^{n \times m}$ of the incidence matrix $D$, i.e. $\rho = \max_{j \in [m]} \|\delta_j^+\|_2$, and is called **inverse scaling factor**.
  For the path graph, we have $m = n - 1$, $\rho \asymp \sqrt{n}$ and, according to Lemma 3 in Hütter and Rigollet (2016), $\kappa_D(S) = \Omega(1)$, if $|S| \geq 2$.

- **Dalalyan, Hebiri and Lederer (2017)** obtain that, $\forall S \neq \emptyset$, for $\delta \in (0, \frac{1}{2})$ and the choice of the tuning parameter $\lambda := 2\sigma\sqrt{2\log(n/\delta)/n}$, the remainder term has rate

$$\mathcal{O}\left(\frac{s\log(n/\delta)}{n}\left(\log n + \frac{n}{W_{\min,S}}\right)\right),$$

with probability $1-2\delta$, where $S = \{i_1, \dots, i_s\}$, $s = |S|$, $W_{\min,S} := \min_{1\leq j\leq s+1}|i_j - i_{j-1}|$, with the convention $i_0 = 1$ and $i_{s+1} = n+1$.
- **Lin et al. (2017)** prove a result similar to the one of Dalalyan, Hebiri and Lederer (2017) using a technique that they call lower interpolant. Their result states that the mean squared error of the Fused Lasso estimator with the choice of the tuning patameter $\lambda = n^{-\frac{3}{4}}W_{\min,S_0}^{\frac{1}{4}}$, for $n$ large enough, has error rate

$$\mathcal{O}\left(\gamma^2\frac{s_0}{n}\left((\log s_0 + \log\log n)\log n + \sqrt{\frac{n}{W_{\min,S_0}}}\right)\right),$$

with probability at least $1 - e^{-C\gamma}$, where $C > 0$ is a constant only depending on $\sigma$ and where $\gamma > 1$.
- **Guntuboyina et al. (2017)** consider the sequence of estimators $\{\widehat{f}_\lambda, \lambda \geq 0\}$, where

$$\widehat{f}_\lambda = \arg\min_{f\in\mathbb{R}^n}\left\{\|Y-f\|_2^2 + 2\sigma\lambda\|Df\|_1\right\},$$

and prove that, when the minimum length condition $W_{\min,S_0} \geq \frac{c_1 n}{s_0+1}, 0 < c_1 \leq 1$, is satisfied, then

$$\inf_{\lambda\geq 0}\mathbb{E}\left[\|\widehat{f}_\lambda - f^0\|_n^2\right] = \mathcal{O}\left(\frac{s_0+1}{n}\log\left(\frac{ne}{s_0+1}\right)\right),$$

where the value $\lambda^0$ at which the infimum is reached depends on $f^0$, (see Corollary 2.8 in Guntuboyina et al. (2017)).
Moreover Guntuboyina et al. (2017) prove, in a newer version of their article, that under the minimum length condition and the maximum length condition, i.e. $\max_{1\leq j\leq s_0+1}|i_j - i_{j-1}| \leq \frac{c_2 n}{s_0+1}, c_2 \geq 1$ with the convention $i_0 = 1$ and $i_{s+1} = n+1$, if $\lambda \asymp C(c_1, c_2)\sqrt{\frac{\log n}{n}}$, then

$$\mathbb{E}\left[\|\widehat{f}_\lambda - f^0\|_n^2\right] = \mathcal{O}\left(\frac{(s_0+1)^2}{n}\log n\right),$$

where $C(c_1, c_2)$ is a constant depending on $c_1$ and $c_2$. Since $c_1$ and $c_2$ depend on $f^0$, so does implicitely the choice of the tuning parameter.
It is worth noticing that the results by Guntuboyina et al. (2017) hold also with a weaker version of the minimum length condition which involves only piecewise constant regions between jumps of opposite signs. However for the ease and simplicity of exposition we exposed the results using the stronger condition involving all the piecewise constant regions. Moreover,

thanks to concentration results for penalized least squares estimators, high probability bounds for the mean squared error are derived by the bounds on its expectation.

## 2. Approach for general tree graphs

The approach we follow is very similar to the one presented in the proof of Theorem 3 of Dalalyan, Hebiri and Lederer (2017). However, we refine their proof by not penalizing the first coefficient of $\beta$ and by adjusting the definition of compatibility constant accordingly. Note that by not penalizing the first coefficient we allow it to be always active. This is a more natural approach to utilize, considering our problem definition.

Let $\beta \in \mathbb{R}^n$ be a vector of coefficients, $S \subseteq \{2, \dots, n\}$ a subset of the indices of $\beta$, called active set with $s := |S|$ being its cardinality.

Let $v \in \mathbb{R}^n$ and $T \subseteq [n]$ be a subset of indices of $v$. By $v_T$ we denote the vector satisfying

$$\begin{cases} (v_T)_i = v_i & ,\forall i \in T, \\ (v_T)_i = 0 & ,\forall i \notin T, \end{cases}$$

and by $v_{-T}$ we denote the vector satisfying

$$\begin{cases} (v_{-T})_i = 0 & ,\forall i \in T, \\ (v_{-T})_i = v_i & ,\forall i \notin T, \end{cases}$$

giving place to the equation $v = v_T + v_{-T}$.

**Definition 2.1** (**Compatibility constant**). *The compatibility constant $\kappa(S)$ is defined as*

$$\kappa^2(S) := \min\left\{(s+1)\|X\beta\|_n^2 : \|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1 = 1\right\}.$$

Let $V_{\{1\}\cup S}$ denote the linear subspace of $\mathbb{R}^n$ spanned by the columns of $X$ with index in $\{1\} \cup S$. Let $\Pi_{\{1\}\cup S}$ be the orthogonal projection matrix onto $V_{\{1\}\cup S}$. We have that $\Pi_{\{1\}\cup S} = X_{\{1\}\cup S}(X'_{\{1\}\cup S}X_{\{1\}\cup S})^{-1}X'_{\{1\}\cup S}$ and that $A_{\{1\}\cup S} = \mathrm{I}_n - \Pi_{\{1\}\cup S}$, where $\mathrm{I}_n$ denotes the $n \times n$ identity matrix.

**Definition 2.2.** *The vector $\omega \in \mathbb{R}^n$ is defined as*

$$\omega_j = \|A_{\{1\}\cup S}X_j\|_n, \forall j \in [n].$$

**Remark.** Note that $\omega_{\{1\}\cup S} = 0$ and $0 \leq \omega \leq 1$, since for tree graphs the maximum $\|\cdot\|_n$-norm of a column of $X$ is 1.

**Definition 2.3.** *Take $\gamma > 1$. The vector of weights $w \in \mathbb{R}^n$ is defined as*

$$w_j = 1 - \frac{\omega_j}{\gamma}, \forall j \in [n].$$

**Remark.** Note that $0 \leq w \leq 1$ and that $w_{\{1\}\cup S} = 1$.

For two vectors $a, b \in \mathbb{R}^k$, $a \odot b := (a_1 b_1, a_2 b_2, \ldots, a_k b_k)'$.

**Definition 2.4** (**Weighted compatibility constant**). *The weighted compatibility constant $\kappa_w(S)$ is defined as*

$$\kappa_w^2(S) := \min\left\{(s+1)\|X\beta\|_n^2 : \|\beta_S\|_1 - \|(w \odot \beta)_{-(\{1\}\cup S)}\|_1 = 1\right\}.$$

**Remark.** Note that the (weighted) compatibility constant depends on the graph through $X$, which is the path matrix of the graph rooted at the vertex 1.

**Remark.** Note that a key point in our approach is the computation of a lower bound for the compatibility constant over the path graph, which is shown to be tight in some special cases. The concept of compatibility constant for total variation estimators over graphs is already presented in Hütter and Rigollet (2016). However, we refer to the (different) definition given in Dalalyan, Hebiri and Lederer (2017), which we slightly modify to adapt it to our problem definition.

**Theorem 2.5** (Oracle inequality for total variation regularized estimators over tree graphs). *Fix $\delta \in (0,1)$ and $\gamma > 1$. Choose $\lambda = \gamma\sigma\sqrt{2\log\left(4(n-s-1)/\delta\right)/n}$. Then, with probability at least $1-\delta$, it holds that*

$$\begin{aligned}\|\widehat{f} - f^0\|_n^2 &\leq \inf_{f\in\mathbb{R}^n}\left\{\|f - f^0\|_n^2 + 4\lambda\|(Df)_{-S}\|_1\right\} \\ &+ \frac{4\sigma^2}{n}\left((s+1) + 2\log(2/\delta) + \frac{\gamma^2(s+1)}{\kappa_w^2(S)}\log\left(4(n-s-1)/\delta\right)\right).\end{aligned}$$

*Proof of Theorem 2.5.* See Appendix A. □

## 3. Notation

Here we expose the notational conventions used for handling the (branched) path graph and later branching points with arbitrarily many ($K$) branches.

### 3.1. Path graph

In the case of the path graph we write the candidate set of active edges $S$, which has cardinality $s$, as

$$S = \{d_1 + 1, d_1 + d_2 + 1, \ldots, d_1 + d_2 + \ldots + d_s + 1\}$$

and we define $d_{s+1} = n - \sum_{i=1}^{s} d_i$. In the next subsection on the branched path graph we are going to introduce some dual notation, because we will have to use different notations depending on the context. It is worth mentioning that in the case of the path graph this is not necessary and we always can utilize the same notation. Indeed, the need of the dual notation is due to the different possible ways to decompose the branched path graph into three smaller path graphs. In the case of the path graph it is evident that this problem is not present.

### *3.2. (Branched) path graph*

We decide to enumerate the vertices of the (branched) path graph starting from the root 1, continuing up to the end of the main branch $n_1$ and then continuing from the vertex $n_1 + 1$ of the side branch attached to vertex $b$ up to the last vertex of the side branch $n = n_1 + n_2$.

We are going to use two different notations: the one is going to be used for finding explicit expressions for quantities related to the projection of a column of $X$ onto some subsets of the columns of $X$. The other is going to be used when calculating the compatibility constant and is based on the decomposition of the (branched) path graph into smaller path graphs. In both notations we let the set $S \subseteq \{2, \dots, n\}$ be a candidate set of active edges.

**First notation** (for calculating projection coefficients and lengths of antiprojections).

We partition $S$ into three mutually disjoint sets $S_1, S_2, S_3$, where $S_1 = \{2, \dots, b\} \cap S$, $S_2 = \{b+1, \dots, n_1\} \cap S$, $S_3 = \{n_1 + 1, \dots, n\} \cap S$. We write the sets $S_1, S_2, S_3$ as:

$$S_1 =: \{i_1, \dots, i_{s_1}\}, S_2 =: \{j_1, \dots, j_{s_2}\}, S_3 =: \{k_1, \dots, k_{s_3}\}.$$

We write $s_i := |S_i|, i \in \{1, 2, 3\}$. Note that $s := |S| = s_1 + s_2 + s_3$.

Let us write $S = \{\xi_1, \dots, \xi_{s_1+s_2+s_3}\}$. Define

$$\begin{aligned} B \;:=\; & \{\xi_1 - 1, \xi_2 - \xi_1, \dots, \xi_{s_1} - \xi_{s_1-1}, b - \xi_{s_1} + 1, \\ & \xi_{s_1+1} - b - 1, \xi_{s_1+2} - \xi_{s_1+1}, \dots, \xi_{s_1+s_2} - \xi_{s_1+s_2-1}, n_1 - \xi_{s_1+s_2} + 1, \\ & \xi_{s_1+s_2+1} - n_1 - 1, \xi_{s_1+s_2+2} - \xi_{s_1+s_2+1}, \dots, n - \xi_{s_1+s_2+s_3} + 1\} \\ =:\; & \{d_1^1, d_2^1, \dots, d_{s_1}^1, \tilde{d}_{s_1+1}^1, \tilde{d}_1^2, d_2^2, \dots, d_{s_2}^2, d_{s_2+1}^2, \\ & \tilde{d}_1^3, d_2^3, \dots, d_{s_3+1}^3\}. \end{aligned}$$

Define $d^* := \tilde{d}_{s_1+1}^1 + \tilde{d}_1^2 + \tilde{d}_1^3$.

This notation implicitely means that we cut the branched path graph into three subgraphs, all of them being paths. These three subgraphs are obtained by cutting the edges $(b, b+1)$ and $(b, n_1 + 1)$. The set $B$ is the set containing the information about how many vertices are between consecutive jumps, i.e. the length of the piecewise constant regions that a signal would have if its true active set were $S$. Since the $\xi_i, i \in [s_1 + s_2 + s_3]$ tell us which edges are in the active set, at the beginning and at the end of each of the subgraphs obtained as explained here above we need to subtract, respectively add, a vertex. Indeed, say that the first element of $S$ is $\xi_1 = 5$. This means that we have a candidate jump on the edge $(4, 5)$ and thus the first candidate piecewise constant region is constituted by the vertices $\{1, 2, 3, 4\}$ and $d_1^1 = \xi_1 - 1 = 4$. Analogously suppose that $\xi_{s_1} = 15$ and $b = 18$. This means that the last candidate active edge of $S_1$ is $(14, 15)$ and thus the last candidate piecewise constant region is $\{15, 16, 17, 18\}$. It follows that $\tilde{d}_{s_1+1}^1 = b - \xi_{s_1} + 1 = 18 - 15 + 1$.

It is thus clear that the three sets $\{\tilde{d}_1^1, \ldots, d_{s_1}^1, d_{s_1+1}^1\}$, $\{\tilde{d}_1^2, d_2^2, \ldots, d_{s_2+1}^2\}$ and $\{\tilde{d}_1^3, d_2^3, \ldots, d_{s_2+1}^3\}$ contain the information on the length of the candidate piecewise constant regoins inside the three subgraphs obtained by cutting the edges $(b, b+1)$ and $(b, n_1+1)$, when we suppose that $S = S_1 \cup S_2 \cup S_3$ is the candidate active set of edges.

**Second notation** (for bounding the compatibility constant).

What is meant with the second notation is that we decompose the branched path graph into three smaller path graphs. However, the end of the first one does not necessarily coincide with the point $b$ and the beginning of the other two does not necessarily coincide with the points $b+1$ and $n_1+1$ respectively, i.e. the three path graphs resulting from this operation are not necessarily obtained by cutting the edges $(b, b+1)$ and $(b, n_1+1)$.

Let us write

$$S_1 = \{d_1^1 + 1, d_1^1 + d_2^1 + 1, \ldots, d_1^1 + d_2^1 + \ldots + d_{s_1}^1 + 1\} = S \cap \{1, \ldots, b\},$$

and

$$S_i = \{p_i + 1, p_i + d_2^i + 1, p_i + d_2^i + d_3^i + 1, \ldots, p_i + d_2^i + d_3^i + \ldots + d_{s_i}^i + 1\}, i = 2, 3,$$

where, using the first notation introduced, $p_2 = j_1 - 1$, $p_3 = k_1 - 1$, $d_{s_2+1}^2 = n_1 - \xi_{s_1+s_2} + 1$ and $d_{s_3+1}^3 = n - \xi_{s_1+s_2+s_3} + 1$. Note that $d^* = d_{s_1+1}^1 + d_1^2 + d_1^3 = \tilde{d}_{s_1+1}^1 + \tilde{d}_1^2 + \tilde{d}_1^3$.

The second notation differs from the first one only in regard to which edges in the proximity of the branching point are cut to obtain a decomposition into three path graphs. The second notation reflects a more flexible choice, where the end of the first path graph does not have to coincide with the branching point $b$. It is clear that the only difference with the first notation is in the length of the piecewise constant pieces of the three subgraphs: $d_{s_1+1}^1, d_1^2, d_1^3$ are not necessarily equal to $\tilde{d}_{s_1+1}^1, \tilde{d}_1^2, \tilde{d}_1^3$. We decide to keep the notation without tilde for the situation where we have to bound the compatibility constant to remain coherent with the notation in van de Geer (2018).

The quantities $d_{s_1+1}^1, d_1^2, d_1^3$ can be seen as

- $d_{s_1+1}^1$: the length of the last piecewise constant region of the first path graph obtained by decomposing the branched path graph into three path graphs by cutting two edges which are not necessarily $(b, b+1)$ and $(b, n_1+1)$, when $S$ is the candidate active set.
- $d_1^2, d_1^3$: the length of the first piecewise constant region of the second, resp. third, path graph obtained by the decomposition procedure explained above for $d_{s_1+1}^1$.

### *3.3. Branching point with arbitrarily many branches*

In Sections 4 and 7 we are going to consider branching points participating in $K+1$ edges. In these cases we are ging go denote by $d_{s_1+1}^1$ the number of vertices

between the branching point and the last vertex in $S$ in the main branch, with these two extreme vertices included, and by $d_1^2, \ldots, d_1^{K+1}$ the number of vertices after the branching point and before the first vertex in $S$ (or the end of the relative branch). In these more complex cases for the sake of simplicity we only consider situations where the first and second notation coincide.

## 4. Calculation of projection coefficients and lengths of antiprojections, a local approach

In this section we are going to present an easy and intuitive way of calculating (anti-)projections and the related projection coefficients of some columns of a path matrix rooted at vertex 1 of a tree onto a subset of the columns of the same matrix. Let this matrix be called $X$. These calculations are motivated by the necessity of finding explicit expressions for the length of the antiprojections (for the weighted compatibility constant) and for the projection coefficients (to check for which signal patterns the irrepresentable condition is satisfied).

In particular consider the task of projecting a column $X_j, j \notin \{1\} \cup S$ onto $X_{\{1\}\cup S}$. This can be seen as finding the following argmin:

$$\hat{\theta}^j := \arg \min_{\theta^j \in \mathbb{R}^{s+1}} \|X_j - X_{\{1\}\cup S}\theta^j\|_2^2.$$

We see that:

- $\hat{\theta}^{j'}$ corresponds to the $j^{\text{th}}$ row of $X'X_{\{1\}\cup S}(X'_{\{1\}\cup S}X_{\{1\}\cup S})^{-1}$;
- $\|X_j - X_{\{1\}\cup S}\hat{\theta}^j\|_2^2 = n\omega_j^2$.

The direct computation of these quantities can be quite laborious. Here, we show an easier way to compute these projections and we prove that they can be computed "locally", i.e. taking into account only some smaller part of the graph.

We start by considering the path graph. Then we treat the more general situation of a branching point with arbitrarily many branches.

### *4.1. Path graph*

Let $j \notin \{1\} \cup S$ be the index of a column of $X$ that we want to project onto $X_{\{1\}\cup S}$. Define

$$j^- := \max\{i < j, i \in \{1\} \cup S\}, \tag{1}$$

$$j^+ := \min\{i > j, i \in \{1\} \cup S \cup \{n+1\}\}, \tag{2}$$

and denote their indices inside $\{1\} \cup S \cup \{n+1\} = \{i_1, \ldots, i_{s+2}\}$ by $l^-$ and $l^+$, i.e. $j^- = i_{l^-}$ and $j^+ = i_{l^+}$. We use the convention $X_{n+1} = 0 \in \mathbb{R}^n$. We are going to show that the projection of $X_j$ onto $X_{\{1\}\cup S}$ is the same as its projection onto $X_{\{j^-\}\cup\{j^+\}}$. This means that the part of the set $\{1\} \cup S$ not bordering with $j$ can be neglected.

The intuition behind this insight can be clarified as follows. Projecting $X_j$ onto $X_{\{1\}\cup S}$ amounts to finding the projection coefficients $\hat{\theta}^j$ minimizing the length of the antiprojection. The projection is then $X_{\{1\}\cup S}\hat{\theta}^j$. Since the columns of $X_{\{1\}\cup S}$ can be seen as indicator functions on $[n]$, this projection problem can be interpreted as the problem of finding the least squares approximation to $1_{\{i\geq j\}}$ by using functions in the class $\{1_{\{i\geq j^*\}}, j^* \in \{1\}\cup S\}$.

We now apply a linear transformation in order to obtain orthogonal design. Note that $\mathrm{I}_{s+1} = \tilde{D}^{(s+1)}X^{(s+1)}$, where $\tilde{D}^{(s+1)}$ is the incidence matrix of a path graph with $s+1$ vertices rooted at vertex 1 and $X^{(s+1)}$ is its inverse, i.e. the corresponding rooted path matrix. We get that

$$\min_{\theta^j\in\mathbb{R}^{s+1}} \|X_j - X_{\{1\}\cup S}\theta^j\|_2^2 = \min_{\tau^j\in\mathbb{R}^{s+1}} \|X_j - X_{\{1\}\cup S}\tilde{D}^{(s+1)}\tau^j\|_2^2,$$

where $\tau^j = X^{(s+1)}\theta^j$, i.e. the progressively cumulative sum of the components of $\theta^j$ and $X_{\{1\}\cup S}\tilde{D}^{(s+1)} \in \mathbb{R}^{n\times(s+1)}$ is a matrix containing as columns the indicator functions $\{1_{\{i_l\leq i<i_{l+1}\}}, l\in\{1,\dots,s+1\}\}$, which are pairwise orthogonal. Because of the orthogonality of the design matrix, we can now solve $s+1$ separate optimization problems to find the components of $\hat{\tau}^j$. It is clear that, to minimize the sum of squared residuals (i.e. the length of the antiprojection), $\hat{\tau}^j$ must be s.t.

$$\{\hat{\tau}_i^j\}_{i<l^-} = 0 \text{ and } \{\hat{\tau}_i^j\}_{i\geq l^+} = 1.$$

It now remains to find $\hat{\tau}^j_{l^-}$ by solving

$$\hat{\tau}^j_{l^-} = \arg\min_{x\in\mathbb{R}} \{(j-j^-)x^2 + (j^+-j)(1-x)^2\} = \frac{j^+-j}{j^+-j^-} = 1 - \frac{j-j^-}{j^+-j^-}.$$

We see that, to get this projection coefficient, we either need to know $j^+$ and $j^-$ or the information on the length of the constant segment in which $j$ lies with its position within this segment. Thus, we obtain that

$$\hat{\tau}^j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{j^+-j}{j^+-j^-} \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad \hat{\theta}^j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{j^+-j}{j^+-j^-} \\ \frac{j-j^-}{j^+-j^-} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and have proved the following Lemma.

**Lemma 4.1** (Localizing the projections)**.** *Let $X$ be the path matrix rooted at vertex 1 of a path graph with $n$ vertices and $S\subseteq\{2,\dots,n\}$. For $j\notin\{1\}\cup S$ define $j^-$ and $j^+$ as in Equations (1) and (2). Then*

$$\min_{\theta^j \in \mathbb{R}^{s+1}} \|X_j - X_{\{1\}\cup S}\theta^j\|_2^2 = \min_{\tilde{\theta}^j \in \mathbb{R}^2} \|X_j - X_{\{j^-\}\cup\{j^+\}}\tilde{\theta}^j\|_2^2,$$

*i.e. the (length of the) (anti-)projections can be computed in a "local" way.*

*Moreover, by writing* $A_{\{1\}\cup S} = I_n - \Pi_{\{1\}\cup S}$ *we have that*

$$\|A_{\{1\}\cup S}X_j\|_2^2 = \frac{(j^+ - j)(j - j^-)}{(j^+ - j^-)}.$$

*Furthermore, for* $j < i_s, j \notin \{1\} \cup S$*, the sum of the entries of* $\hat{\theta}^j$ *is 1.*

### *4.2. General branching point*

Using arguments similar to the ones above we can now focus on a branching point of a general tree graph.

#### *4.2.1. General branching point and* $S = \emptyset$

Let us consider $K$ path graphs of length $\tilde{d}_1^2, \tilde{d}_1^3, \ldots, \tilde{d}_1^{K+1}$ attached at the end of a path graph (which we assume to contain the root) of length $\tilde{d}_{s_1+1}^1$. We define $d^* = \tilde{d}_{s_1+1}^1 + \sum_{l=2}^{K+1} \tilde{d}_1^l$. The path matrix rooted at the first vertex is

$$X = \begin{pmatrix} X^{(\tilde{d}_{s_1+1}^1)} & & & \\ 1 & X^{(\tilde{d}_1^2)} & & \\ \vdots & & \ddots & \\ 1 & & & X^{(\tilde{d}_1^{K+1})} \end{pmatrix} \in \mathbb{R}^{d^* \times d^*}$$

and we want to find the projections of $X_{-1}$ onto $X_1 = (1, \ldots, 1)'$. The entries $X^{(q)}, q \in \{\tilde{d}_{s_1+1}^1, \tilde{d}_1^2, \ldots, \tilde{d}_1^{K+1}\}$ of the matrix $X$ are $q \times q$ lower triangular matrices of ones. Let us write $j = 1 + i, i \in \{1, \ldots, \tilde{d}_{S_1+1}^1 - 1\}$ in the case $j \leq \tilde{d}_{s_1+1}^1$ and $j = \tilde{d}_{s_1+1}^1 + \sum_{l=d}^{i^*} \tilde{d}_1^l - i, i \in \{1, \ldots, \tilde{d}_1^{i^*}\}$ for some $i^* \in \{2, \ldots, K+1\}$ in the case $j \geq \tilde{d}_{s_1+1}^1$. Without loss of generality we can consider only one $i^* \in \{2, \ldots, K+1\}$. Note that the index $i^*$ tells us in which sub-path graph the column we want to project onto $X_1$ lies. We now consider two cases.

- First case: $j \in \{2, \ldots, \tilde{d}_{s_1+1}^1\}$ .

  We write $j = i + 1, i \in \{1, \ldots, \tilde{d}_{s_1+1}^1 - 1\}$. We want to compute the projection coefficient onto $X_1 \in \mathbb{R}^{d^*}$, i.e. we want to find

  $$\hat{\theta}^j = \arg\min_{\theta^j \in \mathbb{R}} \|X_j - X_1\theta^j\|_2^2.$$

  From the first order optimality condition we get that

  $$\hat{\theta}^j = \frac{X_1'X_j}{X_1'X_1} = \frac{d^* - j + 1}{d^*} = \frac{d^* - i}{d^*} = 1 - \frac{i}{d^*}.$$

It follows that

$$\|A_{\{1\}\cup S}X_j\|_2^2 = \frac{i(d^*-i)}{d^*}, 1 \le i \le \tilde{d}^1_{s_1+1} - 1.$$

- Second case: $j \in \{\tilde{d}^1_{s_1+1} + \sum_{l=2}^{i^*-1} \tilde{d}^l_1 + 1, \tilde{d}^1_{s_1+1} + \sum_{l=2}^{i^*} \tilde{d}^l_1\}$, for some $i^* \in \{2, \dots, K+1\}$.

  We write $j = \tilde{d}^1_{s_1+1} + \sum_{l=2}^{i^*-1} \tilde{d}^l_1 + i, i \in \{1, \dots, d^{i^*}_1\}$ and we see that

$$\hat{\theta}^j = \arg\min_{\theta^j} \|X_j - X_1\theta^j\|_2^2 = \frac{X_1'X_j}{X_1'X_1} = \frac{\tilde{d}^1_{s_1+1} + \sum_{l=2}^{i^*} \tilde{d}^l_1 - j + 1}{d^*} = \frac{i}{d^*}.$$

  This is the case because the value of $X_1'X_j$ is the number of nonzero elements in $X_j$. This number can be calculated by seeing that the index $i$ describes the position of $X_j$ inside $X^{(i^*)}$ starting from the left, which is exactly the number of nonzero elements in $X_j$. Alternatively we can see that $X_1'X_j$ can be interpreted as the difference between $\tilde{d}^1_{s_1+1} + \sum_{l=2}^{i^*} \tilde{d}^l_1$ and $j$.
  From the above calculation it follows that the length of the antiprojections is

$$\|A_{\{1\}\cup S}X_j\|_2^2 = \frac{i(d^*-i)}{d^*}, 1 \le i \le \tilde{d}^{i^*}_1.$$

Note that in the last region before the end of one branch, the approximation of the indicator function we implicitely calculate does not have to jump up to one and thus only one coefficient of the respective $\hat{\theta}^j$ will be nonzero and this coefficient will be smaller than one.

### *4.2.2. General branching point and S has elements in all the branches*

Now we focus on the case where each of the branches (path graphs) involved in a branching presents at least one jump (i.e. one element of the set $S$). The length of the antiprojections is calculated in the same way as above. According to the arguments exposed in precedence, we can consider only the jumps surrounding the branching point. Indeed we observed in Subsection 4.1 that what happens in a path graph between two elements of $S$ does not influence and is not influenced by what happens outside that region.

Let us call the jumps surrounding the branching point $j_1, j_2, \dots, j_{k+1}$. We have to find

$$\begin{aligned}
\hat{\theta}^j &= \arg\min_{\theta^j \in \mathbb{R}^{s+1}} \|X_j - X_{\{1\}\cup S}\theta^j\|_2^2 \\
&= \arg\min_{\tilde{\theta}^j \in \mathbb{R}^{K+1}} \|X_j - X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}}\tilde{\theta}^j\|_2^2 \\
&= \arg\min_{\tilde{\theta}^j \in \mathbb{R}^{K+1}} \|X_j - X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}}D^\star X^\star\tilde{\theta}^j\|_2^2,
\end{aligned}$$

where

$$D^\star = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ \vdots & & \ddots & \\ -1 & & & 1 \end{pmatrix} \in \mathbb{R}^{(K+1)\times(K+1)} \text{ and } X^\star = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & & & 1 \end{pmatrix} \in \mathbb{R}^{(K+1)}$$

are respectively the rooted incidence matrix of a star graph with $(K+1)$ vertices and its inverse.

Let us write $j = j_1+i, i \in \{1,\dots,\tilde{d}^1_{s_1+1}-1\}$ and $j = j_l-i, i \in \{1,\dots,\tilde{d}^l_1\}, l \in \{2,\dots,K+1\}$. We define $d^* = \tilde{d}^1_{s_1+1} + \sum_{l=2}^{K+1} \tilde{d}^l_1$. Now let

$$\hat{\tau}^j = \arg\min_{\tau^j \in \mathbb{R}^{K+1}} \|X_j - X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} D^\star \tau^j\|_2^2.$$

Note that in this case our task consists in calculating $K+1$ projection coefficients, whereas we had to calculate only one of them in the preceeding subsection.

The first order optimality conditions translate into

$$D^{\star\prime} X'_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} D^\star \hat{\tau}^j = D^{\star\prime} X'_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} X_j.$$

Note that $X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} D^\star$ differs from $X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}}$ only in the first column which is $X_{j-1} - \sum_{l=2}^{K+1} X_{j_l}$. Thus the columns of $X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} D^\star$ are orthogonal to each other and $D^{\star\prime} X'_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} D^\star$ is a diagonal matrix with first entry $d^*$. The other diagonal entries are respectively the numbers of nonzero elements of $X_{j_2},\dots,X_{j_{K+1}}$.

We can now distinguish two cases:

- First case: $j \in \{j_1+1,\dots,j_1+\tilde{d}^1_{s_1+1}-1\}$.

  We write $j = j_1+i, i \in \{1,\dots,\tilde{d}^1_{s_1+1}-1\}$. Then it follows that the product $(D^{\star\prime} X'_{\{j_1\}\cup\dots\cup\{j_{K+1}\}})_{1,\cdot} X_j$ has value $d^*+j_1-j = d^*-i$ and the product with the other rows of $D^{\star\prime} X'_{\{j_1\}\cup\dots\cup\{j_{K+1}\}}$ is equal to the $2^{\text{nd}}$ to the $(K+1)^{\text{th}}$ diagonal values of $D^{\star\prime} X'_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} X_{\{j_1\}\cup\dots\cup\{j_{K+1}\}} D^\star$. Thus, we get that:

  $$\hat{\tau}^j_1 = 1 - \frac{i}{d^*},$$
  $$\hat{\tau}^j_l = 1, l = \{2,\dots,K+1\},$$

  which translates into

  $$\hat{\theta}^j_1 = 1 - \frac{i}{d^*},$$
  $$\hat{\theta}^j_l = \frac{i}{d^*}, l = \{2,\dots,K+1\}.$$

We thus get that

$$\begin{aligned}\|A_{\{1\}\cup S}X_j\|_2^2 &= \left(1-\frac{i}{d^*}\right)^2 i + \left(1-\left(1-\frac{i}{d^*}\right)\right)^2 (d^*-i) \\ &= \frac{(d^*-i)i}{d^*}, i \in \{1,\ldots,\tilde{d}^1_{s_1+1}-1\}.\end{aligned}$$

- Second case: $j \in \{j_{l'}-\tilde{d}^{l'}_1,\ldots,j_{l'}-1\}$, for some $l' \in \{2,\ldots,K+1\}$.

  We write $j = j_{l'}-i, i \in \{1,\ldots,\tilde{d}^{l'}_1\}$. Then it follows that the product $(D^{\star\prime}X'_{\{j_1\}\cup\ldots\cup\{j_{K+1}\}})_{1,\cdot}X_j$ has value $i$, i.e. the number of nonzero elements that $X_j$ has in addition to $X_{j_{l'}}$. Moreover $(D^{\star\prime}X'_{\{j_1\}\cup\ldots\cup\{j_{K+1}\}})_{i,\cdot}X_j = 0$, $i \in \{2,\ldots,K+1\}\setminus\{l'\}$ and the product $(D^{\star\prime}X'_{\{j_1\}\cup\ldots\cup\{j_{K+1}\}})_{l',\cdot}X_j$ is equal to the $l'^{\text{th}}$ diagonal entry of $D^{\star\prime}X'_{\{j_1\}\cup\ldots\cup\{j_{K+1}\}}X_{\{j_1\}\cup\ldots\cup\{j_{K+1}\}}D^{\star}$. By the diagonality of the above matrix it follows that:

$$\begin{aligned}&\hat{\tau}^j_1 = \frac{i}{d^*}, \\ &\hat{\tau}^j_l = 0, l \in \{2,\ldots,K+1\}\setminus\{l'\}, \\ &\hat{\tau}^j_l = 1, l = l',\end{aligned}$$

  which translates into

$$\begin{aligned}&\hat{\theta}^j_1 = \frac{i}{d^*}, \\ &\hat{\theta}^j_l = -\frac{i}{d^*}, l \in \{2,\ldots,K+1\}\setminus\{l'\}, \\ &\hat{\theta}^j_l = 1-\frac{i}{d^*}, l = l'.\end{aligned}$$

We thus get that

$$\begin{aligned}\|A_{\{1\}\cup S}X_j\|_2^2 &= \left(\frac{i}{d^*}\right)^2 (d^*-i) + \left(1-\frac{i}{d^*}\right)^2 i \\ &= \frac{(d^*-i)i}{d^*}, i \in \{1,\ldots,\tilde{d}^{l'}_1\}.\end{aligned}$$

## 5. Path graph

### 5.1. Compatibility constant

In this section we assume $\mathcal{G}$ to be the path graph with $n$ vertices. We give two lower bounds for the compatibility constant for the path graph without and with weights. The proofs are postponed to the Appendix B, where we present

some elements that allow extension to the branched path graph and to more general tree graphs as well. These bounds are presented in a paper by van de Geer (2018) as well. We use the second notation exposed in Section 3.

Let $S \subseteq [2, \dots, n]$ be a subset of the edges of a path graph with $n$ vertices.

**Assumption 5.1.** Assume that $S$ can be written as in the second notation in Section 3, where we additionally require that $d_1 \geq 2$, $d_i \geq 4, \forall i \in \{2, \dots, s\}$, $d_{s+1} \geq 2$.

**Assumption 5.2.** Assume that $S$ can be written as in the second notation in Section 3, where we additionally require that $d_i \geq 4, \forall i \in [s+1]$.

**Remark.** Assumption 5.1 is required by Lemma B.3, see Appendix B, where the proofs of this section are, while the slightly stronger Assumption 5.2 allows us to obtain a simpler form for the upper bound on the weighted compatibility constant.

**Lemma 5.3** (Lower bound on the compatibility constant for the path graph, part of Theorem 6.1 in van de Geer (2018))**.** *Under Assumption 5.1 on $S$, let $\{u_j\}_{j=2}^s$ be a sequence of integers, s.t. $2 \leq u_j \leq d_j - 2, \forall j \in \{2, \dots, s\}$.*

*Then for the path graph it holds that*

$$\kappa^2(S) \geq \frac{s+1}{n} \frac{1}{K},$$

*where*

$$K = \frac{1}{d_1} + \sum_{j=2}^{s} \left( \frac{1}{u_j} + \frac{1}{d_j - u_j} \right) + \frac{1}{d_{s+1}}.$$

*Proof of Lemma 5.3.* See Appendix B. □

**Corollary 5.4** (The bound can be tight, part of Theorem 6.1 in van de Geer (2018))**.** *Suppose that Assumption 5.1 on $S$ holds and that moreover $d_j$ is even $\forall j \in \{2, \dots, s\}$, so that we can take $u_j = d_j/2$. Let us now define $f^* \in \mathbb{R}^n$ by*

$$f_i^* = \begin{cases} -\frac{n}{d_1} & i = 1, \dots, d_1 \\ \frac{2n}{d_2} & i = d_1 + 1, \dots, d_1 + d_2 \\ \vdots & \\ (-1)^s \frac{2n}{d_s} & i = \sum_{j=1}^{s-1} d_j + 1, \dots, \sum_{j=1}^{s} d_j \\ (-1)^{s+1} \frac{n}{d_{s+1}} & i = \sum_{j=1}^{s} d_j + 1, \dots, n \end{cases}.$$

*Let $\beta^*$ be defined by $f^* = X\beta^*$. Then*

$$\kappa^2(S) = \frac{s+1}{n} \frac{1}{K},$$

*where*

$$K = \frac{1}{d_1} + \sum_{j=2}^{s} \frac{4}{d_j} + \frac{1}{d_{s+1}}.$$

*Proof of Corollary 5.4.* See Appendix B. □

**Remark.** For the compatibility constant we want to find the largest possible lower bound. Thus we have to choose the $u_j$'s s.t. $K$ is minimized. We look at the first order optimality conditions and notice that they reduce to finding the extremes of $(s-1)$ functions of the type $g(x) = \frac{1}{d-x} + \frac{1}{x}$, $x \in (0,d)$, where $d \in \mathbb{N}$ is fixed. The global minimum of $g$ on $(0,d)$ is achieved at $x = \frac{d}{2}$. Because of the restriction $x \in \mathbb{N}$, as soon as at least one $d_j, 2 \leq j \leq s$ is odd, we can not obtain a value of $K$ giving a tight bound for our definition $f^*$.

**Lemma 5.5** (Lower bound on the weighted compatibility constant for the path graph, Lemma 9.1 in van de Geer (2018))**.** *Under Assumption 5.1 on $S$, let $\{u_j\}_{j=2}^s$ be a sequence of integers, s.t. $2 \leq u_j \leq d_j - 2, \forall j \in \{2,\ldots,s\}$.*

*Then for the path graph it holds that*

$$\kappa_w^2(S) \geq \frac{s+1}{n} \frac{1}{(\|w\|_\infty \sqrt{K} + \|Dw\|_2)^2} \geq \frac{s+1}{n} \frac{1}{2(\|w\|_\infty^2 K + \|Dw\|_2^2)},$$

*where $D$ is the incidence matrix of the path graph.*

*Proof of Lemma 5.5.* See Appendix B. □

### 5.2. Oracle inequality

Define the vector

$$\Delta := (d_1, \lfloor d_2/2 \rfloor, \lceil d_2/2 \rceil, \ldots, \lfloor d_s/2 \rfloor, \lceil d_s/2 \rceil, d_{s+1}) \in \mathbb{R}^{s+1}$$

and let $\overline{\Delta}_h$ be its harmonic mean.

We now want to translate the result of Theorem 2.5 to the path graph. To do so we need a lower bound for the weighted compatibility constant, i.e. an explicit upper bound for $\sum_{i=2}^n (w_i - w_{i-1})^2$. In this way we can obtain the following corollary.

**Corollary 5.6** (Sharp oracle inequality for the path graph)**.** *Suppose that $S$ is s.t. Assumption 5.2 holds. Then we have that, with probability at least $1-\delta$,*

$$\begin{aligned} \|\widehat{f} - f^0\|_n^2 &\leq \inf_{f \in \mathbb{R}^n} \left\{ \|f - f^0\|_n^2 + 4\lambda \|(Df)_{-S}\|_1 \right\} \\ &+ \frac{8\log(2/\delta)\sigma^2}{n} + 4\sigma^2 \frac{s+1}{n} \\ &+ 8\sigma^2 \log(4(n-s-1)/\delta) \left( \frac{2\gamma^2 s}{\bar{\Delta}_h} + 5\frac{s+1}{n} \log\left( \frac{n}{s+1} \right) \right). \end{aligned}$$

*Suppose Assumption 5.2 holds for $S_0$. If we choose $f = f^0$ and $S = S_0$ we obtain that, with probability at least $1-\delta$,*

$$\|\widehat{f} - f^0\|_n^2 = \mathcal{O}(\log(n/\delta) s_0 / \bar{\Delta}_h) + \mathcal{O}(\log(n/\delta) \log(n/(s_0+1))(s_0+1)/n).$$

*Proof of Corollary 5.6.* See Appendix B. □

**Remark.** Since the harmonic mean of $\Delta$ is upper bounded by its arithmetic mean, and this upper bound is attained when all the entries of $\Delta$ are the same, we get a lower bound for the order of the mean squared error of

$$\frac{(s+1)\log(n)}{n}\left((s+1)+\log\left(\frac{n}{s+1}\right)\right).$$

**Remark.** Our result differs from the one obtained by Dalalyan, Hebiri and Lederer (2017) in two points:

- We have $\bar{\Delta}_h$, the harmonic mean of the distances between jumps, instead of $\min_j \Delta_j$, the minimum distance between jumps;
- We slightly improve the rate from by reducing a $\log(n)$ to $\log(n/(s+1))$. This is achieved with a more careful bound on the square of the consecutive differences of the weights.

## 6. Path graph with one branch

In this section we consider $\mathcal{G}$ to be the path graph with one branch and $n$ vertices.

We present two assumptions, which are analogous to Assumptions 5.1 and 5.2 presented in Section 5

Let $S$ be a subset of the edges of the branched path graph.

**Assumption 6.1.** Suppose $S$ can be written as in the second notation in Section 3, with

- $d_1^i \geq 2, \forall i \in [3]$;
- $d_{s_i+1}^i \geq 2, \forall i \in [3]$;
- $d_j^i \geq 4, \forall j \in \{2,\dots,s_i\}, \forall i \in [3]$.

**Assumption 6.2.** We require that Assumption 6.1 holds and, in addition, that $d_1^1 \geq 4$, $d_{s_i+1}^i \geq 4, \forall i \in \{2,3\}$.

### *6.1. Compatibility constant*

**Lemma 6.3** (Lower bound for the compatibility constant for the branched path graph)**.** *Under Assumption 6.1 on $S$, let $u_j^i \in \mathbb{N}$ satisfy $2 \leq u_j^i \leq d_j^i - 2$ for $j \in \{2,\dots,s_i\}$ and $i \in \{1,2,3\}$.*

*Then, for the branched path graph it holds that*

$$\kappa^2(S) \geq \frac{s+1}{n}\frac{1}{K^b},$$

*where*

$$K^b = \sum_{i=1}^{3}\left(\frac{1}{d_1^i} + \sum_{j=2}^{s_i}\left(\frac{1}{u_j^i} + \frac{1}{d_j^i - u_j^i}\right) + \frac{1}{d_{s_i+1}^i}\right)$$

*Proof of Lemma 6.3.* See Appendix C. □

**Corollary 6.4** (The bound can be tight)**.** *Suppose that Assumption 6.1 holds and that moreover $d_j^i$ is even $\forall j \in \{2,\dots,s_i\}, i \in \{1,2,3\}$. One can then choose $u_j^i = d_j/2, \forall j \in \{2,\dots,s_i\}, i \in \{1,2,3\}$. Moreover, assume that $d_{s_1+1}^1 = d_1^2 = d_1^3$. Let $f^i, i \in \{1,2,3\}$ be the restriction of $f$ to the three path graphs of length $q_i$ each implicitely obtained when using the second notation. Let us now define $f^{*i} \in \mathbb{R}^{q_i}$ by*

$$f_j^{*i} = \begin{cases} -\frac{n}{d_1^1} & j = 1,\dots,d_1^1 \\ \frac{2n}{d_2^1} & j = d_1^1+1,\dots,d_1^1+d_2^1 \\ \vdots & \\ (-1)^{s_1}\frac{2n}{d_{s_1}^1} & j = \sum_{j=1}^{s_1-1} d_j^1 + 1,\dots,\sum_{j=1}^{s_1} d_j^1 \\ (-1)^{s_1+1}\frac{n}{d_{s_1+1}^1} & j = \sum_{j=1}^{s_1} d_j^1 + 1,\dots,q_1 \end{cases}$$

*and for $i \in \{2,3\}$*

$$f_j^{*i} = \begin{cases} (-1)^{s_1+1}\frac{n}{d_1^i} & j = 1,\dots,d_1^i \\ (-1)^{s_1+2}\frac{2n}{d_2^i} & j = d_1^i+1,\dots,d_1^i+d_2^i \\ \vdots & \\ (-1)^{s_1+s_i+1}\frac{2n}{d_{s_i}^i} & j = \sum_{j=1}^{s_i-1} d_j^i + 1,\dots,\sum_{j=1}^{s_i} d_j^i \\ (-1)^{s_1+s_1+1}\frac{n}{d_{s_i+1}^i} & j = \sum_{j=1}^{s_i} d_j^i + 1,\dots,q_i. \end{cases}$$

*Let $\beta^*$ be defined by $f^* = X\beta^*$. Then*

$$\kappa^2(S) = \frac{s+1}{n}\frac{1}{K^b},$$

*where*

$$K^b = \sum_{i=1}^{3}\left(\frac{1}{d_1^i} + \sum_{j=2}^{s_i}\frac{4}{d_j^i} + \frac{1}{d_{s_i+1}^i}\right).$$

*Proof of Corollary 6.4.* See Appendix C. □

Consider the decomposition of the branched path graph into three path graphs, implicitely done by using the second notation in Section 3. Let $D^*$ denote the incidence matrix of the branched path graph, where the entries in the rows corresponding to the edges connecting the three above mentioned path graphs have been substituted with zeroes.

**Lemma 6.5** (Lower bound on the weighted compatibility constant for the branched path graph)**.** *Under Assumption 6.1 on $S$, let $u_j^i \in \mathbb{N}$ satisfy $2 \le u_j^i \le d_j^i - 2$ for $j \in \{2,\dots,s_i\}$ and $i \in \{1,2,3\}$.*

*Then, for the branched path graph it holds that*

$$\begin{aligned}\kappa_w^2(S) &\geq \frac{s+1}{n}\frac{1}{(\sqrt{K^b}\|w\|_\infty + \|D^* w\|_2)^2} \geq \frac{s+1}{n}\frac{1}{2(K^b\|w\|_\infty^2 + \|D^* w\|_2^2)} \\ &\geq \frac{s+1}{n}\frac{1}{2(K^b\|w\|_\infty^2 + \|Dw\|_2^2)}.\end{aligned}$$

*Proof of Lemma 6.5.* See Appendix C. □

### 6.2. Oracle inequality

As in the case of the path graph, to prove an oracle inequality for the branched path graph, we need to find an explicit expression to control the weighted compatibility constant to insert in Theorem 2.5. The resulting bound is similar to the one obtained in the Proof of Corollary 5.6, up to a difference: we now have to handle with care the region around the branching point $b$.

For the branched path graph we define the vectors

$$\Delta^i := (d_1^i, \lfloor d_2^i/2 \rfloor, \lceil d_2^i/2 \rceil, \ldots, \lfloor d_{s_i}^i/2 \rfloor, \lceil d_{s_i}^i/2 \rceil, d_{s_i+1}^i) \in \mathbb{R}^{2s_i}, i \in [3],$$

and $\Delta := (\Delta^1, \Delta^2, \Delta^3) \in \mathbb{R}^{2s}$. Let $\bar{\Delta}_h$ be the harmonic mean of $\Delta$.

#### 6.2.1. Jumps far away from the branching point

We first prove a result where all the jumps surrounding the branching point are far enough from it.

**Corollary 6.6** (Sharp oracle inequality for the branched path graph)**.** *Suppose that $S$ is s.t. Assumption 6.2 holds. Moreover assume that $\tilde{d}_{s_1+1}^1, \tilde{d}_1^2, \tilde{d}_1^3 \geq 2$. Choose $d_{s_1+1}^1 = \tilde{d}_{s_1+1}^1, d_1^2 = \tilde{d}_1^2, d_1^3 = \tilde{d}_1^3$. Then we have that, with probability at least $1-\delta$,*

$$\begin{aligned}\|\widehat{f} - f^0\|_n^2 &\leq \inf_{f \in \mathbb{R}^n} \left\{ \|f - f^0\|_n^2 + 4\lambda \|(Df)_{-S}\|_1 \right\} \\ &+ \frac{8\log(2/\delta)\sigma^2}{n} + 4\sigma^2 \frac{s+1}{n} \\ &+ 8\sigma^2 \log(4(n-s-1)/\delta) \left( \frac{2\gamma^2 s}{\bar{\Delta}_h} + \frac{5(2s+3)}{2n} \log\left( \frac{n+1}{2s+3} \right) \right),\end{aligned}$$

*Suppose that Assumption 6.2 holds for $S_0$. If we choose $f = f^0$ and $S = S_0$ we get that, with probability at least $1-\delta$,*

$$\|\hat{f} - f^0\|_n^2 = \mathcal{O}(\log(n/\delta) s_0/\bar{\Delta}_h) + \mathcal{O}(\log(n/\delta)\log(n/(2s_0+3))(2s_0+3)/n).$$

*Proof of Corollary 6.6.* See Appendix C. □

This case with slightly stronger assumptions will be used in Section 7 to further extend the result to more complex tree structures.

*6.2.2. Some jump close to the branching point*

Our approach allows us to handle cases where some jump occurs close to the branching point too. As made clear in the second notation in Section 3, we require that all $d^1_{s_1+1}, d^2_1, d^3_1 \geq 2$, i.e. $d^* = \tilde{d}^1_{s_1+1} + \tilde{d}^2_1 + \tilde{d}^3_1 \geq 6$. This means that our approach can handle the case where at most one of the jumps surrounding the bifurcation point occurs directly at the bifurcation point. Note that neither $\tilde{d}^1_{s_1+1} = 0$ nor $\tilde{d}^2_1 + \tilde{d}^3_1 = 0$ are allowed.

We can distinguish the following three cases:

1) $\tilde{d}^2_1 = 0$ or $\tilde{d}^3_1 = 0$;
2) $\tilde{d}^1_{s_1+1} = 1$;
   a) $\tilde{d}^2_1 \wedge \tilde{d}^3_1 = 2$;
   b) $\tilde{d}^2_1 \wedge \tilde{d}^3_1 \geq 3$;
3) $\tilde{d}^2_1 = 1$ or $\tilde{d}^3_1 = 1$;

**Corollary 6.7** (Sharp oracle inequality for the branched path graph)**.** *Suppose that $S$ is s.t. Assumption 6.2 holds.*

*Then we have that, with probability at least $1-\delta$,*

$$\begin{aligned}\|\widehat{f} - f^0\|_n^2 &\leq \inf_{f\in\mathbb{R}^n}\left\{\|f-f^0\|_n^2 + 4\lambda\|(Df)_{-S}\|_1\right\} \\ &+ \frac{8\log(2/\delta)\sigma^2}{n} + 4\sigma^2\frac{s+1}{n} \\ &+ 8\sigma^2\log(4(n-s-1)/\delta)\left(\frac{2\gamma^2 s}{\bar{\Delta}_h} + \frac{5(2s+3)}{2n}\log\left(\frac{n+1}{2s+3}\right) + \frac{\zeta}{n}\right),\end{aligned}$$

*where*

$$\zeta = \begin{cases} d^*/2 & , \text{ Case 1)} \\ 3 & , \text{ Case 2)a)} \\ d^*/4 & , \text{ Case 2)b)} \\ d^*/4 & , \text{ Case 3)} \end{cases}.$$

*Suppose that Assumption 6.2 holds for $S_0$. If we choose $f = f^0$ and $S = S_0$ we get that, with probability at least $1-\delta$,*

$$\begin{aligned}\|\hat{f} - f^0\|_n^2 &= \mathcal{O}(\log(n/\delta)s_0/\bar{\Delta}_h) + \mathcal{O}(\log(n/\delta)\log(n/(2s_0+3))(2s_0+3)/n) \\ &+ \mathcal{O}(\log(n/\delta)\zeta/n).\end{aligned}$$

*Proof of Corollary 6.7.* See Appendix C. □

## 7. Extension to more general tree graphs

In this section we consider only situations corresponding to Corollary 6.6. This means that we assume that, even when at the branching point more than one branch is attached, the edge connecting the additional branch to the branching point and the consecutive one do not present jumps (i.e. are not elements of the set $S$).

### *7.1. Oracle inequality for general tree graphs*

With the insights gained in Section 4 we can, by availing ourselves of simple means, prove an oracle inequality for a general tree graph, where the jumps in $S$ are far enough from the branching points, in analogy to Corollary 6.6.

Here as well, we utilize the general approach exposed in Theorem 2.5 and we need to handle with care the weighted compatibility constant and find a lower bound for it.

We know that, when we are in (the generalization of) the situation of Corollary 6.6, to prove bounds for the compatibility constant, the tree graph can be seen as a collection of path graphs glued together at (some of) their extremities. As seen in Section 4, the length of the antiprojections for the vertices around branching points depends on all the branches attached to the branching point in question. Here, for the sake of simplicity, we only consider situations where the first and the second notation exposed in Section 3 coincide.

**Assumption 7.1.** Assume that $\mathcal{G}$ is a tree graph composed of $g$ path graphs glued together at (some of) their extremities and assume $S$ is s.t. $d^i_j \geq 4, \forall j \in \{1, \dots, s_i + 1\}, \forall i \in \{1, \dots, g\}$, i.e. between consecutive jumps there are at least four vertices as well as there are at least four vertices before the first and after the last jump of each path graph resulting from the decomposition of the tree graph.

Indeed, for $d^i_j \geq 4$, we have that $\log(d^i_j) \leq 2\log(d^i_j/2)$ and this helps us find a nice bound on the weighted compatibility constant.

Let $\mathcal{G}$ be a tree graph and $S$ a candidate active set with the properties exposed in Assumption 7.1 above. In particular it can be decomposed into $g$ path graphs. For each of these path graphs, by using the second notation in Subsection 3, we define the vectors

$$\Delta^i = (d^i_1, \lceil d^i_2/2 \rceil, \lfloor d^i_2/2 \rfloor, \dots, \lceil d^i_{s_i}/2 \rceil, \lfloor d^i_{s_i}/2 \rfloor, d^i_{s_i+1}) \in \mathbb{R}^{2s_i}, i \in \{1, \dots, g\}$$

and

$$|\Delta|^i = (\lceil d^i_1/2 \rceil, \lfloor d^i_1/2 \rfloor, \dots, \lceil d^i_{s_i+1}/2 \rceil, \lfloor d^i_{s_i+1}/2 \rfloor) \in \mathbb{R}^{2s_i+2}, i \in \{1, \dots, g\}.$$

Moreover we write

$$\Delta = (\Delta^1, \dots, \Delta^g) \in \mathbb{R}^{2s} \text{ and } |\Delta| = (|\Delta|^1, \dots, |\Delta|^g) \in \mathbb{R}^{2(s+g)}.$$

We have that for $\mathcal{G}$,

$$\kappa^2(S) \geq \frac{s+1}{n}\frac{1}{K}, K \leq \frac{2s}{\bar{\Delta}_h},$$

where $\bar{\Delta}_h$ is the harmonic mean of $\Delta$. Moreover an upper bound for the inverse of the weighted compatibility constant can be computed by upper bounding the squared consecutive pairwise differences of the weigths for the $g$ path graphs.

Assumption 7.1 allows us to use the standard machinery exposed in Section 5, and in particular in Corollary 5.6, for each of the $g$ path graphs into which

the more complex tree graph can be decomposed. In analogy to Corollary 6.6 we can neglect the edges connecting these path graphs when we look for an explicit lower bound on the weighted compatibility constant.

Let $D^*$ be the incidence matrix of a tree graph that can be decomposed into $g$ path graphs, where the rows corresponding to the $g-1$ edges connecting these $g$ path graphs have been deleted. In analogy to the proof of Corollary 5.6 it follows that

$$\begin{aligned}\|D^*w\|_2^2 &\leq \frac{5}{2\gamma^2 n}\log\left(\prod_{j=1}^{g}\prod_{i=1}^{2(s_j+1)}|\Delta|_i^j\right)\\ &= \frac{5}{\gamma^2 n}(s+g)\log(|\bar{\Delta}|) \leq \frac{5}{\gamma^2 n}(s+g)\log(n/(2s+2g))\\ &\leq \frac{5}{\gamma^2 n}(s+g)\log(n/(s+g)),\end{aligned}$$

where Assumption 7.1 is used.

We thus get that, in analogy to Corollary 5.6,

$$\frac{1}{\kappa_w^2(S)} \leq \frac{2n}{s+1}\left(\frac{2s}{\bar{\Delta}_h}+\frac{5}{\gamma^2}\frac{s+g}{n}\log\left(\frac{n}{s+g}\right)\right).$$

We therefore get the following corollary.

**Corollary 7.2** (Oracle inequality for a general tree graph)**.** *Suppose that the tree graph $\mathcal{G}$ and the candidate active set $S$ satisfy Assumption 7.1. Then, with probability at least $1-\delta$,*

$$\begin{aligned}\|\widehat{f}-f^0\|_n^2 &\leq \inf_{f\in\mathbb{R}^n}\left\{\|f-f^0\|_n^2+4\lambda\|(Df)_{-S}\|_1\right\}\\ &+ \frac{8\log(2/\delta)\sigma^2}{n}+4\sigma^2\frac{s+1}{n}\\ &+ 8\sigma^2\log(4(n-s-1)/\delta)\left(\frac{2\gamma^2 s}{\bar{\Delta}_h}+5\frac{(s+g)}{n}\log\left(\frac{n}{s+g}\right)\right).\end{aligned}$$

**Remark.** Notice that it is advantageous to choose a decomposition where the path graphs are as large as possible, s.t. $g$ is small and fewer requirements on the $d_j^i$'s are posed, because we have less extremities. Indeed, in Assumption 7.1 we require $d_j^i \geq 4, \forall j \in \{1,\dots,s_i+1\}, \forall i \in \{1,\dots,g\}$. This requirement is weaker if we choose a decomposition into longer path graphs.

**Remark.** This approach is of course not optimal, however it allows us to prove in a simple way a theoretical guarantee for the Edge Lasso estimator if some (not extremely restrictive) requirement on $\mathcal{G}$ and $S$ is satisfied.

## 8. Asymptotic signal pattern recovery: the irrepresentable condition

### *8.1. Review of the literature on pattern recovery*

Let $Y = X\beta^0 + \epsilon, \epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{I}_n)$, where $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}, \beta^0 \in \mathbb{R}^p, \epsilon \in \mathbb{R}^n$. Let $S_0 := \left\{j \in [p] : \beta_j^0 \neq 0\right\}$ be the active set of $\beta^0$ and $-S_0$ its complement. We are interested in the asymptotic sign recovery properties of the Lasso estimator

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \left\{\|Y - X\beta\|_n^2 + 2\lambda\|\beta\|_1\right\}.$$

**Definition 8.1** (**Sign recovery**, Definition 1 in Zhao and Yu (2006))**.** *We say that an estimator $\hat{\beta}$ recovers the signs of the true coefficients $\beta^0$ if*

$$sgn(\hat{\beta}) = sgn(\beta^0).$$

*We then write*

$$\hat{\beta} =_s \beta^0.$$

**Definition 8.2** (**Pattern recovery**)**.** *We say that an estimator $\hat{f}$ of a signal $f^0$ on a graph $\mathcal{G}$ with incidence matrix $D$ recovers the signal pattern if*

$$D\hat{f} =_s Df^0.$$

**Definition 8.3** (**Strong sign consistency**, Definition 2 in Zhao and Yu (2006))**.** *We say that the Lasso estimator $\hat{\beta}$ is strongly sign consistent if $\exists\lambda = \lambda(n)$ :*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\beta}(\lambda) =_s \beta^0\right) = 1$$

**Definition 8.4** (**Strong irrepresentable condition**, Zhao and Yu (2006))**.** *Without loss of generality we can write*

$$\beta^0 = \begin{pmatrix} \beta^0_{S_0} \\ \beta^0_{-S_0} \end{pmatrix} = \begin{pmatrix} \beta^0_{S_0} \\ 0 \end{pmatrix} =: \begin{pmatrix} \beta^0_1 \\ \beta^0_2 \end{pmatrix},$$

*where 1 and 2 are shorthand notations for $S_0$ and $-S_0$, and*

$$\hat{\Sigma} := \frac{X'X}{n} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}.$$

*Assume that $\hat{\Sigma}_{11}$ is invertible. The strong irrepresentable condition is satisfied if $\exists\eta \in (0, 1]$ :*

$$\|\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1} sgn(\beta_1^0)\|_\infty \leq 1 - \eta$$

Zhao and Yu (2006) prove (in their Theorem 4) that under Gaussian noise the strong irrepresentable condition implies strong sign consistency of the Lasso estimator, if $\exists 0 \leq c_1 < c_2 \leq 1$ and $C_1 > 0 : s_0 = \mathcal{O}(n^{c_1})$ and $n^{\frac{1-c_2}{2}} \min_{j\in S_0}|\beta_j^0| \geq$

$C_1$. For our setup this means that $s_0$ has to grow more slowly than $\mathcal{O}(n)$ and that the magnitude of the smallest nonzero coefficient has to decay (much) slower than $\mathcal{O}(n^{-1/2})$.

In the literature, considerable attention has been given to the question whether or not it is possible to consistently recover the pattern of a piecewise constant signal contaminated with some noise, say Gaussian noise. In that regard, Qian and Jia (2016) highlight the so called **staircase problem**: as soon as there are two consecutive jumps in the same direction in the underlying signal separated by a constant segment, no consistent pattern recovery is possible, since the irrepresentable condition (cfr. Zhao and Yu (2006)) is violated.

Some cures have been proposed to mitigate the staircase problem. Rojas and Wahlberg (2015); Ottersten, Wahlberg and Rojas (2016) suggest to modify the algorithm for computing the Fused Lasso estimator. Their strategy is based on the connection made by Rojas and Wahlberg (2014) between the Fused Lasso estimator and a sequence of discrete Brownian Bridges. Owrang et al. (2017) propose instead to normalize the design matrix of the associated Lasso problem, to comply with the irrepresentable condition. Another proposal aimed at complying with the irrepresentable condition is the one by Qian and Jia (2016), based on the preconditioning of the design matrix with the puffer transformation defined in Jia and Rohe (2015), which results in estimating the jumps of the true signal with the soft-thresholded differences of consecutive observations.

### 8.2. *Approach to pattern recovery for total variation regularized estimators over tree graphs*

Let us now consider the case of the Edge Lasso on a tree graph rooted at vertex 1. We saw in Section 1 that the problem can be transformed into an ordinary Lasso problem where the first coefficient is not penalized.

We start with the following remark.

**Remark** (The irrepresentable condition when some coefficients are not penalized)**.** Let us consider the Lasso problem where some coefficients are not penalized, i.e. the estimator

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \|\beta_{-U}\|_1 \right\},$$

where $U, R, S$ are three subsets partitioning $[p]$. In particular $U$ is the set of the unpenalized coefficients, $R$ is the set of truly zero coefficients and $S$ is the set of truly nonzero (active) coefficients. We assume the linear model $Y = X\beta^0 + \epsilon, \epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{I}_n)$. The vector of true coefficients $\beta^0$ can be written as

$$\beta^0 = \begin{pmatrix} \beta_U^0 \\ \beta_S^0 \\ 0 \end{pmatrix}.$$

Moreover we write

$$\frac{X'X}{n} =: \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{UU} & \hat{\Sigma}_{US} & \hat{\Sigma}_{UR} \\ \hat{\Sigma}_{SU} & \hat{\Sigma}_{SS} & \hat{\Sigma}_{SR} \\ \hat{\Sigma}_{RU} & \hat{\Sigma}_{RS} & \hat{\Sigma}_{RR} \end{pmatrix}.$$

Assume that $|U| \leq n$ and that $\hat{\Sigma}_{UU}, \hat{\Sigma}_{SS}$ and are invertible. We can write the irrepresentable condition as

$$\|X'_R A_U X_S (X'_S A_U X_S)^{-1} z_S^0\|_\infty \leq 1 - \eta,$$

where $z_S^0 = \text{sgn}(\beta_S^0)$, $A_U = \mathrm{I}_n - \Pi_U$ is the antiprojection matrix onto $V_U$, the linear subspace spanned by $X_U$, and $\Pi_U := X_U(X'_U X_U)^{-1} X'_U$ is the orthogonal projection matrix onto $V_U$.

Indeed, write $\delta := \hat{\beta} - \beta^0$. The KKT conditions can be written as

$$\hat{\Sigma}_{UU}\delta_U + \hat{\Sigma}_{US}\delta_S + \hat{\Sigma}_{UR}\delta_R - \frac{X'_U \epsilon}{n} = 0; \tag{3}$$

$$\hat{\Sigma}_{SU}\delta_U + \hat{\Sigma}_{SS}\delta_S + \hat{\Sigma}_{SR}\delta_R - \frac{X'_S \epsilon}{n} + \lambda \hat{z}_S = 0, \hat{z}_S \in \delta \|\hat{\beta}_S\|_1; \tag{4}$$

$$\hat{\Sigma}_{RU}\delta_U + \hat{\Sigma}_{RS}\delta_S + \hat{\Sigma}_{RR}\delta_R - \frac{X'_R \epsilon}{n} + \lambda \hat{z}_R = 0, \hat{z}_R \in \delta \|\hat{\beta}_R\|_1. \tag{5}$$

By solving Equation 3 with respect to $\delta_U$, then inserting into Equation 4 and solving with respect to $\delta_S$, then inserting the expression for $\delta_R$ in the expression for $\delta_U$ to get $\delta_U(\delta_R)$ and $\delta_S(\delta_R)$ and by finally inserting them into Equation 5 by analogy with the proof proposed by Zhao and Yu (2006), we find the irrepresentable condition when some coefficients are not penalized, which writes as follows: $\exists \eta > 0$ :

$$\|\left(\hat{\Sigma}_{RS} - \hat{\Sigma}_{RU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{US}\right)\left(\hat{\Sigma}_{SS} - \hat{\Sigma}_{SU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{US}\right)^{-1} z_S^0\|_\infty \leq 1 - \eta,$$

where $z_S^0 = \text{sgn}(\beta_S^0)$.

Note that $\Pi_U = \frac{1}{n} X_U \hat{\Sigma}_{UU}^{-1} X'_U$ and we obtain the above expression.

Thus, by using the notation of the remark above we let $U = \{1\}$, $S = S_0$ and $R = [n] \setminus (S_0 \cup \{1\})$.

**Lemma 8.5.** *We have that*

$$\|X'_R X_{\{1\}\cup S_0}(X'_{\{1\}\cup S_0} X_{\{1\}\cup S_0})^{-1} z^0_{\{1\}\cup S_0}\|_\infty = \|X'_R A_1 X_{S_0}(X'_{S_0} A_1 X_{S_0})^{-1} z^0_{S_0}\|_\infty.$$

*Proof of Lemma 8.5.* See Appendix D. □

This means that for tree graphs the irrepresentable condition can be checked for the "active set" $\{1\} \cup S_0$ instead of $S_0$, but then the first column has to be neglected. This fact is justified, however in a different way than the one we propose, in Qian and Jia (2016) as well.

**Remark** (The irrepresentable condition for asymptotic pattern recovery of a signal on a graph does not depend on the orientation of the edges of the graph)**.** We assume the linear model $Y = f^0 + \epsilon, \epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{I}_n)$. Then the Edge Lasso can be written as

$$\hat{f} = \arg\min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_n^2 + 2\lambda \|(\tilde{I}\tilde{D}f)_{-1}\|_1 \right\},$$

where

$$\tilde{I} \in \mathcal{I} = \left\{ \tilde{I} \in \mathbb{R}^n, \tilde{I} \text{ diagonal}, \mathrm{diag}(\tilde{I}) \in \{1, -1\}^n \right\}.$$

Define $\beta = \tilde{I}\tilde{D}f$. Then $f = X\tilde{I}\beta$. The linear model assumed becomes $Y = X\tilde{I}\beta^0 + \epsilon$ and the estimator

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \left\{ \|Y - X\tilde{I}\beta\|_n^2 + 2\lambda \|\beta_{-1}\|_1 \right\}, \tilde{I} \in \mathcal{I}.$$

It is clear that now the design matrix is $X\tilde{I}$. Let us write, without loss of generality,

$$\tilde{I} = \begin{pmatrix} \tilde{I}_{\{1\} \cup S_0} & 0 \\ 0 & \tilde{I}_{-(\{1\} \cup S_0)} \end{pmatrix}.$$

According to the Lemma 8.5 we can check if $\exists \eta \in (0, 1]$:

$$\|\tilde{I}_{-(\{1\} \cup S_0)} X'_{-(\{1\} \cup S_0)} (X'_{\{1\} \cup S_0} X_{\{1\} \cup S_0})^{-1} \tilde{I}_{\{1\} \cup S_0} \tilde{z}^0_{\{1\} \cup S_0}\|_\infty \leq 1 - \eta,$$

where $\tilde{z}^0_{\{1\} \cup S_0} = \begin{pmatrix} 0 \\ \tilde{z}^0_{S_0} \end{pmatrix}$ and $\tilde{z}^0_{S_0} = \mathrm{sgn}(\beta^0_{S_0}) = \tilde{I}_{S_0} \mathrm{sgn}(\tilde{D}f^0) = \tilde{I}_{S_0} \mathrm{sgn}(\bar{\beta}^0)$, where $\bar{\beta}^0 = \tilde{D}f^0$, i.e. the vector of truly nonzero jumps when the root has sign $+1$ and the edges are oriented away from it.

Note that $\tilde{I}_{-(\{1\} \cup S_0)}$ does not change the $\ell^\infty$-norm and by inserting the expression for $\tilde{z}^0_{\{1\} \cup S_0}$ we get that, $\forall \tilde{I} \in \mathcal{I}$,

$$\left\| \tilde{I}_{-(\{1\} \cup S_0)} X'_{-(\{1\} \cup S_0)} (X'_{\{1\} \cup S_0} X_{\{1\} \cup S_0})^{-1} \tilde{I}_{\{1\} \cup S_0} \begin{pmatrix} 0 & \\ & \tilde{I}_{S_0} \end{pmatrix} \begin{pmatrix} 0 \\ \bar{z}^0_{S_0} \end{pmatrix} \right\|_\infty \leq 1 - \eta,$$

where $\bar{z}^0_{S_0} = \mathrm{sgn}(\bar{\beta}^0)$. This means that it is enough to check that $\exists \eta > 0$:

$$\left\| X'_{-(\{1\} \cup S_0)} (X'_{\{1\} \cup S_0} X_{\{1\} \cup S_0})^{-1} \begin{pmatrix} 0 \\ \bar{z}^0_{S_0} \end{pmatrix} \right\|_\infty \leq 1 - \eta, \forall \tilde{I} \in \mathcal{I}$$

to know, for all the orientations of the graph, whether the irrepresentable condition holds. The intuition behind this is that, by choosing the orientation of the edges of the graph, we choose at the same time the sign that the true jumps have across the edges.

### 8.3. Irrepresentable condition for the path graph

**Theorem 8.6** (Irrepresentable condition for the transformed Fused Lasso, Theorem 2 in Qian and Jia (2016))**.** *Consider the model for a piecewise constant signal and let $S_0$ denote the set of indices of the jumps in the true signal, i.e.*

$$S_0 = \left\{ j : f_j^0 \neq f_{j-1}^0, j = 2, \cdots, n \right\} = \{ i_1, \cdots, i_{s_0} \},$$

*with $s_0 = |S_0|$ denoting its cardinality. The irrepresentable condition for the Edge Lasso on the path graph holds if and only if one of the two following conditions hold:*

- *The jump points are consecutive, i.e. $s_0 = 1$ or $\max_{2 \leq k \leq s_0}(i_k - i_{k-1}) = 1$.*
- *All the jumps between constant signal blocks have alternating signs, i.e.*

$$(f_{i_k}^0 - f_{i_k - 1}^0)(f_{i_{k+1}}^0 - f_{i_{k+1}-1}^0) < 0, k = 2, \cdots, s_0 - 1.$$

**Remark.** This fact can as well be easily read out from the consideration made in Section 4 and in particular in Lemma 4.1.

### 8.4. Irrepresentable condition for the path graph with one branch

**Corollary 8.7** (Irrepresentable condition for the branched path graph)**.** *Assume $S_0 \neq 0$. The irrepresentable condition for the branched path graph is satisfied if and only if one of the following cases holds,*

- *$s_0 = n - 1$ or $s_0 = 1$;*
- *$sgn(\beta_{i_{s_1}}^0) = -sgn(\beta_{j_1}^0) = -sgn(\beta_{k_1}^0)$ and in the subvectors $\beta_{1:n_1}^0$ and $\beta_{(b,n_1+1:n)}^0$ there are no two consecutive nonzero entries of $\beta^0$ with the same sign being separated by some zero entry.*

*Note that:*

- *If $i_{s_1} = b$, then the requirement above is relaxed to $sgn(\beta_{j_1}^0) = sgn(\beta_{k_1}^0)$;*
- *If $j_1 = b + 1$, then the requirement above is relaxed to $sgn(\beta_{i_{s_1}}^0) = -sgn(\beta_{k_1}^0)$;*
- *If $k_1 = n_1 + 1$, then the requirement above is relaxed to $sgn(\beta_{i_{s_1}}^0) = -sgn(\beta_{j_1}^0)$.*

*Proof of Lemma 8.7.* This is a special case of Theorem 8.8 and follows directly from it. □

### 8.5. The irrepresentable condition for general branching points

When the graph $\mathcal{G}$ has a branching point where arbitrarily many branches are attached, for the irrepresentable condition to be satisfied it is required, in addition to the absence of staircase patterns along the path graphs building $\mathcal{G}$, that

the last jump in the path graph containing the branching point has sign + (resp. −) and all the first jumps in the other path graphs glued to this branching point have sign − (resp. +), with respect to the orientation of the edges away from the root. For the index of the $K+1$ jumps surrounding the branching point we use the same notation as in Subesction 4.2, i.e we denote them by $\{j_1, \dots, j_{K+1}\}$.

**Theorem 8.8.** *Consider the Edge Lasso estimator on a general "large enough" tree graph. The irrepresentable condition for the corresponding (almost) ordinary Lasso problem is satisfied if and only if for the signal on the path graphs connected at the branching points the conditions of Theorem 8.6 hold and for the true signal around any branching point involving $K+1$ edges, the jump just before it and the jumps right after it have opposite signs. More formally, this last condition writes:*

1. $sgn(j_1)sgn(j_l) < 0, \forall l \in \left\{ l^* \in \{2, \dots, K+1\}, \tilde{d}_1^{l^*} \neq 0 \right\}$
2. *and* $sgn(j_l)sgn(j_{l'}) > 0, \forall l, l' \in \left\{ l^* \in \{2, \dots, K+1\}, \tilde{d}_1^{l^*} \neq 0 \right\}$.
3. *and* $\tilde{d}^1_{s_1+1} - 1, \tilde{d}_1^2, \dots, \tilde{d}_1^{K+1} < \frac{2}{K+1} d^*$.

*Note that if $\tilde{d}^1_{s_1+1} = 1$, then the condition requiring that $sgn(j_1)sgn(j_l) < 0$, for all $l \in \left\{ l^* \in \{2, \dots, K+1\}, \tilde{d}_1^{l^*} \neq 0 \right\}$ is removed.*

*Proof of Theorem 8.8.* See Appendix D. □

## 9. Conclusion

We refined some details of the approach of Dalalyan, Hebiri and Lederer (2017) for proving a sharp oracle inequality for the total variation regularized estimator over the path graph. In particular we decided to follow an approach where a coefficient is left unpenalized and we gave a proof of a lower bound on the compatibility constant which does not use probabilistic arguments. The key point of this article is that we proved that the approach applied on the path graph can indeed be generalized to a branched graph and further to more general tree graphs. In particular we found a lower bound on the compatibility constant and we generalized the result concerning the irrepresentable condition obtained for the path graph by Qian and Jia (2016).

## Appendix A: Proofs of Section 2

*Proof of Theorem 2.5.* **Deterministic part**

Recall the definition of the estimator

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \|\beta_{-1}\|_1 \right\}.$$

The KKT conditions are

$$\frac{1}{n} X'(Y - X\widehat{\beta}) = \lambda \widehat{z}_{-1}, \widehat{z}_{-1} \in \partial \|\widehat{\beta}_{-1}\|_1,$$

where $\widehat{z}_{-1} \in \mathbb{R}^n$ is a vector with the first entry equal to zero and the remaining ones equal to the subdifferential of the absolute value of the corresponding entry of $\widehat{\beta}$. Inserting $Y = X\beta^0 + \epsilon$ into the KKT conditions and multiplying them once by $\widehat{\beta}$ and once by $\beta$ we obtain

$$-\frac{1}{n}\widehat{\beta}' X'(X(\widehat{\beta} - \beta^0) - \epsilon) = \lambda \|\widehat{\beta}_{-1}\|_1$$

and

$$-\frac{1}{n}\beta' X'(X(\widehat{\beta} - \beta^0) - \epsilon) = \lambda \beta'_{-1}\widehat{z}_{-1} \leq \lambda \|\beta_{-1}\|_1,$$

where the last inequality follows by the dual norm inequality and the fact that $\|\widehat{z}_{-1}\|_\infty \leq 1$. Subtracting the first inequality from the second we get

$$\frac{1}{n}(\widehat{\beta} - \beta)' X'(X(\widehat{\beta} - \beta^0) - \epsilon) \leq \lambda(\|\beta_{-1}\|_1 - \|\widehat{\beta}_{-1}\|_1).$$

Using polarization we obtain

$$\begin{aligned} \|X(\widehat{\beta} - \beta)\|_n^2 + \|X(\widehat{\beta} - \beta^0)\|_n^2 &\leq \|X(\beta - \beta^0)\|_n^2 + \frac{2}{n}(\widehat{\beta} - \beta)' X'\epsilon \\ &+ 2\lambda\left(\|\beta_{-1}\|_1 - \|\widehat{\beta}_{-1}\|_1\right). \end{aligned}$$

Let $S \subset \{2, \ldots, n\}$. We have that

$$\begin{aligned} \|\beta_{-1}\|_1 - \|\widehat{\beta}_{-1}\|_1 &= \|\beta_S\|_1 - \|\widehat{\beta}_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1 - \|\widehat{\beta}_{-(\{1\}\cup S)}\|_1 \\ &+ 2\|\beta_{-(\{1\}\cup S)}\|_1 \\ &\leq \|\beta_S - \widehat{\beta}_S\|_1 - \|\beta_{-(\{1\}\cup S)} - \widehat{\beta}_{-(\{1\}\cup S)}\|_1 \\ &+ 2\|\beta_{-(\{1\}\cup S)}\|_1. \end{aligned}$$

Thus we get the "basic" inequality

$$\begin{aligned} &\|X(\widehat{\beta} - \beta)\|_n^2 + \|X(\widehat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta - \beta^0)\|_n^2 + 4\lambda\|\beta_{-(\{1\}\cup S)}\|_1 \\ &+ \underbrace{\frac{2}{n}(\widehat{\beta} - \beta)' X'\epsilon + 2\lambda\left(\|(\beta - \widehat{\beta})_S\|_1 - \|(\beta - \widehat{\beta})_{-(\{1\}\cup S)}\|_1\right)}_{\mathrm{I}}. \end{aligned}$$

We are going to utilize the approach described by Dalalyan, Hebiri and Lederer (2017) to handle the remainder term I with care. Since $\mathrm{I}_n = \Pi_{\{1\}\cup S} + A_{\{1\}\cup S}$, it follows that

$$(\widehat{\beta} - \beta)' X'\epsilon = (\widehat{\beta} - \beta)' X'\Pi_{\{1\}\cup S}\epsilon + (\widehat{\beta} - \beta)'_{-(\{1\}\cup S)} X'_{-(\{1\}\cup S)} A i_{\{1\}\cup S}\epsilon.$$

Indeed the antiprojection of elements of $V_{\{1\}\cup S}$ is zero. Note that

$$(\widehat{\beta} - \beta)'_{-(\{1\}\cup S)} X'_{-(\{1\}\cup S)} A_{\{1\}\cup S}\epsilon \leq \sum_{j \in -(\{1\}\cup S)} |\widehat{\beta} - \beta|_j |\epsilon' A_{\{1\}\cup S} X_j|.$$

Restricting ourselves to the set

$$F = \left\{ |\epsilon' A_{\{1\}\cup S} X_j| \leq \frac{\lambda n}{\gamma} \|A_{\{1\}\cup S} X_j\|_n, \forall j \notin (\{1\} \cup S) \right\},$$

for $\gamma \geq 1$ we obtain

$$\begin{aligned} \mathrm{I} &\leq \frac{2}{n}(\widehat{\beta}-\beta)' X' \Pi_{\{1\}\cup S}\epsilon \\ &+ 2\lambda \left( \|(\widehat{\beta}-\beta)_S\|_1 - \|(\widehat{\beta}-\beta)_{-(\{1\}\cup S)}\|_1 + \|(\frac{\omega}{\gamma} \odot (\widehat{\beta}-\beta))_{-(\{1\}\cup S)}\|_1 \right) \\ &\leq 2\frac{\|X(\widehat{\beta}-\beta)\|_2}{\sqrt{n}} \frac{\|\Pi_{\{1\}\cup S}\epsilon\|_2}{\sqrt{n}} \\ &+ 2\lambda \left( \|(\widehat{\beta}-\beta)_S\|_1 - \|(w \odot (\widehat{\beta}-\beta))_{-(\{1\}\cup S)}\|_1 \right), \end{aligned}$$

Using the definition of the weighted compatibility constant and the convex conjugate inequality we obtain

$$\begin{aligned} \mathrm{I} &\leq 2\frac{\|X(\widehat{\beta}-\beta)\|_2}{\sqrt{n}} \left( \frac{\|\Pi_{\{1\}\cup S}\epsilon\|_2}{\sqrt{n}} + \lambda \frac{\sqrt{s+1}}{\kappa_w(S)} \right) \\ &\leq \|X(\widehat{\beta}-\beta)\|_n^2 + \left( \frac{\|\Pi_{\{1\}\cup S}\epsilon\|_2}{\sqrt{n}} + \lambda \frac{\sqrt{s+1}}{\kappa_w(S)} \right)^2. \end{aligned}$$

We see that $\|X(\widehat{\beta}-\beta)\|_n^2$ cancels out and we are left with

$$\begin{aligned} \|X(\widehat{\beta}-\beta^0)\|_n^2 &\leq \inf_{\beta\in\mathbb{R}^n} \left\{ \|X(\beta-\beta^0)\|_n^2 + 4\lambda \|\beta_{-(\{1\}\cup S)}\|_1 \right\} \\ &+ \left( \frac{\|\Pi_{\{1\}\cup S}\epsilon\|_2}{\sqrt{n}} + \lambda \frac{\sqrt{s+1}}{\kappa_w(S)} \right)^2. \end{aligned}$$

It now remains to find a lower bound for $\mathbb{P}(F)$ and a high-probability upper bound for $\|\Pi_{\{1\}\cup S}\epsilon\|_n^2$.

**Random part**

- First, we lower bound $\mathbb{P}(F)$, thanks to the following lemma.

  **Lemma A.1** (The maximum of $p$ random variables, Lemma 17.5 in van de Geer (2016))**.** *Let $V_1, \ldots, V_p$ be real valued random variables. Assume that $\forall j \in \{1, \ldots, p\}$ and $\forall r > 0$*

  $$\mathbb{E}\left[ e^{r|V_j|} \right] \leq 2e^{\frac{r^2}{2}}.$$

  *Then, $\forall t > 0$*

  $$\mathbb{P}\left( \max_{1\leq j\leq p} |V_j| \geq \sqrt{2\log(2p) + 2t} \right) \leq e^{-t}.$$

We now apply Lemma A.1 to $F$. Note that $F$ can be written as

$$F = \left\{ \max_{j \in -(\{1\} \cup S)} \left| \frac{\epsilon' A_{\{1\} \cup S} X_j}{\sigma \| A_{\{1\} \cup S} X_j \|_2} \right| \leq \frac{\lambda \sqrt{n}}{\gamma \sigma} \right\}.$$

Since $X_j' A_{\{1\} \cup S} \epsilon \sim \mathcal{N}(0, \sigma^2 \| X_j' A_{\{1\} \cup S} \|_2^2)$, we obtain that for the standard normal random variables $V_1, \ldots, V_{n-s-1} \sim \mathcal{N}(0,1)$

$$F = \left\{ \max_{1 \leq j \leq n-s-1} |V_j| \leq \frac{\lambda \sqrt{n}}{\gamma \sigma} \right\}.$$

The moment generating function of $|V_j|$ is

$$\mathbb{E}\left[ e^{r|Vj|} \right] = 2(1 - \Phi(-r)) e^{\frac{r^2}{2}} \leq 2 e^{\frac{r^2}{2}}, \forall r > 0$$

Choosing, for some $\delta \in (0,1)$, $\lambda = \gamma \sigma \sqrt{2 \log\left(4(n-s-1)/\delta\right)/n}$ and applying Lemma A.1 with $p = n - s - 1$ and $t = \log\left(\frac{2}{\delta}\right)$, we obtain

$$\mathbb{P}(F) \geq 1 - \delta/2.$$

- Second, we are going to find an high probability upper bound for

$$\| \Pi_{\{1\} \cup S} \epsilon \|_n^2 = \frac{\sigma^2}{n} \underbrace{\frac{\| \Pi_{\{1\} \cup S} \epsilon \|_2^2}{\sigma^2}}_{\sim \chi^2_{s+1}},$$

where $\operatorname{rank}(\Pi_{\{1\} \cup S}) = s + 1$. We use Lemma 8.6 in van de Geer (2016), which reproves part of Lemma 1 in Laurent and Massart (2000).

**Lemma A.2** (The special case of $\chi^2$ random variables, Lemma 1 in Laurent and Massart (2000), Lemma 8.6 in van de Geer (2016))**.** *Let* $X \sim \chi^2_d$. *Then,* $\forall t > 0$

$$\mathbb{P}\left( X \geq d + 2\sqrt{dt} + 2t \right) \leq e^{-t}$$

Note that from Lemma A.2 it follows that

$$\mathbb{P}\left( \sqrt{X} \leq \sqrt{d} + \sqrt{2t} \right) \geq \mathbb{P}\left( X \leq d + 2\sqrt{dt} + 2t \right) \geq 1 - e^{-t}$$

Define

$$G := \left\{ \frac{\| \Pi_{\{1\} \cup S} \epsilon \|_2}{\sqrt{n}} \leq \sqrt{\frac{\sigma^2}{n}} \left( \sqrt{s+1} + \sqrt{2 \log(2/\delta)} \right) \right\}.$$

By applying Lemma A.2 with $t = \log(2/\delta)$, for some $\delta \in (0,1)$, we get

$$\mathbb{P}(G) \geq 1 - \delta/2.$$

If we choose $\lambda = \gamma\sigma\sqrt{2\log\left(4(n-s-1)/\delta\right)/n}$ and apply twice the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we get that with probability $\mathbb{P}(F \cap G) \geq 1-\delta$ the following oracle inequality holds

$$\begin{aligned}\|X(\widehat{\beta}-\beta^0)\|_n^2 &\leq \inf_{\beta\in\mathbb{R}^n}\left\{\|X(\beta-\beta^0)\|_n^2 + 4\lambda\|\beta_{-(\{1\}\cup S)}\|_1\right\} \\ &+ \frac{8\sigma^2}{n}\log(2/\delta) \\ &+ \frac{4\sigma^2}{n}\left((s+1)+\frac{\gamma^2(s+1)}{\kappa_w^2(S)}\log\left(4(n-s-1)/\delta\right)\right).\end{aligned}$$

The statement of the theorem is obtained using the identity $f = X\beta$. □

## Appendix B: Proofs of Section 5

Let $f \in \mathbb{R}^n$ be a function defined at every vertex of a connected nondegenerate graph $\mathcal{G}$. Moreover let

$$f_{(n)} \geq \ldots \geq f_{(1)}$$

be an ordering of $f$, with arbitrary order within tuples. Let $D$ denote the incidence matrix of the graph $\mathcal{G}$.

**Lemma B.1** (Lemma 11.9 in van de Geer (2018)). *It holds that*

$$\|Df\|_1 \geq f_{(n)} - f_{(1)}.$$

**Remark.** For the special case of $\mathcal{G}$ being the path graph, we have equality in Lemma B.1 when $f$ is nonincreasing or nondecreasing on the graph.

*Proof of Lemma B.1.* Since $\mathcal{G}$ is connected there is a path between any two vertices. Therefore there is a path connecting the vertices where $f$ takes the values $f_{(n)}$ and $f_{(1)}$. The total variation of a function defined on a graph is nondecreasing in the number of edges of the graph. Let us now consider $f_P$, the restriction of $f$ on a path $P$ connecting $f_{(1)}$ to $f_{(n)}$. If $f$ is nondecreasing on the path $P$, then $\|Df_P\|_1 = f_{(n)} - f_{(1)}$, otherwise $\|Df_P\|_1 \geq f_{(n)} - f_{(1)}$. Since $\mathcal{G}$ has at least as many edges as $P$:

$$\|Df\|_1 \geq \|Df_P\|_1 \geq f_{(n)} - f_{(1)}.$$ □

**Lemma B.2** (Lemma 11.10 in van de Geer (2018)). *It holds for any* $j \in \{1,\ldots,n\}$ *that*

$$f_j - \|Df\|_1 \leq f_{(1)} \leq \frac{1}{n}\sum_{i=1}^n |f_i|,$$

*and*

$$-f_j - \|Df\|_1 \leq -f_{(n)} \leq \frac{1}{n}\sum_{i=1}^n |f_i|.$$

*Proof of Lemma B.2.* For completeness and readability, we report the proof which can also be found in van de Geer (2018).

We have from Lemma B.2 that $\|Df\|_1 \geq f_{(n)} - f_{(1)}$. Moreover, $f_j \leq f_{(n)}$. Thus

$$\begin{aligned} f_j - \|Df\|_1 &\leq f_j - (f_{(n)} - f_{(1)}) \\ &\leq f_{(n)} - (f_{(n)} - f_{(1)}) \\ &= f_{(1)}. \end{aligned}$$

Case 1: if $f_{(1)} < 0$, obviously $f_{(1)} < \frac{1}{n}\sum_{i=1}^n |f_i|$.

Case 2: if $f_{(1)} \geq 0$, then $f_i \geq 0$ for all $i$ and then

$$f_{(1)} \leq \sum_{i=1}^n f_i/n = \sum_{i=1}^n |f_i|/n.$$

In the same way

$$\begin{aligned} -f_j - \|Df\|_1 &\leq -f_j - (f_{(n)} - f_{(1)}) \\ &\leq -f_{(1)} - (f_{(n)} - f_{(1)}) \\ &= -f_{(n)}. \end{aligned}$$

Case 1: if $f_{(n)} > 0$, then $-f_{(n)} < \frac{1}{n}\sum_{i=1}^n |f_i|$.

Case 2: if $f_{(n)} \leq 0$, then $f_i \leq 0$ for all $i$ and then

$$-f_{(n)} \leq -\sum_{i=1}^n f_i/n = \sum_{i=1}^n |f_i|/n. \qquad \square$$

**Lemma B.3** (Lemma 11.11 in van de Geer (2018))**.** *Let $f \in \mathbb{R}^n$ be defined over a connected graph $\mathcal{G}_f$ whose incidence matrix is $D_f$. The total variation of $f$ is $\|D_f f\|_1$. Analogously, let $g \in \mathbb{R}^m$ be defined over a connected graph $\mathcal{G}_g$ whose incidence matrix is $D_g$. The total variation of $g$ is $\|D_g g\|_1$. Then for any $j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$*

$$|f_j - g_k| - \|D_f f\|_1 - \|D_g g\|_1 \leq \frac{1}{n}\sum_{i=1}^n |f_i| + \frac{1}{m}\sum_{i=1}^m |g_i|.$$

*Proof of Lemma B.3.* Suppose without loss of generality that $f_j \geq g_k$. Then by Lemma B.2

$$\begin{aligned} |f_j - g_k| - \|D_f f\|_1 - \|D_g g\|_1 &= \underbrace{(f_j - \|D_f f\|_1)}_{\leq \sum_{i=1}^n |f_i|/n} + \underbrace{(-g_k - \|D_g g\|_1)}_{\leq \sum_{i=1}^m |g_i|/m} \\ &\leq \frac{1}{n}\sum_{i=1}^n |f_i| + \frac{1}{m}\sum_{i=1}^m |g_i|. \end{aligned}$$

$\square$

*Proof of Lemma 5.3.* For completeness and readability, we report the proof by van de Geer (2018).

We may write for $f = X\beta$,

$$
\begin{aligned}
& \|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1 \\
\leq\ & |f_{d_1+1} - f_{d_1}| - \sum_{i=2}^{d_1} |f_i - f_{i-1}| - \sum_{i=d_1+2}^{d_1+u_2} |f_i - f_{i-1}| \\
+\ & |f_{d_1+d_2+1} - f_{d_1+d_2}| - \sum_{i=d_1+u_2+2}^{d_1+d_2} |f_i - f_{i-1}| - \sum_{i=d_1+d_2+2}^{d_1+d_2+u_3} |f_i - f_{i-1}| \\
& \dots \\
+\ & |f_{d_1+\cdots+d_{s-1}+1} - f_{d_1+\cdots+d_{s-1}}| \\
-\ & \sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+2}^{d_1+\cdots+d_{s-1}} |f_i - f_{i-1}| - \sum_{i=d_1+\cdots+d_{s-1}+2}^{d_1+\cdots+d_{s-1}+u_s} |f_i - f_{i-1}| \\
+\ & |f_{d_1+\cdots+d_s+1} - f_{d_1+\cdots+d_s}| \\
-\ & \sum_{i=d_1+\cdots+d_{s-1}+u_s+2}^{d_1+\cdots+d_s} |f_i - f_{i-1}| - \sum_{i=d_1+\cdots+d_s+2}^{n} |f_i - f_{i-1}|
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & \frac{1}{d_1} \sum_{i=1}^{d_1} |f_i| + \frac{1}{u_2} \sum_{i=d_1+1}^{d_1+u_2} |f_i| \\
+\ & \frac{1}{d_2 - u_2} \sum_{i=d_1+u_2+1}^{d_1+d_2} |f_i| + \frac{1}{u_3} \sum_{i=d_1+d_2+1}^{d_1+d_2+u_3} |f_i| \\
& \dots \\
+\ & \frac{1}{d_{s-1} - u_{s-1}} \sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}} |f_i| + \frac{1}{u_s} \sum_{i=d_1+\cdots+d_{s-1}+1}^{d_1+\cdots+d_{s-1}+u_s} |f_i| \\
+\ & \frac{1}{d_s - u_s} \sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s} |f_i| + \frac{1}{d_{s+1}} \sum_{i=d_1+\cdots+d_s+1}^{n} |f_i|
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & \sqrt{\underbrace{\frac{1}{d_1} + \frac{1}{u_2} + \frac{1}{d_2 - u_2} + \cdots + \frac{1}{d_{s-1} - u_{s-1}} + \frac{1}{u_s} + \frac{1}{d_s - u_s} + \frac{1}{d_{s+1}}}_{=:K}} \\
& \times \sqrt{\sum_{i=1}^{n} |f_i|^2},
\end{aligned}
$$

where the last step follows from the Cauchy-Schwarz inequality. We thus infer Lemma 5.3, since

$$\frac{s+1}{nK} \leq \frac{(s+1)\|f\|_2^2}{n(\|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1)^2}, \forall \beta \in \mathbb{R}^n$$

implies that

$$\frac{s+1}{nK} \leq \kappa^2(S).$$

Note that Lemma B.3 is used in the proof. The idea behind the proof is to cut the graph into smaller pieces of length $u_j$ and $d_j - u_j$ respectively. The places of these cuts are the edges indexed by $S$. Then Lemma B.3 is applied to obtain terms to which one can apply the Cauchy-Schwarz inequality to finally obtain the term $\|f\|_2$ multiplied by some factor. Notice also that in the first inequality some edges indexed by $-(\{1\} \cup S)$ are left out. Indeed we want each $f_i$ to be part of the average of only one piece of graph when we apply Lemma B.3, to get a more convenient expression after applying the Cauchy-Schwarz inequality. Thus the path graph is cut into $2s$ smaller path graphs. Consecutive pairs of path graphs are then used to apply Lemma B.3. □

*Proof of Corollary 5.4.* We report, for completeness and readability, the proof of Theorem 6.1 in van de Geer (2018).

Note that by the definition of $f^* = X\beta^*$,

$$\begin{aligned}
\|\beta_S^*\|_1 = \sum_{j=1}^{s} |f_{d_j+1}^* - f_{d_j}^*| &= \frac{n}{d_1} + \frac{2n}{d_2} \\
&+ \frac{2n}{d_2} + \frac{2n}{d_3} \\
&\vdots \\
&+ \frac{2n}{d_{s-1}} + \frac{2n}{d_s} \\
&+ \frac{2n}{d_s} + \frac{n}{d_{s+1}} \\
&= \frac{n}{d_1} + 4\sum_{j=2}^{s} \frac{n}{d_j} + \frac{n}{d_{s+1}},
\end{aligned}$$

and also

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} f_i^{*2} &= \frac{d_1}{n} {f_{d_1}^*}^2 + \cdots + \frac{d_{s+1}}{n} {f_{d_{s+1}}^*}^2 \\
&= \frac{n}{d_1} + 4\sum_{j=2}^{s} \frac{n}{d_j} + \frac{n}{d_{s+1}}.
\end{aligned}$$

Note also that

$$\begin{aligned}
&\|\beta^*_{-(\{1\}\cup S)}\|_1 \\
= &\sum_{i=2}^{d_1} |f_i^* - f_{i-1}^*| + \sum_{i=d_1+2}^{d_2} |f_i^* - f_{i-1}^*| + \cdots + \sum_{i=d_1+\cdots+d_s+2}^{n} |f_i^* - f_{i-1}^*| \\
= &\ 0
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{(s+1)\|X\beta^*\|_2^2}{n(\|\beta^*_S\|_1 - \|\beta^*_{-(\{1\}\cup S)}\|_1)^2} &= \frac{(s+1)\sum_{i=1}^n f_i^{*2}}{n\left(\sum_{j=1}^s |f^*_{d_j+1} - f^*_{d_j}|\right)^2} \\
&= \frac{s+1}{\frac{n}{d_1} + \sum_{j=2}^s \frac{4n}{d_j} + \frac{n}{d_{s+1}}}.
\end{aligned}$$

□

*Proof of Lemma 5.5.* For completeness and readability, we report the proof of Lemma 9.1 in van de Geer (2018). See Appendix B.1 for an intuition.

Let $g_i := w_i f_i$, $i = 1, \ldots, n$.

We have that

$$\begin{aligned}
&\|(w \odot \beta)_S\|_1 - \|(w \odot \beta)_{-(\{1\}\cup S)}\|_1 \\
&= \textstyle\sum_{j=1}^s w_{d_1+\ldots+d_j+1} |f_{d_1+\ldots+d_j+1} - f_{d_1+\ldots+d_j}| \\
&- \textstyle\sum_{i=2}^{d_1} w_i |f_i - f_{i-1}| - \sum_{j=2}^{s-1} \sum_{i=d_1+\ldots+d_{j+1}}^{d_1+\ldots+d_{j+1}} w_i |f_i - f_{i-1}| \\
&- \textstyle\sum_{i=d_1+\ldots+d_s+1}^{n} w_i |f_i - f_{i-1}| \\
&\left.\begin{aligned}
&\leq |g_{d_1+1} - g_{d_1}| - \sum_{i=2}^{d_1} |g_i - g_{i-1}| - \sum_{i=d_1+2}^{d_1+u_2} |g_i - g_{i-1}| \\
&+ |g_{d_1+d_2+1} - g_{d_1+d_2}| - \sum_{i=d_1+u_2+2}^{d_1+d_2} |g_i - g_{i-1}| - \sum_{i=d_1+d_2+2}^{d_1+d_2+u_3} |g_i - g_{i-1}| \\
&\ldots \\
&+ |g_{d_1+\cdots+d_{s-1}+1} - g_{d_1+\cdots+d_{s-1}}| \\
&- \sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+2}^{d_1+\cdots+d_{s-1}} |g_i - g_{i-1}| - \sum_{i=d_1+\cdots+d_{s-1}+2}^{d_1+\cdots+d_{s-1}+u_s} |g_i - g_{i-1}| \\
&+ |g_{d_1+\cdots+d_s+1} - g_{d_1+\cdots+d_s}| \\
&- \sum_{i=d_1+\cdots+d_{s-1}+u_s+2}^{d_1+\cdots+d_s} |g_i - g_{i-1}| - \sum_{i=d_1+\cdots+d_s+2}^{n} |g_i - g_{i-1}|
\end{aligned}\right\} \text{I} \\
&+ \underbrace{\sum_{i\in I} |w_i - w_{i-1}||f_{i-1}|}_{\text{II}},
\end{aligned}$$

where

$$J = [n] \setminus \{1, d_1 + u_2 + 1, d_1 + d_2 + u_3 + 1, \ldots, d_1 + \ldots + d_{s-1} + u_s\}$$

Moreover, by Lemma B.3 and by the Cauchy-Schwarz inequality

$$\begin{aligned}
\mathrm{I} \;\leq\; & \frac{1}{d_1}\sum_{i=1}^{d_1}|g_i| + \frac{1}{u_2}\sum_{i=d_1+1}^{d_1+u_2}|g_i| + \frac{1}{d_2-u_2}\sum_{i=d_1+u_2+1}^{d_1+d_2}|g_i| \\
+ \;& \frac{1}{u_3}\sum_{i=d_1+d_2+1}^{d_1+d_2+u_3}|g_i| + \cdots + \\
+ \;& \frac{1}{d_{s-1}-u_{s-1}}\sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}}|g_i| + \frac{1}{u_s}\sum_{i=d_1+\cdots+d_{s-1}+1}^{d_1+\cdots+d_{s-1}+u_s}|g_i| \\
+ \;& \frac{1}{d_s-u_s}\sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s}|g_i| + \frac{1}{d_{s+1}}\sum_{i=d_1+\cdots+d_s+1}^{n}|g_i| \\
\leq \;& \Bigg(\frac{1}{d_1^2}\sum_{i=1}^{d_1}w_i^2 + \frac{1}{u_2^2}\sum_{i=d_1+1}^{d_1+u_2}w_i^2 \\
+ \;& \frac{1}{(d_2-u_2)^2}\sum_{i=d_1+u_2+1}^{d_1+d_2}w_i^2 + \frac{1}{u_3^2}\sum_{i=d_1+d_2+1}^{d_1+d_2+u_3}w_i^2 + \cdots + \\
+ \;& \frac{1}{(d_{s-1}-u_{s-1})^2}\sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}}w_i^2 + \frac{1}{u_s^2}\sum_{i=d_1+\cdots+d_{s-1}+1}^{d_1+\cdots+d_{s-1}+u_s}w_i^2 \\
+ \;& \frac{1}{(d_s-u_s)^2}\sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s}w_i^2 \\
+ \;& \frac{1}{d_{s+1}^2}\sum_{i=d_1+\cdots+d_s+1}^{n}w_i^2\Bigg)^{1/2} \times \Bigg(\sum_{i=1}^{n}f_i^2\Bigg)^{1/2} \\
\leq \;& \sqrt{\frac{1}{d_1}+\frac{1}{u_2}+\frac{1}{d_2-u_2}+\cdots+\frac{1}{d_{s-1}-u_{s-1}}+\frac{1}{u_s}+\frac{1}{d_s-u_s}+\frac{1}{d_{s+1}}} \\
& \times\|w\|_\infty \times \sqrt{\sum_{i=1}^{n}|f_i|^2}.
\end{aligned}$$

and by the Cauchy-Schwarz inequality

$$\begin{aligned}
\mathrm{II} \;&\leq\; \sqrt{\sum_{i\in J}(w_i - w_{i-1})^2}\sqrt{\sum_{i\in J}f_{i-1}^2} \\
&\leq\; \sqrt{\sum_{i=2}^{n}(w_i - w_{i-1})^2}\sqrt{\sum_{i=1}^{n}f_i^2}.
\end{aligned}$$

We thus infer Lemma 5.5. □

*Proof of Corollary 5.6.* Let $A_{\{1\}\cup S} = \mathrm{I}_n - \Pi_{\{1\}\cup S}$ denote the antiprojection matrix on the coulmns of $X$ indexed by $\{1\} \cup S$. By using the definition of $w_i$ and $\omega_i$, we have that

$$\begin{aligned}\|Dw\|_2^2 = \sum_{i=2}^{n}(w_i - w_{i-1})^2 &= \frac{1}{\gamma^2}\sum_{i=2}^{n}(\omega_i - \omega_{i-1})^2 \\ &= \frac{1}{\gamma^2 n}\sum_{i=2}^{n}(\|A_{\{1\}\cup S}X_i\|_2 - \|A_{\{1\}\cup S}X_{i-1}\|_2)^2.\end{aligned}$$

Let us define the function $f(x) = -2x^2 + 2(c+1)x - (c+1)$, where $c > 0$ is a positive constant.

For the path graph we have, thanks to Section 4,

$$\begin{aligned}\sum_{i=2}^{n}(w_i - w_{i-1})^2 &= \frac{1}{n\gamma^2}\sum_{j=1}^{s+1}\sum_{i=1}^{d_j}\frac{(\sqrt{i(d_j - i)} - \sqrt{(i-1)(d_j - (i-1))})^2}{d_j} \\ &= \frac{1}{n\gamma^2}\sum_{j=1}^{s+1}\sum_{i=1}^{d_j}\frac{(i(d_j - i) - (i-1)(d_j - (i-1)))^2}{d_j(\sqrt{i(d_j - i)} + \sqrt{(i-1)(d_j - (i-1))})^2} \\ &\leq \frac{1}{n\gamma^2}\sum_{j=1}^{s+1}\sum_{i=1}^{d_j}\frac{(-2i + d_j + 1)^2}{d_j(-2i^2 + 2(d_j+1)i - (d_j+1))} \\ &\leq \frac{1}{n\gamma^2}\sum_{j=1}^{s+1} d_j \sum_{i=1}^{d_j}\frac{1}{f(i)}.\end{aligned}$$

Now note that the function $f(x)$ is strictly concave, has two zeroes at $x = \frac{c+1}{2} \pm \frac{\sqrt{c^2-1}}{2}$ and a global maximum at $x = \frac{c+1}{2}$. We also note that $\forall c \geq 1$ the left zero point $\frac{c+1}{2} - \frac{\sqrt{c^2-1}}{2} \leq 1$. Moreover, $\forall x \in [1, c/2]$, $f(x) \geq cx$. Using the symmetry of quadratic functions around the global maximum we obtain, in partial analogy to Dalalyan, Hebiri and Lederer (2017),

$$\begin{aligned}\|Dw\|_2^2 = \sum_{i=2}^{n}(w_i - w_{i-1})^2 &\leq \frac{1}{\gamma^2 n}\sum_{j=1}^{s+1} d_j\left(\sum_{i=1}^{\lceil d_j/2\rceil}\frac{1}{d_j i} + \sum_{i=1}^{\lfloor d_j/2\rfloor}\frac{1}{d_j i}\right) \\ &\leq \frac{5}{2\gamma^2 n}\sum_{j=1}^{s+1}\log(\lceil d_j/2\rceil\lfloor d_j/2\rfloor) = \frac{5}{2\gamma^2 n}\log\left(\prod_{i=1}^{2(s+1)}|\Delta|_i\right) \\ &= \frac{5}{\gamma^2 n}(s+1)\log(|\bar{\Delta}|) \leq \frac{5}{\gamma^2 n}(s+1)\log(n/(2s+2)) \\ &\leq \frac{5}{\gamma^2 n}(s+1)\log(n/(s+1)),\end{aligned}$$

where $|\bar{\Delta}|$ is the geometric mean of $|\Delta|$, which is upper bounded by the arithmetic mean of $|\Delta|$, which is $n/(2s+2)$. Moreover the constant $5/2$ and the assumption $d_j \geq 4, \forall j \in [s+1]$ come from the fact that

$$\frac{\sum_{i=1}^{k} i^{-1}}{\log i}$$

is finite only if $i \geq 2$, is decreasing in $i$ and has value approximately 2.16 when $i = 2$. Moreover the vector $|\Delta| \in \mathbb{R}^{2s+2}$ is defined as

$$|\Delta| \in \mathbb{R}^{2s+2} = (\lfloor d_1/2 \rfloor, \lceil d_1/2 \rceil, \ldots, \lfloor d_{s+1}/2 \rfloor, \lceil d_{s+1}/2 \rceil)\,.$$

We now have to find an upper bound for $K$. Since the choice of $u_j$ is arbitrary, we choose $u_j = \lfloor d_j/2 \rfloor, j \in \{2, \ldots, s\}$, which minimize the upper bound among the integers. We thus have that $K \leq \frac{2s}{\bar{\Delta}_h}$, where $\bar{\Delta}_h$ is the harmonic mean of $\Delta$.

Finaly, for the path graph we have

$$\begin{aligned} \frac{1}{\kappa_w^2(S)} &\leq \frac{2n}{s+1}(K + \|Dw\|_2^2) \\ &\leq \frac{2n}{\gamma^2(s+1)}\left(\frac{2\gamma^2 s}{\bar{\Delta}_h} + 5\frac{s+1}{n}\log(n/(s+1))\right), \end{aligned}$$

and we obtain the Corollary 5.6. □

### *B.1. Outline of proofs by means of a minimal toy example*

For giving an intuition to the reader we present a minimal toy example. Consider the path graph with $n = 8$ and let $S = \{3, 7\}$. In this example $d_1 = 2, d_2 = 4, u_2 = 2, d_3 = 2$. We write

$$\begin{aligned} \|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1 &= |f_3 - f_2| - |f_2 - f_1| - |f_4 - f_3| \\ &\quad + |f_7 - f_6| - |f_6 - f_5| - |f_8 - f_7| \\ &\quad - |f_5 - f_4| \end{aligned}$$

The idea now is to apply Lemma B.3 twice. The first time we apply it to the path graphs $(\{1, 2\}, (1, 2))$ and $(\{3, 4\}, (3, 4))$. The second time we apply it to the path graphs $(\{5, 6\}, (5, 6))$ and $(\{7, 8\}, (7, 8))$. Note that the term $|f_5 - f_4|$ is not needed to apply Lemma B.3 and thus can be left out. We get

$$\|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1 \leq \frac{1}{2}\sum_{i=1}^{8}|f_i| \leq \sqrt{2}\|f\|_2,$$

where the last step follows by the Cauchy-Schwarz inequality. We thus see that we can handle graphs built by modules consisting of small path graphs containing an edge in $S$ and at least one vertex not involved in this edge on each side. The edges connecting these modules can then be neglected when upperbounding $\|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1$.

In the weighted case we define $g_i = w_i f_i, i = 1, \dots, 8$ and write

$$
\begin{aligned}
&\|(w \odot \beta)_S\|_1 - \|(w \odot \beta)_{-(\{1\} \cup S)}\|_1 \\
&\leq w_3|f_3 - f_2| - w_2|f_2 - f_1| - w_4|f_4 - f_3| \\
&+w_7|f_7 - f_6| - w_6|f_6 - f_5| - w_8|f_8 - f_7| \\
&\leq |g_3 - g_2| - |g_2 - g_1| - |g_4 - g_3| \\
&+|g_7 - g_6| - |g_6 - g_5| - |g_8 - g_7| \\
&+\sum_{i=2}^{4} |w_i - w_{i-1}||f_{i-1}| + \sum_{i=6}^{8} |w_i - w_{i-1}||f_{i-1}| \\
&\leq \sqrt{1/4 \|w\|_2^2} \|f\|_2 \\
&+\sqrt{\sum_{i=2}^{4} (w_i - w_{i-1})^2 + \sum_{i=6}^{8} (w_i - w_{i-1})^2} \sqrt{\sum_{i=1}^{3} f_i^2 + \sum_{i=5}^{7} f_i^2} \\
&\leq \left( \sqrt{2} \|w\|_\infty + \sqrt{\sum_{i=2}^{4} (w_i - w_{i-1})^2 + \sum_{i=6}^{8} (w_i - w_{i-1})^2} \right) \|f\|_2.
\end{aligned}
$$

Here as well, note that the squared difference of the weights across the edge connecting the two modules (smaller but large enough path graphs containing an element of $S$) can be neglected. The procedure exemplified here can be used to handle larger tree graphs, as long as one is able to decompose them in such smaller modules. The fact that squared weights differences can be neglected at the junction of modules will be of use in the proof of Corollary 6.6.

**Remark.** The limits of this approach are given by Lemma B.3, since its use requires the presence of at least a distinct edge not in $S$ on the left and on the right for each edge in $S$ not sharing vertices with edges used to handle other elements of $S$. Thus $s \leq n/4$. However, this limitation is very likely to be of scarce relevance if some kind of minimal length condition holds, see for instance Dalalyan, Hebiri and Lederer (2017); Guntuboyina et al. (2017); Lin et al. (2017).

## Appendix C: Proofs of Section 6

*Proof of Lemma 6.3.* The result follows directly by the proof of Lemma 5.3 (i.e. Theorem 6.1 in van de Geer (2018)), by the decomposition of the branched path graph into three path graphs. See Appendix B.1 for an intuition.

We consider here the case where the first and the second notation introduced in Section 3 coincide. The case where the two notations do not coincide differs from the case exposed here only in the choice of the edges around the branching points which are chosen to bound the last jump in $S_1$ and the first jumps in $S_2$ and $S_3$ by the mean of the signal values at some vertices surrounding these

candidate active edges. The case we expose here can be seen as an analogous to Corollary 6.6.

Let us define $b^1 = 0$, $e^1 = b^2 = b$, $e^2 = b^3 = n_1$, $e^3 = n$. In analogy with the proof of Lemma 5.3,

$$
\begin{aligned}
&\|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\| \\
&\leq \sum_{i=1}^{3} \Bigg\{ |f_{b^i+d_1^i+1} - f_{b^i+d_1^i}| - \sum_{k=2}^{b^i+d_1^i} |f_k - f_{k-1}| - \sum_{k=b^i+d_1^i+2}^{b^i+d_1^i+u_2} |f_k - f_{k-1}| \\
&+ |f_{b^i+d_1^i+d_2^i+1} - f_{b^i+d_1^i+d_2^i}| - \sum_{k=b^i+d_1^i+u_2+2}^{b^i+d_1^i+d_2^i} |f_k - f_{k-1}| \\
&- \sum_{k=b^i+d_1^i+d_2^i+2}^{b^i+d_1^i+d_2^i+u_3} |f_k - f_{k-1}| \\
&\dots \\
&+ |f_{b^i+d_1^i+\cdots+d_{s-1}^i+1} - f_{b^i+d_1^i+\cdots+d_{s-1}^i}| \\
&- \sum_{k=b^i+d_1^i+\cdots+d_{s-2}^i+u_{s-1}+2}^{b^i+d_1^i+\cdots+d_{s-1}^i} |f_k - f_{k-1}| - \sum_{k=b^i+d_1^i+\cdots+d_{s-1}^i+2}^{b^i+d_1^i+\cdots+d_{s-1}^i+u_s} |f_k - f_{k-1}| \\
&+ |f_{b^i+d_1^i+\cdots+d_s^i+1} - f_{b^i+d_1^i+\cdots+d_s^i}| \\
&- \sum_{k=b^i+d_1^i+\cdots+d_{s-1}^i+u_s+2}^{b^i+d_1^i+\cdots+d_s^i} |f_k - f_{k-1}| - \sum_{k=b^i+d_1^i+\cdots+d_s^i+2}^{e^i} |f_k - f_{k-1}| \Bigg\}
\end{aligned}
$$

We can thus infer Lemma 6.3 by applying exactly the same passages applied in the proof of Lemma 5.3. It is crucial to notice here that the term $-|f_{b+1} - f_b| - |f_{n_1+1} - f_b|$ is upper bounded by zero, i.e. simply discarded, since is is not used when we apply Lemma B.3. Indeed, in the case considered here, the edges $(b, b+1)$ and $(b, n_1+1)$ are cut to obtain three path graphs and thanks to our tools do not have to participate in the bound of the compatibility constant and can be discarded. The same reasoning can be applied in the case when other edges are cut to obtain a decomposition into three path graphs. The proof of these cases is essentially the same. It only requires the introduction of additional heavy notation. □

*Proof of Corollary 6.4.* The proof follows by direct calculations in analogy to the one of Corollary 5.4 (i.e. Theorem 6.1 in van de Geer (2018)). □

*Proof of Lemma 6.5.* In the proof of Lemma 5.3 and Lemma 5.5 (i.e. Theorem 6.1 and Lemma 9.1 in van de Geer (2018)) and in Appendix B.1 it is made clear, that the use of Lemma B.3 requires that the edges connecting the smaller pieces into which the path graph is partitioned are taken out of consideration when upper bounding $\|\beta_S\|_1 - \|\beta_{-(\{1\}\cup S)}\|_1$ resp. $\|(\beta \odot w)_S\|_1 - \|(\beta \odot w)_{-(\{1\}\cup S)}\|_1$.

This results in an upper bound containing only the square of some of the consecutive pairwise differences between the entries of $w$, the vector of weights. This "incomplete" sum can then of course be upper bounded by $\|Dw\|_2$, where $D$ is the incidence matrix of the path graph.

In the case of the branched path graph the same reasoning applies in particular to the two edges connecting together the three path graphs defined by the second notation. Indeed, these can be left out. Thus, in full analogy to the procedure exposed in the proofs of Lemma 5.3 and 5.5 (i.e. Theorem 6.1 and Lemma 9.1 in van de Geer (2018)) for the path graph, the statement of Lemma 6.5 follows. See Appendix B.1 for an intuition.

We expose here the idea for the case where the first and the second notation introduced in Section 3 coincide, as we did in the proof of Lemma 6.3.

Let us define $b^1 = 0$, $e^1 = b^2 = b$, $e^2 = b^3 = n_1$, $e^3 = n$. Let $g_i = f_i w_i$. In analogy to Lemma 5.5, we can apply the calculations performed for the nonweighted case to $g = w \odot f$, i.e.

$$
\begin{aligned}
&\|b_S \odot w_S\|_1 - \|b_{-(\{1\}\cup S)} \odot w_{-(\{1\}\cup S)}\|_1 \\
&\leq \sum_{i=1}^{3} \Bigg\{ \Bigg\{ |g_{b^i+d_1^i+1} - g_{b^i+d_1^i}| - \sum_{k=2}^{b^i+d_1^i} |g_k - g_{k-1}| - \sum_{k=b^i+d_1^i+2}^{b^i+d_1^i+u_2} |g_k - g_{k-1}| \\
&+ |g_{b^i+d_1^i+d_2^i+1} - g_{b^i+d_1^i+d_2^i}| - \sum_{k=b^i+d_1^i+u_2+2}^{b^i+d_1^i+d_2^i} |g_k - g_{k-1}| \\
&- \sum_{k=b^i+d_1^i+d_2^i+2}^{b^i+d_1^i+d_2^i+u_3} |g_k - g_{k-1}| \\
&\dots \\
&+ |g_{b^i+d_1^i+\dots+d_{s-1}^i+1} - g_{b^i+d_1^i+\dots+d_{s-1}^i}| \\
&- \sum_{k=b^i+d_1^i+\dots+d_{s-2}^i+u_{s-1}+2}^{b^i+d_1^i+\dots+d_{s-1}^i} |g_k - g_{k-1}| - \sum_{k=b^i+d_1^i+\dots+d_{s-1}^i+2}^{b^i+d_1^i+\dots+d_{s-1}^i+u_s} |g_k - g_{k-1}| \\
&+ |g_{b^i+d_1^i+\dots+d_s^i+1} - g_{b^i+d_1^i+\dots+d_s^i}| \\
&- \sum_{k=b^i+d_1^i+\dots+d_{s-1}^i+u_s+2}^{b^i+d_1^i+\dots+d_s^i} |g_k - g_{k-1}| - \sum_{k=b^i+d_1^i+\dots+d_s^i+2}^{e^i} |g_k - g_{k-1}| \Bigg\} \Bigg\} \text{ I} \\
&+ \sum_{i=2}^{b} |w_i - w_{i-1}||f_{i-1}| + \sum_{i=b+2}^{n_1} |w_i - w_{i-1}||f_{i-1}| \\
&+ \sum_{i=n_1+2}^{n} |w_i - w_{i-1}||f_{i-1}|. \Bigg\} \text{ II}
\end{aligned}
$$

For the first term we have, in analogy with Lemma 5.5 and the procedure illustrated in Lemma 6.3

$$\text{I} \leq \sqrt{K^b}\|w\|_\infty \|f\|_2.$$

Let $D^*$ denote the incidence matrix of the branched path graph, where the entries in the rows corresponding to the edges cut to obtain the three path graphs according to the second notation exposed in Section 3 have been substituted with zeroes. For the second term we have, by the Cauchy-Schwarz inequality

$$\begin{aligned} \text{II} &\leq \left(\sum_{i=1}^{b-1} f_i^2 + \sum_{i=b+1}^{n_1-1} f_i^2 + \sum_{i=n_1+1}^{n-1} f_i^2\right)^{1/2} \|D^* w\|_2 \\ &\leq \|f\|_2 \|D^* w\|_2 \leq \|f\|_2 \|Dw\|_2. \end{aligned}$$

Combining the inequalities for I and for II we can infer Lemma 6.5. □

*Proof of Corollary 6.6.* We use the calculations done in Section 4. By writing

$$(a_j^i)_k = \sqrt{\frac{(d_j^i - k)k}{d_j^i}}, i \in \{0, 1, \ldots, d_j^i\},$$

where $j \in [s_1]$ for $i = 1$ and $j \in [s_i + 1] \setminus \{1\}$ for $i \in \{2, 3\}$, and

$$a_k^* = \sqrt{\frac{(d^* - i)i}{d^*}}, i \in \{0, 1, \ldots, d^*\} \text{ where } d^* = \tilde{d}_{s_1+1}^1 + \tilde{d}_1^2 + \tilde{d}_1^3$$

we obtain that

$$\begin{aligned} \|Dw\|_2^2 &= \frac{1}{\gamma^2 n} \left\{ \sum_{j=1}^{s_1} \sum_{k=1}^{d_j^1} ((a_j^i)_k - (a_j^i)_{k-1})^2 + \sum_{j=2}^{s_2+1} \sum_{k=1}^{d_j^2} ((a_j^i)_k - (a_j^i)_{k-1})^2 \right. \\ &+ \sum_{j=2}^{s_3+1} \sum_{k=1}^{d_j^3} ((a_j^i)_k - (a_j^i)_{k-1})^2 \\ &+ \sum_{k=1}^{\tilde{d}_{s_1+1}^1 - 1} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{\tilde{d}_1^2} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{\tilde{d}_1^3} (a_k^* - a_{k-1}^*)^2 \\ &+ \left. (a_{\tilde{d}_{s_1+1}^1 - 1}^* - a_{\tilde{d}_1^2}^*)^2 + (a_{\tilde{d}_{s_1+1}^1 - 1}^* - a_{\tilde{d}_1^3}^*)^2 \right\}. \end{aligned}$$

Indeed we can bound all the terms except the last two ones by applying the reasoning developed for the path graph.

We have that

$$\begin{aligned} \|Dw\|_2^2 &= \frac{1}{\gamma^2 n} \left\{ \sum_{j=1}^{s_1} \sum_{k=1}^{d_j^1} ((a_j^i)_k - (a_j^i)_{k-1})^2 + \sum_{j=2}^{s_2+1} \sum_{k=1}^{d_j^2} ((a_j^i)_k - (a_j^i)_{k-1})^2 \right. \\ &+ \left. \sum_{j=2}^{s_3+1} \sum_{k=1}^{d_j^3} ((a_j^i)_k - (a_j^i)_{k-1})^2 + z \right\}, \end{aligned}$$

where

$$\begin{aligned} z &= \sum_{k=1}^{\tilde{d}^1_{s_1+1}-1}(a^*_k - a^*_{k-1})^2 + \sum_{k=1}^{\tilde{d}^2_1}(a^*_k - a^*_{k-1})^2 + \sum_{k=1}^{\tilde{d}^3_1}(a^*_k - a^*_{k-1})^2 \\ &+ (a^*_{\tilde{d}^1_{s_1+1}-1} - a^*_{\tilde{d}^2_1})^2 + (a^*_{\tilde{d}^1_{s_1+1}-1} - a^*_{\tilde{d}^3_1})^2 \end{aligned}$$

We are now interested in upper bounding $\|D^*w\|_2^2$ rather than $\|Dw\|_2^2$. The form of $D^*$ depends of course on which edges are cut out to obtain three path graphs satisfying Assumption 6.2.

In the case we consider in this corollary we have that

$$\begin{aligned} \|D^*w\|_2^2 &= \frac{1}{\gamma^2 n}\left\{\sum_{j=1}^{s_1}\sum_{k=1}^{d^1_j}((a^i_j)_k - (a^i_j)_{k-1})^2 + \sum_{j=2}^{s_2+1}\sum_{k=1}^{d^2_j}((a^i_j)_k - (a^i_j)_{k-1})^2 \right. \\ &+ \left. \sum_{j=2}^{s_3+1}\sum_{k=1}^{d^3_j}((a^i_j)_k - (a^i_j)_{k-1})^2 + z\right\}, \end{aligned}$$

where

$$\begin{aligned} z &= \sum_{k=1}^{\tilde{d}^1_{s_1+1}-1}(a^*_k - a^*_{k-1})^2 + \sum_{k=1}^{\tilde{d}^2_1}(a^*_k - a^*_{k-1})^2 + \sum_{k=1}^{\tilde{d}^3_1}(a^*_k - a^*_{k-1})^2 \\ &\leq 5/2\log(\lfloor d^*/3\rfloor \lceil d^*/3\rceil (d^* - \lfloor d^*/3\rfloor - \lceil d^*/3\rceil)). \end{aligned}$$

Now define the vectors

$$|\Delta|^i := \begin{cases} \left(\lfloor d^i_1/2\rfloor, \lceil d^i_1/2\rceil, \ldots, \lfloor d^i_{s_i}/2\rfloor, \lceil d^i_{s_i}/2\rceil, \delta^i\right), i = 1 \\ \left(\delta^i, \lfloor d^i_2/2\rfloor, \lceil d^i_2/2\rceil, \ldots, \lfloor d^i_{s_i+1}/2\rfloor, \lceil d^i_{s_i+1}/2\rceil\right), i = 2,3 \end{cases}, \in \mathbb{R}^{2s_i+1},$$

where $(\delta^1, \delta^2, \delta^3) = (\lfloor d^*/3\rfloor, \lceil d^*/3\rceil, d^* - \lfloor d^*/3\rfloor - \lceil d^*/3\rceil)$ in any order.

Let $|\Delta| := (|\Delta|^1, |\Delta|^2, |\Delta|^3) \in \mathbb{R}^{2s+3}$.

In analogy to the case of the path graph, see Proof of Corollary 5.6 in Appendix B, we can find the bound

$$\begin{aligned} \|D^*w\|_2^2 &\leq (5/2)\log\left(\prod_{i=1}^{2s+3}|\Delta|_i\right) + \\ &\leq (5/2)(2s+3)\log\left(\frac{n+1}{2s+3}\right) \end{aligned}$$

For the compatibility constant we have that $K_b \leq \frac{2s}{\bar{\Delta}_h}$ and we obtain an upper bound for the reciprocal of the weighted compatibility constant

$$\frac{1}{\kappa^2_w(S)} \leq \frac{2n}{\gamma^2(s+1)}\left(\frac{2\gamma^2 s}{\bar{\Delta}_h} + \frac{5(2s+3)\log(n+1)}{2n}\right).$$

We therefore get Corollary 6.6. □

*Proof of Corollary 6.7.* We recycle the initial considerations of the proof of Corollary 6.6. The proof of the three cases we consider deviates from the one of Corollary 6.6 by the way $z$ is bounded, i.e. which differences of consecutive weights can be left out.

We can distinguish three cases:

1) Assume without loss of generality that $\tilde{d}_1^3 = 0$.

$$\begin{aligned} z &= \sum_{k=1}^{\tilde{d}^1_{s_1+1}-1} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{\tilde{d}_1^2} (a_k^* - a_{k-1}^*)^2 + (a^*_{\tilde{d}^1_{s_1+1}-1})^2 \\ &\leq \sum_{k=1}^{d^*} (a_k^* - a_{k-1}^*)^2 + \max_{k\in[d^*]} (a_k^*)^2 \\ &\leq 5/2 \log(\lfloor d^*/2 \rfloor \lceil d^*/2 \rceil) + d^*/2 \end{aligned}$$

2) a) Assume without loss of generality that $\tilde{d}_1^3 = 2$.

$$\begin{aligned} z &\leq \sum_{k=1}^{\tilde{d}_1^2} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{\tilde{d}_1^3} (a_k^* - a_{k-1}^*)^2 + (a^*_{d^*-3})^2 \\ &\leq 5/2 \log(\lfloor d^*/2 \rfloor \lceil d^*/2 \rceil) + 3 \end{aligned}$$

b) We have the choice, which edge we can leave out of our consideration: either the edge $(b, b+1)$ or the edge $(b, n_1+1)$. In both cases

$$z \leq \sum_{k=1}^{\tilde{d}_1^2} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{\tilde{d}_1^3} (a_k^* - a_{k-1}^*)^2 + [(a^*_{\tilde{d}_1^2})^2 \wedge (a^*_{\tilde{d}_1^3})^2].$$

Denote $y := \tilde{d}_1^2$. Then $\tilde{d}_1^3 = d^* - 1 - y$. We get that

$$(a_y^*)^2 \wedge (a^*_{b-y-1})^2 = \begin{cases} (a_y^*)^2 & , 3 \leq y \leq (d^*-1)/2 \\ (a^*_{b-y-1})^2 & , (d^*-1)/2 \leq w \leq d^* - 3, \end{cases} \leq d^*/4.$$

Thus

$$z \leq 5/2 \log(\lfloor d^*/2 \rfloor \lceil d^*/2 \rceil) + d^*/4$$

3) Assume without loss of generality $\tilde{d}_1^3 = 1$, then

$$\begin{aligned} z &\leq \sum_{k=1}^{\tilde{d}^1_{s_1+1}-1} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{\tilde{d}_1^2} (a_k^* - a_{k-1}^*)^2 + \sum_{k=1}^{1} (a_k^* - a_{k-1}^*)^2 \\ &+ (a_1^* - a^*_{\tilde{d}^1_{s_1+1}-1})^2. \end{aligned}$$

Let $x := \tilde{d}^1_{s_1+1}$. We have that

$$\max_{3\leq x\leq d^*-3} \left( \sqrt{\frac{d^*-1}{d^*}} - \sqrt{\frac{(d^*-x+1)(x-1)}{d^*}} \right)^2$$

$$= \frac{1}{d^*}(d^*/2 - \sqrt{d^* - 1})^2 \leq d^*/4,$$

where the maximum is attained at $x = \frac{d^*+2}{2}$ and the last inequality holds since $d^*/2 \geq \sqrt{d^* - 1}, \forall d^* \geq 1$. Therefore

$$z \leq 5/2 \log(\lfloor d^*/3 \rfloor \lceil d^*/3 \rceil (d^* - \lfloor d^*/3 \rfloor - \lceil d^*/3 \rceil)) + d^*/4$$

Now define the vectors

$$|\Delta|^i := \begin{cases} \left(\lfloor d_1^i/2 \rfloor, \lceil d_1^i/2 \rceil, \ldots, \lfloor d_{s_i}^i/2 \rfloor, \lceil d_{s_i}^i/2 \rceil, \delta^i\right), i = 1 \\ \left(\delta^i, \lfloor d_2^i/2 \rfloor, \lceil d_2^i/2 \rceil, \ldots, \lfloor d_{s_i+1}^i/2 \rfloor, \lceil d_{s_i+1}^i/2 \rceil\right), i = 2, 3 \end{cases}, \in \mathbb{R}^{2s_i+1}.$$

We can distinguish the following four cases:

1) $\delta^2 = 1$ or $\delta^3 = 1$ and the nonzero $\delta$'s take values $\lfloor d^*/2 \rfloor$ and $\lceil d^*/2 \rceil$;
2) See Case 1), however with $\delta^1 = 1$;
3) $(\delta^1, \delta^2, \delta^3) = (\lfloor d^*/3 \rfloor, \lceil d^*/3 \rceil, d^* - \lfloor d^*/3 \rfloor - \lceil d^*/3 \rceil)$ in any order.

In these four cases we replace by ones potential zeroes, since this still allows us to obtain an upper bound.

Let $|\Delta| := (|\Delta|^1, |\Delta|^2, |\Delta|^3) \in \mathbb{R}^{2s+3}$.

In analogy to the case of the path graph, see Proof of Corollary 5.6 in Appendix B, we can find the bound

$$\begin{aligned} \|D^* w\|_2^2 &\leq (5/2) \log \left( \prod_{i=1}^{2s+3} |\Delta|_i \right) + \zeta \\ &\leq (5/2)(2s+3) \log \left( \frac{n+1}{2s+3} \right) + \zeta, \end{aligned}$$

where

$$\zeta = \begin{cases} d^*/2 & , \text{ Case 1)} \\ 3 & , \text{ Case 2)a)} \\ d^*/4 & , \text{ Case 2)b)} \\ d^*/4 & , \text{ Case 3)} \end{cases}.$$

For the compatibility constant we have that $K_b \leq \frac{2s}{\bar{\Delta}_h}$ and we obtain an upper bound for the reciprocal of the weighted compatibility constant

$$\frac{1}{\kappa_w^2(S)} \leq \frac{2n}{\gamma^2(s+1)} \left( \frac{2\gamma^2 s}{\bar{\Delta}_h} + \frac{5(2s+3)\log(n+1)}{2n} + \frac{\zeta}{n} \right),$$

where $\zeta$ is as above. We therefore get Corollary 6.7. □

## Appendix D: Proofs of Section 8

### *D.1. Preliminaries*

We will need the following results.

**Lemma D.1** (The inverse of a partitioned matrix)**.** *Let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

*where $A_{11}$ and $A_{22}$ are invertible matrices and $A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $A_{22} - A_{21}A_{11}^{-1}A_{12}$ are invertible as well. Then*

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}$$

**Lemma D.2** (The inverse of the sum of two matrices, Miller (1981))**.** *Let $G$ and $G + E$ be invertible matrices, where $E$ is a matrix of rank one. Let $g := trace(EG^{-1})$.*

*Then $g \neq -1$ and*

$$(G + E)^{-1} = G^{-1} - \frac{1}{1+g}G^{-1}EG^{-1}.$$

**Inverse of symmetric matrices** It is known that the inverse of a symmetric matrix is symmetric as well. This fact has relevance in Lemma D.1, where

$$(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1},$$

if $A$ is symmetric.

### *D.2. Proofs*

*Proof of Lemma 8.5.* Then $U = \{1\}$ and $X$ is the path matrix with reference vertex 1 of the graph. It follows that $X_1 = 1_n$, $X_1'X_1 = n$ and $\Pi_1 = \frac{\mathbb{I}_n}{n}$, where $\mathbb{I}_n \in \mathbb{R}^{n\times n}$ is a matrix only consisting of ones.

We want to show that the last $s$ conlumns of

$$X_R'X_{\{1\}\cup S_0}(X_{\{1\}\cup S_0}'X_{\{1\}\cup S_0})^{-1}$$

are the same as

$$X_R'A_1X_{S_0}(X_{S_0}'A_1X_{S_0})^{-1},$$

i.e. that the last $s$ columns of

$$X_{\{1\}\cup S_0}(X_{\{1\}\cup S_0}'X_{\{1\}\cup S_0})^{-1}$$

are the same as

$$A_1X_{S_0}(X_{S_0}'A_1X_{S_0})^{-1}.$$

We start by writing

$$X_{\{1\}\cup S_0}'X_{\{1\}\cup S_0} = n\begin{pmatrix} 1 & \mu' \\ \mu & \hat{\Sigma}_{S_0S_0} \end{pmatrix},$$

where $\mu$ (resp. $\mu'$) is the first column (resp. row) of $\hat{\Sigma}_{S_0S_0}$. Note that

$$X'_{S_0}A_1X_{S_0} = n(\hat{\Sigma}_{S_0S_0} - \mu\mu').$$

By using the formula for the inverse of a partitioned matrix (see Lemma D.1) we get that

$$(X'_{\{1\}\cup S_0}X_{\{1\}\cup S_0})^{-1} = \begin{pmatrix} \frac{1}{n(1-\mu_1)} & \frac{-1}{n(1-\mu_1)}e_1' \\ \frac{-1}{n(1-\mu_1)}e_1 & (X'_{S_0}A_1X_{S_0})^{-1} \end{pmatrix},$$

where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^s$. As a consequence we can perform the following multiplication:

$$X_{\{1\}\cup S_0}(X'_{\{1\}\cup S_0}X_{\{1\}\cup S_0})^{-1} =$$
$$\left(\tfrac{1}{n(1-\mu_1)}(X_1 - X_{S_0}e_1) \quad X_{S_0}(X'_{S_0}A_1X_{S_0})^{-1} - \tfrac{1}{n(1-\mu_1)}X_1e_1'\right).$$

We now develop $A_1X_{S_0}(X'_{S_0}A_1X_{S_0})^{-1}$ to see if it coincides with the second entry of the matrix we have obtained. In particular

$$\begin{aligned} A_1X_{S_0}(X'_{S_0}A_1X_{S_0})^{-1} &= (\mathrm{I}_n - \Pi_1)X_{S_0}(X'_{S_0}A_1X_{S_0})^{-1} \\ &= X_{S_0}(X'_{S_0}A_1X_{S_0})^{-1} - \frac{X_1\mu'}{n}(\hat{\Sigma} - \mu\mu')^{-1}. \end{aligned}$$

By using Lemma D.2 we can write the second term as

$$\begin{aligned} -\frac{X_1\mu'}{n}(\hat{\Sigma}_{S_0S_0} - \mu\mu')^{-1} &= -\frac{X_1\mu'}{n}(\hat{\Sigma}^{-1}_{S_0S_0} + \frac{1}{1-\mu_1})\begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix} \\ &= \frac{-X_1e_1'}{n}\left(1 + \frac{\mu_1}{1-\mu_1}\right) = \frac{-1}{n(1-\mu_1)}X_1e_1'. \end{aligned}$$

In the KKT conditions we note that $z_1^0 = 0$ (indeed we have the usual normal equations for coefficients not penalized) and thus we establish the desired equality. □

*Proof of Theorem 8.8.* We refer to Section 4 for the calculation of the projection coefficients.

Let us define

$$\alpha(i) := \frac{i}{d^*}, i = \{1, \ldots, d^* - 1\}.$$

We now select an $i$ and write $\alpha = \alpha(i)$. We get that the irrepresentable condition is satisfied for a signal pattern $z \in \{-1, 1\}^{K+1}$ if $\forall i \leq \min\{\tilde{d}^1_{s_1+1} - 1, \tilde{d}^2_1, \ldots, \tilde{d}^{K+1}_1\}$

1. $|(1-\alpha, \alpha, \ldots, \alpha)z| < 1$ and
2. $|(\alpha, 1-\alpha, -\alpha, \ldots, -\alpha)z| < 1$ as well as this has to hold for any of the $K$ possible permutations of the last $K$ elements of the vector $(\alpha, 1-\alpha, -\alpha, \ldots, -\alpha)' \in \mathbb{R}^{K+1}$.

We now want to find the signal patterns $z$ for which the irrepresentable condition is satisfied.

Consider the first condition: it excludes the signal pattern where all the jumps have the same sign.

Thus, in the following assume w.l.o.g. that $z_1 = 1$. Now we look at the second condition. We are going to consider the cases where $p$ of the $K$ last elements of the vector $(\alpha, 1-\alpha, -\alpha, \ldots, -\alpha)$ get the sign $+$ and $K-p$ get the sign $-$. We look for the linear combination with the highest absolute value. This can be seen as finding the linear combination $L$ of $(\alpha, -\alpha, \ldots, -\alpha)$ determined by $p$ and then adding $\operatorname{sgn}(L)$ to it. We scan the cases $p = 1, \ldots, K-1$, since the case $p = K$ is already discarded by looking at the first condition.

For $p = 1, \ldots, \lfloor (K+1)/2 \rfloor$, we have that $K+1-2p > 0$, thus we assign a $+$ sign to $(1-\alpha)$ and get $1 + (K+1-2p)\alpha > 1$ and the irrepresentable condition is violated.

For $p = \lceil (K+1)/2 \rceil, \ldots, K-1$, we have that $K+1-2p < 0$, thus we assign a $-$ sign to $(1-\alpha)$ and get $-1 + (K+1-2p)\alpha < -1$ and the irrepresentable condition is violated.

If $K$ is odd, for $p = (K+1)/2$, we have that $K+1-2p = 0$ and the irrepresentable condition is violated, since the linear combination gives $\pm 1$.

Thus, it only remains to consider $p = 0$. For $p = 0$ we get the condition $|1-(K+1)\alpha| < 1$ from the first as well as from the second condition above. This condition is satisfied whenever $\alpha < 2/(K+1)$, i.e.

$$i < \frac{2}{K+1} d^*$$

This means that if any of $\tilde{d}^1_{s_1+1} - 1, \tilde{d}^2_1, \ldots, \tilde{d}^{K+1}_1$ exceeds $\frac{2d^*}{K+1}$, then the irrepresentable condition is not satisfied. □

## References


Bapat, R. B. (2014). *Graphs and Matrices*. Springer, London. MR3289036

Dalalyan, A. S., Hebiri, M. and Lederer, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. MR3556784

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics* **26** 879–921. MR1635414

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1** 302–332. MR2415737

Guntuboyina, A., Lieu, D., Chatterjee, S. and Sen, B. (2017). Adaptive Risk Bounds in Univariate Total Variation Denoising and Trend Filtering. *ArXiv ID 1702.05113*.

Hütter, J.-C. and Rigollet, P. (2016). Optimal rates for total variation denoising. *JMLR: Workshop and Conference Proceedings* **49** 1–32.

Jacobs, D. P., Machado, C. M. S., Pereira, E. C. and Trevisan, V. (2008). Computing the inverse of a tree's incidence matrix. *Congr. Numerantium* **189** 169–176. MR2489782

Jia, J. and Rohe, K. (2015). Preconditioning the Lasso for sign consistency. *Electronic Journal of Statistics* **9** 1150–1172. MR3354334

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics* **28** 1302–1338. MR1805785

Lin, K., Sharpnack, J., Rinaldo, A. and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. *Neural Information Processing Systems (NIPS)* **3** 42.

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

Miller, K. S. (1981). On the inverse of the sum of matrices. *Mathematics Magazine* **54** 67–72. MR0617892

Ottersten, J., Wahlberg, B. and Rojas, C. R. (2016). Accurate changing point detection for l1 mean filtering. *IEEE Signal Processing Letters* **23** 297–301.

Owrang, A., Malek-Mohammadi, M., Proutiere, A. and Jansson, M. (2017). Consistent change point detection for piecewise constant signals with normalized fused LASSO. *IEEE Signal Processing Letters* **24** 799–803.

Qian, J. and Jia, J. (2016). On stepwise pattern recovery of the fused Lasso. *Computational Statistics and Data Analysis* **94** 221–237. MR3412821

Rojas, C. R. and Wahlberg, B. (2014). On change point detection using the fused lasso method. *ArXiv ID 1401.5408v1.*

Rojas, C. R. and Wahlberg, B. (2015). How to monitor and mitigate staircasing in l1 trend filtering. *ICASSP* 3946–3950.

Sadhanala, V., Wang, Y.-X. and Tibshirani, R. J. (2016). Total variation classes beyond 1d: minimax rates, and the limitations of linear smoothers. *NIPS.*

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized LASSO. *The Annals of Statistics* **39** 1335–1371. MR2850205

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

van de Geer, S. (2016). *Estimation and testing under sparsity* **2159**. Springer. MR3526202

van de Geer, S. (2018). On tight bounds for the Lasso. *ArXiv ID 1804.00989.*

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7** 2541–2563. MR2274449