Fused Lasso Nearly-Isotonic Signal Approximation in General Dimensions

Vladimir Pastukhov

vmpastukhov@yahoo.com

Research Article

Keywords: Constrained inference, isotonic regression, nearly-isotonic regression, fused lasso

Posted Date: October 24th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3467665/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Statistics and Computing on May 22nd, 2024. See the published version at https://doi.org/10.1007/s11222-024-10432-6.

Fused Lasso Nearly-Isotonic Signal Approximation in General Dimensions

Vladimir Pastukhov[1*]

[1*]Department of Computer Science and Engineering, Chalmers University of Technology, Rannvagen 6, Chalmers University of Technology, Gothenburg, 41258, Sweden.

Corresponding author(s). E-mail(s): vlapas@chalmers.se;

Abstract

In this paper we introduce and study fused lasso nearly-isotonic signal approximation, which is a combination of fused lasso and generalized nearly-isotonic regression. We show how these three estimators relate to each other and derive solution to a general problem. Our estimator is computationally feasible and provides a trade-off between monotonicity, block sparsity and goodness-of-fit. Next, we prove that fusion and near-isotonisation in one dimensional case can be applied interchangably, and this step-wise procedure gives the solution to the original optimization problem. This property of the estimator is very important, because it provides a direct way to construct path solution when one of the penalization parameters is fixed. Also, we derive unbiased estimator of degrees of freedom of the estimator.

Keywords: Constrained inference, isotonic regression, nearly-isotonic regression, fused lasso

1 Introduction

This work is motivated by recent papers in nearly-constrained estimation in several dimensions and by the papers in generalised penalized least squared regression. The subject of penalized estimators starts with L_1-penalisation, cf. [1], which is called lasso signal approximation, and L_2-penalisation, which is usually addressed as ridge regression [2] or sometimes as Tikhonov-Philips regularization [3, 4]. The first generalisation of lasso is L_1-penalisation imposed on the successive differences of the coefficients. For a given sequence of data points $\boldsymbol{y} \in \mathbb{R}^n$ the fusion approximator (cf. [5]) is given by

$$\hat{\boldsymbol{\beta}}^F(\boldsymbol{y}, \lambda_F) = \underset{\boldsymbol{\beta} \in \mathsf{R}^n}{\arg\min} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_F \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|. \tag{1}$$

The combination of fusion approximator and lasso is called fused lasso estimator and is given by:

$$\hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y}, \lambda_F, \lambda_L) = \underset{\boldsymbol{\beta} \in \mathsf{R}^n}{\arg\min} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_F \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}| + \lambda_L||\boldsymbol{\beta}||_1. \tag{2}$$

The fused lasso was introduced in [6] and its asymptotic properties were studied in detail in [5]. Also, it is worth to note that in the paper [7] the estimator in (1) is called the fused lasso, while the estimator in (2) is addressed as the sparse fused lasso.

In the area of constrained inference the basic and simplest problem is isotonic regression in one dimension. For a given sequence of data points $\boldsymbol{y} \in \mathbb{R}^n$ isotonic regression is the following approximation

$$\hat{\boldsymbol{\beta}}^I = \underset{\boldsymbol{\beta} \in \mathsf{R}^n}{\arg\min}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2, \quad \text{subject to} \quad \beta_1 \leq \beta_2 \leq \cdots \leq \beta_n, \tag{3}$$

i.e. it is ℓ^2-projection of the vector $\boldsymbol{y}$ onto the set of non-increasing vectors in $\mathbb{R}^n$. The notion of isotonic "regression" in this context might be confusing. Nevertheless, it is a standard notion in this subject, cf., for example, the papers [8, 9], where the notation "isotonic regression" is used for the isotonic projection of a general vector. Also, in this paper we use notations "regression", "estimator" and "approximator" interchangeably. A general introduction to isotonic regression can be found, for example, in [10].

The nearly-isotonic regression, introduced in [11] and studied in detail in [12], is a less restrictive version of isotonic regression and is given by the following optimization problem

$$\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) = \underset{\boldsymbol{\beta} \in \mathsf{R}^n}{\arg\min} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_{NI} \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|_+, \tag{4}$$

where $x_+ = x \cdot 1\{x > 0\}$.

In this paper we combine fused lasso estimator with nearly-isotonic regression and call the resulting estimator as *fused lasso nearly-isotonic signal approximator*, i.e. for a given sequence of data points $\boldsymbol{y} \in \mathbb{R}^n$ the problem in one dimensional case is the following optimization:

$$\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \\ \underset{\boldsymbol{\beta} \in \mathsf{R}^n}{\arg\min} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_F \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}| + \lambda_L||\boldsymbol{\beta}||_1 + \lambda_{NI} \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|_+. \tag{5}$$

Also, in the case of $\lambda_F \neq 0$ and $\lambda_{NI} \neq 0$, with $\lambda_L = 0$, we call the estimator as *fused nearly-isotonic regression*, i.e.

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) \equiv \hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) = \\
&\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\arg\min}\, \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_F \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}| + \lambda_{NI} \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|_+.
\end{aligned} \tag{6}$$

This generalisation of nearly-isotonic regression in (6) was proposed in the conclusion of the paper [11]. Next, one-dimensional fused nearly-isotonic regression was considered and numerically solved in [13] with time complexity $\mathcal{O}(n)$. Nevertheless, first, in this paper we consider and solve the problem in general dimensions. Second, for fixed penalisation parameters in one-dimensional case we also provide solution with linear complexity and exact partly path solution (when one of the parameters is fixed and the path is with respect the other one) with complexity $\mathcal{O}(n \log(n))$.

It is also worth to mention the paper [14], where the authors studied nearly-isotonic approximator with extra penalisation term

$$(\beta_i - \beta_{i+1})^2 \cdot 1\{(\beta_i - \beta_{i+1}) > 0\}$$

with additial lasso penalty. Also, in the paper [15] the authors did a comparison of the algorithms to solve lasso with linear constraints, which is called constrained lasso.

In the next step we state the problem defined in (5) for the general case of isotonic constraints with respect to a general partial order. First, we have to introduce the notation.

1.1 Notation

We start with basic definitions of partial order and isotonic regression. Let $\mathcal{I} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n\}$ be some index set. Next, we define the following binary relation $\preceq$ on $\mathcal{I}$.

A binary relation $\preceq$ on $\mathcal{I}$ is called partial order if

- it is reflexive, i.e. $\boldsymbol{j} \preceq \boldsymbol{j}$ for all $\boldsymbol{j} \in \mathcal{I}$;
- it is transitive, i.e. $\boldsymbol{j}_1, \boldsymbol{j}_2, \boldsymbol{j}_3 \in \mathcal{I}$, $\boldsymbol{j}_1 \preceq \boldsymbol{j}_2$ and $\boldsymbol{j}_2 \preceq \boldsymbol{j}_3$ imply $\boldsymbol{j}_1 \preceq \boldsymbol{j}_3$;
- it is antisymmetric, i.e. $\boldsymbol{j}_1, \boldsymbol{j}_2 \in \mathcal{I}$, $\boldsymbol{j}_1 \preceq \boldsymbol{j}_2$ and $\boldsymbol{j}_2 \preceq \boldsymbol{j}_1$ imply $\boldsymbol{j}_1 = \boldsymbol{j}_2$.

Further, a vector $\boldsymbol{\beta} \in \mathbb{R}^n$ indexed by $\mathcal{I}$ is called isotonic with respect to the partial order $\preceq$ on $\mathcal{I}$ if $\boldsymbol{j}_1 \preceq \boldsymbol{j}_2$ implies $\beta_{\boldsymbol{j}_1} \leq \beta_{\boldsymbol{j}_2}$. We denote the set of all isotonic vectors in $\mathbb{R}^n$ with respect to the partial order $\preceq$ on $\mathcal{I}$ by $\boldsymbol{\mathcal{B}}^{is}$, which is closed convex cone in $\mathbb{R}^n$ and it is also called isotonic cone. Next, a vector $\boldsymbol{\beta}^I \in \mathbb{R}^n$ is isotonic regression of an arbitrary vector $\boldsymbol{y} \in \mathbb{R}^n$ over the pre-ordered index set $\mathcal{I}$ if

$$\boldsymbol{\beta}^I = \underset{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}^{is}}{\arg\min} \sum_{\boldsymbol{j}\in\mathcal{I}} (\beta_{\boldsymbol{j}} - y_{\boldsymbol{j}})^2. \tag{7}$$

For any partial order relation $\preceq$ on $\mathcal{I}$ there exists directed graph $G = (V, E)$, with $V = \mathcal{I}$ and E is the minimal set of edges such that

$$E = \{(\boldsymbol{j}_1, \boldsymbol{j}_2), \text{ where } (\boldsymbol{j_1}, \boldsymbol{j_2}) \text{ is the ordered pair of vertices from } \mathcal{I}\}, \tag{8}$$

such that an arbitrary vector $\boldsymbol{\beta} \in \mathbb{R}^n$ is isotonic with respect to $\preceq$ iff $\beta_{\boldsymbol{l_1}} \leq \beta_{\boldsymbol{l_2}}$, given that E contains the chain of edges from $\boldsymbol{l}_1 \in V$ to $\boldsymbol{l}_2 \in V$.

Now we can generalise the estimators discussed above. First, equivalently to the definition in (7), a vector $\boldsymbol{\beta}^I \in \mathbb{R}^n$ is isotonic regression of an arbitrary vector $\boldsymbol{y} \in \mathbb{R}^n$ indexed by the partially ordered index set $\mathcal{I}$ if

$$\boldsymbol{\beta}^I = \arg\min_{\boldsymbol{\beta}} \sum_{\boldsymbol{j} \in \mathcal{I}} (\beta_{\boldsymbol{j}} - y_{\boldsymbol{j}})^2, \tag{9}$$

subject to $\beta_{\boldsymbol{l_1}} \leq \beta_{\boldsymbol{l_2}}$, whenever E contains the chain of edges from $\boldsymbol{l}_1 \in V$ to $\boldsymbol{l}_2 \in V$.

Second, for the directed graph $G = (V, E)$, which corresponds to the partial order $\preceq$ on $\mathcal{I}$, the nearly-isotonic regression of $\boldsymbol{y} \in \mathbb{R}^n$ indexed by $\mathcal{I}$ is given by

$$\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_{NI} \sum_{(\boldsymbol{i}, \boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}|_+. \tag{10}$$

This generalisation of nearly-isotonic regression was introduced and studied in [12].

Next, fused and fused lasso approximators for a general directed graph $G = (V, E)$ are given by

$$\hat{\boldsymbol{\beta}}^{F}(\boldsymbol{y}, \lambda_{F}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_{F} \sum_{(\boldsymbol{i}, \boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}|, \tag{11}$$

and

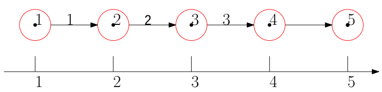
$$\hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y}, \lambda_{F}, \lambda_L) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_{F} \sum_{(\boldsymbol{i}, \boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}| + \lambda_L ||\boldsymbol{\beta}||_1. \tag{12}$$

These optimization problems were introduced and solved for a general graph in [7, 16, 17].

Further, let D denote the oriented incidence matrix for the directed graph $G = (V, E)$, corresponding to $\preceq$ on $\mathcal{I}$. We choose the orientation of D in the following way. Assume that the graph G with n vertexes has m edges. Next, assume we label the vertexes by $\{1, \dots, n\}$ and edges by $\{1, \dots, m\}$. Then D is $m \times n$ matrix with

$$D_{i,j} = \begin{cases} 1, & \text{if vertex } j \text{ is the source of edge } i, \\ -1, & \text{if vertex } j \text{ is the target of edge } i, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

In order to clarify the notations we consider the following examples of partial order relation. First, let us consider the monotonic order relation in one dimensional case. Let $\mathcal{I} = \{1, \dots, n\}$, and for $j_1 \in \mathcal{I}$ and $j_2 \in \mathcal{I}$ we naturally define $j_1 \preceq j_2$ if $j_1 \leq j_2$. Further, if we let $V = \mathcal{I}$ and $E = \{(i, i+1) : i = 1, \dots, n-1\}$, then $G = (V, E)$ is the directed graph which correspond to the one dimensional order relation on $\mathcal{I}$. Figure 1 displays the graph and the incidence matrix for the graph.



(a) Graph $G = (V, E)$

$$D = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

(b) Oriented incidence matrix D

Fig. 1: Graph for monotonic contstraints and oriented incidence matrix

Next, we consider two dimensional case with bimonotonic constraints. The notion of bimonotonicity was first introduced in [18] and it means the following. Let us consider the index set

$$\mathcal{I} = \{\boldsymbol{i} = (i^{(1)}, i^{(2)}) : i^{(1)} = 1, 2, \dots, n_1,\ i^{(2)} = 1, 2, \dots, n_2\}$$

with the following order relation $\preceq$ on it: for $\boldsymbol{j}_1, \boldsymbol{j}_2 \in \mathcal{I}$ we have $\boldsymbol{j}_1 \preceq \boldsymbol{j}_2$ iff $j_1^{(1)} \leq j_2^{(1)}$ and $j_1^{(2)} \leq j_2^{(2)}$. Then, a vector $\boldsymbol{\beta} \in \mathbb{R}^n$, with $n = n_1 n_2$, indexed by $\mathcal{I}$ is called bimonotone if it is isotonic with respect to bimonotone order $\preceq$ defined on its index $\mathcal{I}$. Further, we define the directed graph $G = (V, E)$ with vertexes $V = \mathcal{I}$, and the edges

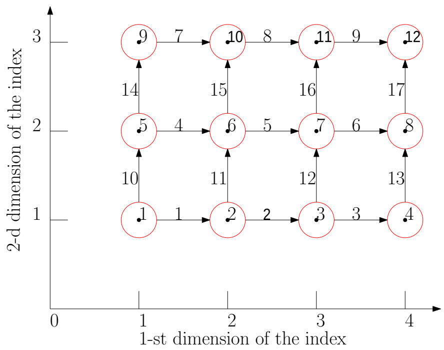
$$\begin{aligned} E = &\{((l, k), (l, k+1)) : 1 \leq l \leq n_1, 1 \leq k \leq n_2 - 1\} \\ &\cup \{((l, k), (l+1, k)) : 1 \leq l \leq n_1 - 1, 1 \leq k \leq n_2\}. \end{aligned}$$

The labeled directed graph for bimonotone constraints and its incidence matrix are displayed on Figure 2.

1.2 General statement of the problem

Now we can state the general problem studied in this paper. Let $\boldsymbol{y} \in \mathbb{R}^n$ be a signal indexed by the index set $\mathcal{I}$ with the partial order relation $\preceq$ defined on $\mathcal{I}$. Next, let $G = (V, E)$ be the directed graph corresponding to $\preceq$ on $\mathcal{I}$. The fused lasso nearly-isotonic signal approximation with respect to $\preceq$ on $\mathcal{I}$ (or, equivalently, to the directed graph $G = (V, E)$, corresponding to $\preceq$) is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min}\, &\frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_F \sum_{(\boldsymbol{i},\boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}| \\ &+ \lambda_L ||\boldsymbol{\beta}||_1 + \lambda_{NI} \sum_{(\boldsymbol{i},\boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}|_+. \end{aligned} \tag{14}$$



(a) Graph $G = (V, E)$

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & -1 \\ 1 & 0 & 0 & 0 & -1 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & -1 \end{pmatrix}$$

(b) Oriented incidence matrix $D \in \mathbb{R}^{17\times 12}$ for the graph $G = (V, E)$

Fig. 2: Graph for bimonotonic contstraints and oriented incidence matrix

Therefore, the estimator in (14) is a combination of the estimators in (10) and (12).

Equivalently, we can rewrite the problem in the following way:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda_F||D\boldsymbol{\beta}||_1 \\ + \lambda_L||\boldsymbol{\beta}||_1 + \lambda_{NI}||D\boldsymbol{\beta}||_+, \end{aligned} \tag{15}$$

where D is the oriented incidence matrix of the graph $G = (V, E)$. Here we clarify that in the case of penalisation with the incidence matrix D we assume that $\boldsymbol{\beta}$ is indexed according to the indexing of the edges in the graph $G = (V, E)$. Analogously to the definition in one dimensional case, if $\lambda_L = 0$ we call the estimator as fused nearly-isotonic approximator and denote it by $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$.

Here it is worth to mention recent papers in constrained estimation [19–21], where the authors studied the asymptotic properties of the isotonic regression in general dimensions. Also, in paper [22] ℓ_1-trend filterin was generalised for the case of a general graph.

1.3 Organisation of the paper

The rest of the paper is organized as follows. In Section 2 we provide the numerical solution to the fussed lasso nearly-isotonic signal approximator. Section 3 is dedicated to the theoretical properties of the estimator. We show how the solutions to the fussed lasso nearly-isotonic regression, fussed lasso and nearly-isotonic regression are related to each other. Also, we prove that in one-dimensional case the new estimator has agglomerative property and the procedures of near-isotonisation and fusion can be swaped and provide the solution to the original problem. Next, in Section 4 we derive the unbiased estimator of the degrees of freedom of the estimator. Further, in Section 5 we discuss the computational aspects, do the simulation study and show that the estimator is computationally feasible for moderately large data sets. Also, we illustrate the usage of the estimator for the real data set. The article closes with a conclusion and a discussion of possible generalisations in Section 6. The proofs of all results are given in Appendix. The R and Python implementations of the estimator are available upon request.

2 Solution to the fused lasso nearly-isotonic signal approximator

First, we consider fused nearly-isotonic regression, i.e. in (15) we assume that $\lambda_L = 0$.

Theorem 1. *For a fixed data vector $\boldsymbol{y} \in \mathbb{R}^n$ indexed by the index set $\mathcal{I}$ with the partial order relation $\preceq$ defined on $\mathcal{I}$ the solution to the fused nearly-isotonic problem in (15) is given by*

$$\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \boldsymbol{y} - D^T \hat{\boldsymbol{\nu}}(\lambda_F, \lambda_{NI}) \tag{16}$$

with

$$\hat{\boldsymbol{\nu}}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \underset{\boldsymbol{\nu} \in \mathbb{R}^m}{\arg\min} \frac{1}{2} ||\boldsymbol{y} - D^T \boldsymbol{\nu}||_2^2 \quad s.\ t. \quad -\lambda_F \mathbf{1} \leq \boldsymbol{\nu} \leq (\lambda_F + \lambda_{NI})\mathbf{1}, \tag{17}$$

where D is the incidence matrix of the directed graph $G = (V, E)$ with n vertices and m edges corresponding to $\preceq$ on $\mathcal{I}$, $\mathbf{1} \in \mathbb{R}^m$ is the vector whose all elements are equal to 1 and the notation $\boldsymbol{a} \leq \boldsymbol{b}$ for vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$ means $a_i \leq b_i$ for all $i = 1, \ldots, m$.

Next, we provide the solution to the fused lasso nearly-isotonic regression.

Theorem 2. *For a given vector $\boldsymbol{y}$ indexed by $\mathcal{I}$ the solution to the fused lasso nearly-isotonic signal approximator $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI})$ is given by soft thresholding the fused nearly-isotonic regression $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$, i.e.*

$$\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \begin{cases} \hat{\beta}_{\boldsymbol{i}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) - \lambda_L, & \text{if } \hat{\beta}_{\boldsymbol{i}}^{FNI} \geq \lambda_L, \\ 0, & \text{if } |\hat{\beta}_{\boldsymbol{i}}^{FNI}| \leq \lambda_L, \\ \hat{\beta}_{\boldsymbol{i}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) + \lambda_L, & \text{if } \hat{\beta}_{\boldsymbol{i}}^{FNI} \leq -\lambda_L, \end{cases} \tag{18}$$

for $\boldsymbol{i} \in \mathcal{I}$.

From this result we can conclude that adding lasso penalisation does not add much to the computational complexity of the solution. The computational aspects of fussed nearly-isotonic approximator will be discussed in the Section 5 below. In the next section we discuss properties of the fussed lasso nearly-isotonic regression.

3 Properties of the fused lasso nearly-isotonic signal approximator

We start with a proposition which shows how the solutions to the optimization problems (11), (10) and (15) are related to each other. This result will be used in the next section to derive degrees of freedom of the fused lasso nearly-isotonic signal approximator.

Proposition 3. *For a fixed data vector $\boldsymbol{y}$ indexed by $\mathcal{I}$ and penalisation parameters λ_{NI} and λ_F the following relations between estimators $\hat{\boldsymbol{\beta}}^F$, $\hat{\boldsymbol{\beta}}^{NI}$ and $\hat{\boldsymbol{\beta}}^{FNI}$ hold*

$$\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^F(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI}), \tag{19}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) &= \hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y} + \lambda_F D^T \mathbf{1}, \lambda_{NI} + 2\lambda_F) \\ &= \hat{\boldsymbol{\beta}}^F(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F) \end{aligned} \tag{20}$$

and

$$\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F, \lambda_L), \tag{21}$$

where D is the oriented incidence matrix for the graph $G = (V, E)$ corresponding to the partial order relation $\preceq$ on $\mathcal{I}$.

Further, let us introduce two "naive" versions of $\hat{\boldsymbol{\beta}}^{FNI}$. Instead of simultaniously penalise by fusion and isotonisation we consider the following two-step procedures:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) &= \hat{\boldsymbol{\beta}}^{NI}(\hat{\boldsymbol{\beta}}^F(\boldsymbol{y}, \lambda_F), \lambda_{NI}) \\ &\equiv \arg\min_{\boldsymbol{\beta} \in \mathsf{R}^n} \frac{1}{2} ||\hat{\boldsymbol{\beta}}^F(\boldsymbol{y}, \lambda_F) - \boldsymbol{\beta}||_2^2 + \lambda_{NI} \sum_{(\boldsymbol{i}, \boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}|_+, \end{aligned} \tag{22}$$

and

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{NI\to F}(\boldsymbol{y}, \lambda_{NI}, \lambda_F) &= \hat{\boldsymbol{\beta}}^F(\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}), \lambda_F) \\ &\equiv \arg\min_{\boldsymbol{\beta} \in \mathsf{R}^n} \frac{1}{2} ||\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) - \boldsymbol{\beta}||_2^2 + \lambda_F \sum_{(\boldsymbol{i}, \boldsymbol{j}) \in E} |\beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}}|. \end{aligned} \tag{23}$$

Below we prove that both "naive" methods in one dimensional case with a simple monotonic restriction defined above are not only equivalent, but both methods provide the solution to the fused nearly-isotonic regression.

First, we have to prove that, analogously to fused lasso and nearly-isotonic regression, as one of the penalization parameters increases the constant regions in the solution $\hat{\boldsymbol{\beta}}^{FLNI}$ can only be joined together and not split apart. In the paper [12] this property of the estimator was called as agglomerative property. We prove this result only for one dimensional monotonic order, and the general case is an open question.

Proposition 4. *(Agglomerative property of FLNI estimator) Let $\mathcal{I} = \{1, \dots, n\}$ with the natural order for integers defined on it. Next, let $\boldsymbol{\lambda} = (\lambda_F, \lambda_L, \lambda_{NI})$ and $\boldsymbol{\lambda}^* = (\lambda_F^*, \lambda_L^*, \lambda_{NI}^*)$ are the triples of penalisation parameters such that one of the elements of $\boldsymbol{\lambda}^*$ is greater than the corresponding element in $\boldsymbol{\lambda}$, while two others are the same. Next, assume that for some i the solution $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda})$ satisfies*

$$\hat{\beta}_i^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}) = \hat{\beta}_{i+1}^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}).$$

Then for $\boldsymbol{\lambda}^*$ we have

$$\hat{\beta}_i^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*) = \hat{\beta}_{i+1}^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*).$$

Now we can prove the commutability property of the "naive" estimators and the equivalence of the approach to the fussed nearly-isotonic regression.

Theorem 5. *(Commutability property of FNI estimator)*
Let $\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$ and $\hat{\boldsymbol{\beta}}^{NI\to F}(\boldsymbol{y}, \lambda_{NI}, \lambda_F)$ be the "naive" versions of the fussed nearly-isotonic approximator, defined in (22) and (23), in the case of one-dimensional monotonic constraint. Then, we have

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{NI\to F}(\boldsymbol{y}, \lambda_{NI}, \lambda_F) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}).$$

One of the first conclusions of Theorem 5 is commutability of strict isotonisation (which corresponds to the large values of λ_{NI}) and fusion. For big values of λ_{NI} fussed lasso nearly-isotonic signal approximation is, in principle, analogous to the approach studied in [23], where the authors studied estimation of isotonic piecewise constant signals solving the following optimization problem

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}^{is}_{n,k}}{\arg\min} \sum_{j=1}^{n} (\beta_j - y_j)^2 + pen(n, k), \tag{24}$$

where

$$\begin{aligned}\boldsymbol{\mathcal{B}}^{is}_{n,k} = \{\boldsymbol{\beta} \in \mathbb{R}^n :\ & \text{there exists } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that} \\ & 0 \le a_0 \le a_1 \le \cdots \le a_k = n, \\ & \mu_1 \le \mu_2 \le \cdots \le \mu_k, \text{ and } \beta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j]\},\end{aligned}$$

and $pen(n, k)$ is a penalization term which depends on n and k but not on $\boldsymbol{y}$. Therefore, the result of Theorem 5 provides an alternative approach to obtain exact solution in the estimation isotonic piecewise constant signals.

4 Degrees of freedom

In this section we discuss the estimation of the degrees of freedom for the fused nearly-isotonic regression and the fused lasso nearly-isotonic signal approximator. Let us consider the following nonparametric model

$$\boldsymbol{Y} = \mathring{\boldsymbol{\beta}} + \boldsymbol{\varepsilon},$$

where $\mathring{\boldsymbol{\beta}} \in \mathbb{R}^n$ is an unknown signal, and the error term $\boldsymbol{\varepsilon} \in \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$, with $\sigma < \infty$.

The degrees of freedom is a measure of complexity of the estimator, and following [24], for the fixed values of λ_F, λ_L and λ_{Ni} the degrees of freedom of $\hat{\boldsymbol{\beta}}^{FNI}$ and $\hat{\boldsymbol{\beta}}^{FLNI}$ are given by

$$df(\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI})) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}[\hat{\beta}_i^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI}), Y_i] \tag{25}$$

and

$$df(\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI})) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}[\hat{\beta}_i^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI}), Y_i]. \tag{26}$$

The next theorem provides the unbiased estimators of the degrees of freedom $df(\hat{\boldsymbol{\beta}}^{FNI})$ and $df(\hat{\boldsymbol{\beta}}^{FLNI})$.

Theorem 6. *For the fixed values of λ_F, λ_L and λ_{Ni} let*

$$K^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \#\{\textit{fused groups in } \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})\},$$

and

$$K^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \#\{\textit{non-zero fused groups in } \hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI})\}.$$

Then we have

$$\mathbb{E}[K^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI})] = df(\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI})),$$

and

$$\mathbb{E}[K^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI})] = df(\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI})).$$

We can potentially use the estimate of degrees of freedom for unbiased estimation of the true risk $\mathbb{E}[\sum_{i=1}^{n}(\mathring{\beta}_i - \hat{\beta}_i^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI}))^2]$, which is given by $\hat{C}_p$ statistic

$$\hat{C}_p(\lambda_F, \lambda_L, \lambda_{NI}) = \sum_{i=1}^{n}(y_i - \hat{\beta}_i^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}))^2 - n\sigma^2 + 2\sigma^2 K^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI}).$$

Though, we note that in a real application the variance σ^2 is unknown. The variance estimator for the case of one-dimensional isotonic regression was introduced in [25]. To the authors' knowledge, the variance estimator even for one dimensional nearly-isotonic regression is an open problem.

5 Computational aspects, simulation study and application to a real data set

First of all, recall that the dual of (6) is given by

$$\hat{\boldsymbol{\nu}}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \operatorname*{arg\,min}_{\boldsymbol{\nu}\in\mathsf{R}^m} \frac{1}{2}||\boldsymbol{y} - D^T\boldsymbol{\nu}||_2^2 \quad \text{subject to} \quad -\lambda_F\mathbf{1} \leq \boldsymbol{\nu} \leq (\lambda_F + \lambda_{NI})\mathbf{1},$$

where D is the incidence matrix displayed in Figure 1 (a) for one-dimensional case. The matrix D is full raw ranked, therefore, the problem is strictly convex. Next, we have similar box-type constraints as in the problem of L_1-trend filtering example and we can solve the problem with $\mathcal{O}(n)$ time complexity.

Second, note that in one-dimensional case the time complexities of path solution algorithms for nearly-isotonic regression and fusion approximator are equal to $\mathcal{O}(n\log(n))$, cf. [11, 17, 26] with the references therein. Therefore, if we have λ_F fixed, then using the result of Theorem 5 we can get the solution path with respect to λ_{NI} with the time complexity $\mathcal{O}(n\log(n))$. Further, if we fix λ_{NI} then, again, using Theorem 5 we can obtain the solution path with respect to λ_F with complexity $\mathcal{O}(n\log(n))$. In the paper [13] one-dimensional fussed nearly-isotonic regression was solved for fixed values of penalisation parameters. Therefore, one dimentional fussed lasso and nearly-isotonic regression have been studied in detail, therefore, in our paper we focus in two-dimensional case.

The case of several dimensions is more complicated. Note, that, for example, even in the case of two dimensions the matrix D, displayed on Figure 2, is not full raw ranked. Therefore, the dual problem is not strictly convex. At the same time one can see that the matrix D is sparse digaonal. Therefore, we apply recently developed algorithm OSQP algorithm, cf. [27]. The time complexity of the solution is linear with respect to the number of edges in the graph, i.e. it is $\mathcal{O}(|E|)$.

The exact solution for fixed values of penalisation parameters can be obtained using results of the paper [12], where the author proposed the algorithm for a general graph with computational complexity $\mathcal{O}(n|E|\log(\frac{n^2}{|E|}))$. Therefore, in principle, using the relation between fused nearly-isotonic regression and nearly-isotonic regression proved in Proposition 3 it is possible to obtain exact solution to the fussed nearly-isotonic approximation for a general graph.

First, recall that from Theorem 2 it follows that the solution with $\lambda_L \neq 0$ is given by soft-thresholding of the solution with $\lambda_L = 0$. Therefore, lasso penalization does not add much to the complexity, and we concentrate on the case with $\lambda_L = 0$. Following [12], we use the following bi-monotone functions (bisigmoid and bicubic) to test the performance of the fused nearly-isotonic approximator:

$$\begin{aligned} f_{bs}(x^{(1)}, x^{(2)}) &= \frac{1}{2}\Big(\frac{e^{16x^{(1)}-8}}{1+e^{16x^{(1)}-8}} + \frac{e^{16x^{(2)}-8}}{1+e^{16x^{(2)}-8}}\Big), \\ f_{bc}(x^{(1)}, x^{(2)}) &= \frac{1}{2}\Big((2x^{(1)}-1)^3 + (2x^{(2)}-1)^3\Big) + 2, \end{aligned}$$

where $x^{(1)} \in [0,1)$ and $x^{(2)} \in [0,1)$.

The simulation experiment is performed in the following way. First, we generate homogeneous grid $k \times k$:

$$x_k^{(1)} = \frac{k-1}{d} \quad \text{and} \quad x_k^{(2)} = \frac{k-1}{d},$$

for $k = 1, \ldots, d$. The size of the side d varies in $\{2\times10^2, 4\times10^2, 6\times10^2, 8\times10^2, 10^3\}$. Next, we uniformly generate penalisation parameters λ_F and λ_{NI} from $U(0,5)$. We perform 10 runs and compute computational times for each d. Analogously to [27], we consider two cases of OSQP algorithm: low precision case with $\varepsilon_{abs} = \varepsilon_{rel} = 10^{-3}$, and high precision case with $\varepsilon_{abs} = \varepsilon_{rel} = 10^{-5}$ (for the details of the settings in OSQP we refer to [27]). Figure 3 below provides these computational times. All the computations were performed on MacBook Air (Apple M1 chip), 16 GB RAM. From these results we can conclude that the estimator is computationally feasible for moderate sized data sets (i.e. for the grids with millions of nodes).

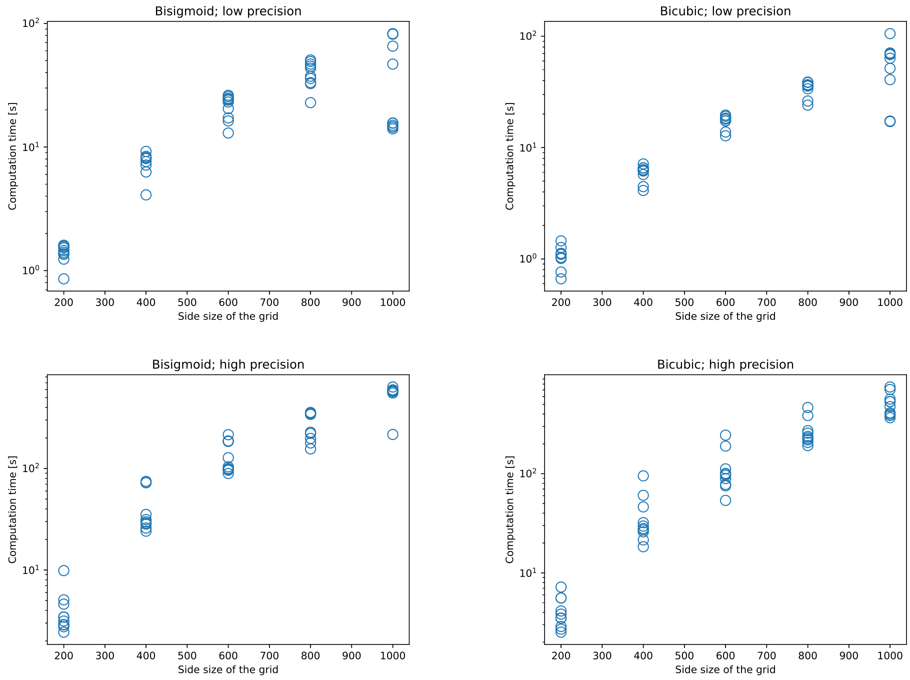


Fig. 3: Computational times vs side size of a square grid for OSQP solution of fussed nearly-isotonic approximator in two dimensions

Next, Figure 4 visualizes the fussed nearly-isotonic approximator. We use Adult data set, available from the UCI Machine Learning repository [28]. The target variable in this data set is either a person's salary is greater than 50 000 dollars per year or less. We use two features (education number and working hours per week) and each bar at the figure is the proportion of people making more that the amount of money mentioned above. This data set was used, for example, in [29].

From Figure 4 we can see that fussed nearly-isotonic regression provides a trade-off between monotonicity, block sparsity and goodness-of-fit.

6 Conclusion and discussion

In this paper we introduced and studied fussed lasso nearly-isotonic signal approiximator in general dimensions. The main result is that the estimator is computationally feasible and it provides interplay between fusion and monotonisation. Also, we proved

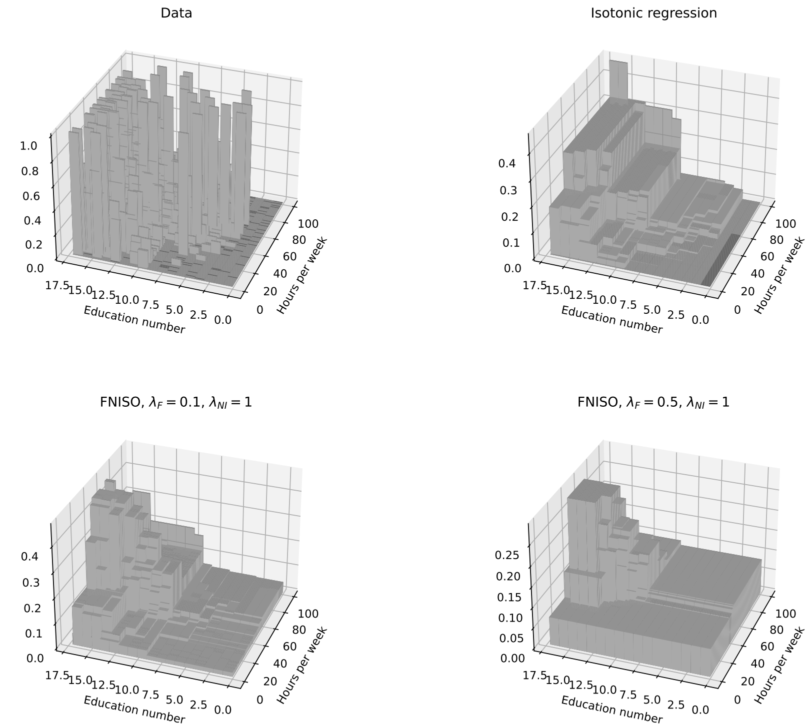


Fig. 4: Data visualisation for different levels of fusion and isotonisation

that the properties of new estimator are very similar to the properties of fusion estimator and nearly-isotonic regression.

In our opinion, one of the most important results is Theorem 5, where we proved the commutability property of fusion and nearly-isotonisation, because for the fixed values of one of the penalisation parameters we can immediately obtain the path solution with respect to the other one. Path algorithm for fussed lasso exists [7, 17]. At the same time, to the authors' knowledge, path algorithm for nearly-isotonic regression in general dimensions has not been developed yet. Therefore, further direction could be the solution for the nearly-isotonic regression, and, next, to prove if commutability holds in a general dimensional case.

One of the other possible direction is to study the asymptotic properties. In particular, it is interesting to understand the rate of convergence for different model selection and cross-validation procedures of choosing penalisation parameters.

Another direction is to study properties of the solution when λ_F and λ_{NI} are not the same for each vertex. An example where one must use different penalisation parameters is the case when the data points are measured along non-homogeneously spaced grid. It is important to note that, as discussed in [12], this case is different and even in one dimensional case the estimator will behave differently. In particular, agglomerative property of the nearly-isotonic regression holds if the penalisation parameters satisfy the certain relatio, cf. Proposition A.1. in [12], which is crucial for the solution path.

Finally, in our opinion, it is interesting to study different combinations of penalisation estimators, even though, practically, in this case one needs more data, because there will be more penalisation parameters to estimate.

Supplementary information. Not applicable

Acknowledgments. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundataion.

Appendix A Proofs of the results

Proof of Theorem 1. First, following the derivations of ℓ_1 trend filtering and generalised lasso in [30] and [7], respectively, we can write the optimization problem in (6) in the following way

$$\underset{\boldsymbol{\beta},\boldsymbol{z}}{\text{minimize}} \frac{1}{2}||\boldsymbol{y}-\boldsymbol{\beta}||_2^2 + \lambda_F||\boldsymbol{z}||_1 + \lambda_{NI}||\boldsymbol{z}||_+ \quad \text{subject to} \quad D\boldsymbol{\beta} = \boldsymbol{z} \in \mathbb{R}^m.$$

Further, the Lagrangian is given by

$$L(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{\nu}) = \frac{1}{2}||\boldsymbol{y}-\boldsymbol{\beta}||_2^2 + \lambda_F||\boldsymbol{z}||_1 + \lambda_{NI}||\boldsymbol{z}||_+ + \boldsymbol{\nu}^T(D\boldsymbol{\beta}-\boldsymbol{z}), \tag{A1}$$

where $\boldsymbol{\nu} \in \mathbb{R}^m$ is a dual variable.

Note that

$$\min_{\boldsymbol{z}}\big(\lambda_F||\boldsymbol{z}||_1 + \lambda_{NI}||\boldsymbol{z}||_+ - \boldsymbol{\nu}^T\boldsymbol{z}\big) = \begin{cases} 0, & \text{if} \quad -\lambda_F\mathbf{1} \le \boldsymbol{\nu} \le (\lambda_F+\lambda_{NI})\mathbf{1}, \\ -\infty, & \text{otherwise,} \end{cases}$$

and

$$\min_{\boldsymbol{\beta}}\big(\frac{1}{2}||\boldsymbol{y}-\boldsymbol{\beta}||_2^2 + \boldsymbol{\nu}^TD\boldsymbol{\beta}\big) = -\frac{1}{2}\boldsymbol{\nu}^TDD^T\boldsymbol{\nu} + \boldsymbol{y}^TD^T\boldsymbol{\nu} = -\frac{1}{2}||\boldsymbol{y}-D^T\boldsymbol{\nu}||_2^2 + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}.$$

Next, the dual function is given by

$$g(\nu) = \min_{\boldsymbol{\beta},\boldsymbol{z}} L(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{\nu}) = \begin{cases} -\frac{1}{2}||\boldsymbol{y}-D^T\boldsymbol{\nu}||_2^2 + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}, & \text{if} \quad -\lambda_F\mathbf{1} \le \boldsymbol{\nu} \le (\lambda_F+\lambda_{NI})\mathbf{1}, \\ -\infty, & \text{otherwise,} \end{cases}$$

and, therefore, the dual problem is

$$\hat{\boldsymbol{\nu}}(\boldsymbol{y},\lambda_F,\lambda_{NI}) = \arg\max_{\boldsymbol{\nu}} g(\nu) \quad \text{subject to} \quad -\lambda_F\mathbf{1} \le \boldsymbol{\nu} \le (\lambda_F+\lambda_{NI})\mathbf{1},$$

which is equivalent to

$$\hat{\boldsymbol{\nu}}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \arg\min_{\boldsymbol{\nu}} \frac{1}{2}||\boldsymbol{y} - D^T\boldsymbol{\nu}||_2^2 \quad \text{subject to} \quad -\lambda_F \mathbf{1} \leq \boldsymbol{\nu} \leq (\lambda_F + \lambda_{NI})\mathbf{1}.$$

Lastly, taking first derivative of Lagrangian $L(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\nu})$ with respect to $\boldsymbol{\beta}$ we get the following relation between $\hat{\boldsymbol{\beta}}^{FNI}(\lambda_F, \lambda_{NI})$ and $\hat{\boldsymbol{\nu}}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$

$$\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \boldsymbol{y} - D^T\hat{\boldsymbol{\nu}}(\boldsymbol{y}, \lambda_F, \lambda_{NI}).$$

□

Proof of Theorem 2. The proof is similar to the derivation of solution of the fused lasso in [16]. Nevertheless, for completeness of the paper we provide the proof for $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI})$.

The subgradient equations (which are necessary and sufficient conditions for the solution of (5)) for $\beta_{\boldsymbol{i}}$, with $\boldsymbol{i} \in \mathcal{I}$, are

$$\begin{aligned} g_{\boldsymbol{i}}(\lambda_L) = & -(y_{\boldsymbol{i}} - \beta_{\boldsymbol{i}}) + \lambda_{NI}\Big(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}} - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}\Big) \\ & + \lambda_F\Big(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}} - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}\Big) + \lambda_L s_i = 0, \end{aligned} \tag{A2}$$

where

$$q_{\boldsymbol{i},\boldsymbol{j}} : \begin{cases} = 1, & \text{if } \beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}} > 0, \\ = 0, & \text{if } \beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}} < 0, \\ \in [0,1], & \text{if } \beta_{\boldsymbol{i}} = \beta_{\boldsymbol{j}}, \end{cases} \qquad t_{\boldsymbol{i},\boldsymbol{j}} : \begin{cases} = 1, & \text{if } \beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}} > 0, \\ = -1, & \text{if } \beta_{\boldsymbol{i}} - \beta_{\boldsymbol{j}} < 0, \\ \in [-1,1], & \text{if } \beta_{\boldsymbol{i}} = \beta_{\boldsymbol{j}}, \end{cases} \tag{A3}$$

$$s_{\boldsymbol{i}} : \begin{cases} = 1, & \text{if } \beta_{\boldsymbol{i}} > 0, \\ = -1, & \text{if } \beta_{\boldsymbol{i}} < 0, \\ \in [-1,1], & \text{if } \beta_{\boldsymbol{i}} = 0. \end{cases}$$

Next, let $q_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L)$, $t_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L)$ and $s_{\boldsymbol{i}}(\lambda_L)$ denote the values of the parameters defined above at some value of λ_L. Note, the values of λ_{NI} and λ_F are fixed. Therefore, if $\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) \neq 0$ for $s_{\boldsymbol{i}}(0)$ we have

$$s_{\boldsymbol{i}}(0) = \begin{cases} 1, & \text{if } \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) > 0, \\ -1, & \text{if } \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) < 0, \end{cases}$$

and for $\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) = 0$ we can set $s_{\boldsymbol{i}}(0) = 0$.

Next, let $\hat{\boldsymbol{\beta}}^{ST}(\lambda_L)$ denote the soft thresholding of $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})$, i.e.

$$\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) = \begin{cases} \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) - \lambda_L, & \text{if} \quad \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) \geq \lambda_L, \\ 0, & \text{if} \quad |\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})| \leq \lambda_L, \\ \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) + \lambda_L, & \text{if} \quad \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) \leq -\lambda_L. \end{cases}$$

The goal is to prove that $\hat{\boldsymbol{\beta}}^{ST}(\lambda_L)$ provides the solution to (14).

Note, analogously to the proof for the fused lasso estimator in Lemma A.1 at [16], if either $\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) \neq 0$ or $\hat{\beta}_{\boldsymbol{j}}^{ST}(\lambda_L) \neq 0$, and $\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) < \hat{\beta}_{\boldsymbol{j}}^{ST}(\lambda_L)$ or $\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) > \hat{\beta}_{\boldsymbol{j}}^{ST}(\lambda_L)$, then we also have $\hat{\beta}_{\boldsymbol{i}}^{ST}(0) < \hat{\beta}_{\boldsymbol{j}}^{ST}(0)$ or $\hat{\beta}_{\boldsymbol{i}}^{ST}(0) > \hat{\beta}_{\boldsymbol{j}}^{ST}(0)$, respectively. Therefore, soft thresholding of $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})$ does not change the ordering of these pairs and we have $q_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) = q_{\boldsymbol{i},\boldsymbol{j}}(0)$ and $t_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) = t_{\boldsymbol{i},\boldsymbol{j}}(0)$. Next, if for some $(\boldsymbol{i}, \boldsymbol{j}) \in E$ we have $\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) = \hat{\beta}_{\boldsymbol{j}}^{ST}(\lambda_L) = 0$, then $q_{\boldsymbol{i},\boldsymbol{j}} \in [0,1]$ and $t_{\boldsymbol{i},\boldsymbol{j}} \in [-1,1]$, and, again, we can set $t_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) = t_{\boldsymbol{i},\boldsymbol{j}}(0)$, and $q_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) = q_{\boldsymbol{i},\boldsymbol{j}}(0)$.

Now let us insert $\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L)$ into subgradient equations (A2) and show that we can find $s_{\boldsymbol{i}}(\lambda_L) \in [0,1]$, for all $\boldsymbol{i} \in \mathcal{I}$.

First, assume that for some $\boldsymbol{i}$ we have $\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) \geq \lambda_L$. Then

$$\begin{aligned} g_{\boldsymbol{i}}(\lambda_L) = & -(y_{\boldsymbol{i}} - \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})) - \lambda_L \\ & + \lambda_{NI}(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}(\lambda_L)) \\ & + \lambda_F(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}(\lambda_L)) + \lambda_L s_{\boldsymbol{i}}(\lambda_L) \\ = & -(y_{\boldsymbol{i}} - \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})) \\ & + \lambda_{NI}(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}(0)) \\ & + s\lambda_F(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}(0)) + \lambda_L s_{\boldsymbol{i}}(\lambda_L) - \lambda_L = 0. \end{aligned}$$

Note, that

$$\begin{aligned} & -(y_i - \hat{\beta}_i^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})) + \lambda_{NI}(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}(0)) \\ & + \lambda_F(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}(0)) = 0, \end{aligned}$$

because $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) \equiv \hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})$.
Therefore, if $s_i(\lambda_L) = \text{sign}\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) = 1$, then $g_{\boldsymbol{i}}(\lambda_L) = 0$.

The proof for the case when $\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI}) \leq -\lambda_L$ is similar and one can show that $g_{\boldsymbol{i}}(\lambda_L) = 0$ if $s_{\boldsymbol{i}}(\lambda_L) = \text{sign}\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) = -1$.

Second, assume that $|\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})| < \lambda_L$. Then, $\hat{\beta}_{\boldsymbol{i}}^{ST}(\lambda_L) = 0$, and

$$\begin{aligned} g_{\boldsymbol{i}}(\lambda_L) = & -y_{\boldsymbol{i}} + \lambda_{NI}(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}(\lambda_L)) \\ & + \lambda_F(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}}(\lambda_L) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}(\lambda_L)) + \lambda_L s_{\boldsymbol{i}}(\lambda_L) \\ = & -y_{\boldsymbol{i}} + \lambda_{NI}(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}(0)) \\ & + \lambda_F(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}(0)) + \lambda_L s_{\boldsymbol{i}}(\lambda_L) = 0. \end{aligned}$$

Next, if we let $s_{\boldsymbol{i}}(\lambda_L) = \hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})/\lambda_L$, then, again, we have

$$\begin{aligned} g_{\boldsymbol{i}}(\lambda_L) = & -(y_i - \hat{\beta}_i^{FLNI}(\boldsymbol{y}, \lambda_F, 0, \lambda_{NI})) \\ & + \lambda_{NI}(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} q_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} q_{\boldsymbol{j},\boldsymbol{i}}(0)) \\ & + \lambda_F(\sum_{\boldsymbol{j}:(\boldsymbol{i},\boldsymbol{j})\in E} t_{\boldsymbol{i},\boldsymbol{j}}(0) - \sum_{\boldsymbol{j}:(\boldsymbol{j},\boldsymbol{i})\in E} t_{\boldsymbol{j},\boldsymbol{i}}(0)) = 0, \end{aligned}$$

Therefore, we have proved that $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{ST}(\lambda_L)$. □

Proof of Proposition 3. First, from [11] the solution to the nearly-isotonic problem is given by

$$\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) = \boldsymbol{y} - D^T \hat{\boldsymbol{v}}(\boldsymbol{y}, \lambda_{NI}),$$

with

$$\hat{\boldsymbol{v}}(\boldsymbol{y}, \lambda_{NI}) = \underset{\boldsymbol{v}\in\mathsf{R}^{n-1}}{\arg\min} \frac{1}{2}||\boldsymbol{y} - D^T\boldsymbol{v}||_2^2 \quad \text{subject to} \quad \mathbf{0} \le \boldsymbol{v} \le \lambda_{NI}\mathbf{1},$$

and from [7] it follows

$$\hat{\boldsymbol{\beta}}^{F}(\boldsymbol{y}, \lambda_F) = \boldsymbol{y} - D^T \hat{\boldsymbol{w}}(\boldsymbol{y}, \lambda_F),$$

with

$$\hat{\boldsymbol{w}}(\boldsymbol{y}, \lambda_F) = \underset{\boldsymbol{w}\in\mathsf{R}^{n-1}}{\arg\min} \frac{1}{2}||\boldsymbol{y} - D^T\boldsymbol{w}||_2^2 \quad \text{subject to} \quad -\lambda_F\mathbf{1} \le \boldsymbol{w} \le \lambda_F\mathbf{1}.$$

Second, let us introduce a new variable $\boldsymbol{v}^* = \boldsymbol{v} - \frac{\lambda_{NI}}{2}\mathbf{1}$. Then

$$\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) = \boldsymbol{y} - D^T\frac{\lambda_{NI}}{2}\mathbf{1} - D^T \hat{\boldsymbol{v}}^*(\boldsymbol{y}, \lambda_{NI}),$$

where

$$\hat{\boldsymbol{v}}^*(\boldsymbol{y}, \lambda_{NI}) = \underset{\boldsymbol{v}^*\in\mathsf{R}^{n-1}}{\arg\min} \frac{1}{2}||\boldsymbol{y} - D^T\frac{\lambda_{NI}}{2}\mathbf{1} - D^T\boldsymbol{v}^*||_2^2 \quad \text{s. t.} \quad -\frac{\lambda_{NI}}{2}\mathbf{1} \le \boldsymbol{v}^* \le \frac{\lambda_{NI}}{2}\mathbf{1}.$$

Therefore, we have proved that $\hat{\boldsymbol{\beta}}^{NI}(\boldsymbol{y}, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{F}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI})$.

The proof for the fused lasso nearly-isotonic estimator is the same with the change of variable $\boldsymbol{u}^* = \boldsymbol{u} + D^T \lambda_F \mathbf{1}$ in (16) and (17) for the proof of the first equality in (20) and with $\boldsymbol{u}^* = \boldsymbol{u} - \frac{\lambda_{NI}}{2}\mathbf{1}$ for the second equality.

Next, we prove the result for the case of fused lasso nearly-isotonic approximator. From Theorem 2 we have

$$\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \begin{cases} \hat{\beta}_{\boldsymbol{i}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) - \lambda_L, & \text{if} \quad \hat{\beta}_{\boldsymbol{i}}^{FNI} \geq \lambda_L, \\ 0, & \text{if} \quad |\hat{\beta}_{\boldsymbol{i}}^{FNI}| \leq \lambda_L, \\ \hat{\beta}_{\boldsymbol{i}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) + \lambda_L, & \text{if} \quad \hat{\beta}_{\boldsymbol{i}}^{FNI} \leq -\lambda_L, \end{cases}$$

for $\boldsymbol{i} \in \mathcal{I}$.

Further, using (20) we have

$$\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \hat{\beta}_{\boldsymbol{i}}^{F}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F) - \lambda_L,$$

if $\hat{\beta}_{\boldsymbol{i}}^{F}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F) \geq \lambda_L$,

$$\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = 0,$$

if $|\hat{\beta}_{\boldsymbol{i}}^{F}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F)| \leq \lambda_L$,

$$\hat{\beta}_{\boldsymbol{i}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \hat{\beta}_{\boldsymbol{i}}^{F}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F) + \lambda_L,$$

if $\hat{\beta}_{\boldsymbol{i}}^{F}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F) \leq -\lambda_L$.

Therefore, we obtain

$$\begin{aligned} &\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \\ &\qquad \arg\min_{\boldsymbol{\beta} \in \mathsf{R}^n} \frac{1}{2}||\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1} - \boldsymbol{\beta}||_2^2 + (\frac{1}{2}\lambda_{NI} + \lambda_F)||D\boldsymbol{\beta}||_1 + \lambda_L||\boldsymbol{\beta}||_1 \equiv \\ &\qquad \hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F, \lambda_L). \end{aligned}$$

□

Let us consider the following cases separately.

Case 1: λ_{NI} and λ_F are fixed and $\lambda_L^* > \lambda_L$. The result of the proposition for this case follows directly from Theorem 2.

Case 2: λ_F and λ_L are fixed and $\lambda_{NI}^* > \lambda_{NI}$. Let us consider the fused nearly-isotonic regression and write the subgradient equations

$$g_i(\lambda_{NI}) = -(y_i - \beta_i) + \lambda_{NI}(q_i(\lambda_{NI}) - q_{i-1}(\lambda_{NI})) + \lambda_F(t_i(\lambda_{NI}) - t_{i-1}(\lambda_{NI})) = 0,$$

where q_i and t_i, with $i = 1, \dots, n$, are defined in (A3), and, analogously to the proof of Theorem 2, $q(\lambda_{NI})$, $t(\lambda_{NI})$ denote the values of the parameters defined above at some value of λ_{NI}.

Assume that for λ_{NI} in the solution $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$ we have a following constant region

$$\begin{aligned} \hat{\beta}^{FNI}_{j-1}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) \neq \hat{\beta}^{FNI}_{j}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \dots \\ = \hat{\beta}^{FNI}_{j+k}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) \neq \hat{\beta}^{FNI}_{j+k+1}(\boldsymbol{y}, \lambda_F, \lambda_{NI}), \end{aligned} \tag{A4}$$

and in the same way as in [11] for λ^*_{NI} we consider the subset of the subgradient equations

$$\begin{aligned} g_i(\lambda_{NI}) = -(y_i - \beta_i) + \lambda^*_{NI}(q_i(\lambda^*_{NI}) - q_{i-1}(\lambda^*_{NI})) \\ + \lambda_F(t_i(\lambda^*_{NI}) - t_{i-1}(\lambda^*_{NI})) = 0, \end{aligned} \tag{A5}$$

with $i = j, \dots, k$, and show that there exists the solution for which (A4) holds, $q_i \in [0, 1]$ and $t_i \in [-1, 1]$.

Note first that as λ_{NI} increases, (A4) holds until the merge with other groups happens, which means that $q_{j-1}, q_{j+k} \in \{0, 1\}$ and $t_{j-1}, t_{j+k} \in \{-1, 1\}$ will not change their values until the merge of this constant region. Also, as it follows from (A3), for $i \in [j, j+k]$ the value of t_i is in $[-1, 1]$. Therefore, without any violation of the restrictions on t_i we can assume that $t_i(\lambda^*_{NI}) = t_i(\lambda)$ for any $i \in [j, j+k-1]$.

Next, taking pairwise differences between subgradient equations for λ_{NI} we have

$$\lambda_{NI} A\tilde{\boldsymbol{q}}(\lambda_{NI}) + \lambda_F A\tilde{\boldsymbol{t}}(\lambda_{NI}) = D\tilde{\boldsymbol{y}} + \lambda_{NI}\boldsymbol{c}(\lambda_{NI}) + \lambda_F \boldsymbol{d}(\lambda_{NI}),$$

where D is displayed at Figure 1,

$$A = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix}, \tag{A6}$$

and

$$\begin{aligned} \tilde{\boldsymbol{y}} &= (y_j, \dots, y_{j+k}), \\ \tilde{\boldsymbol{q}}(\lambda_{NI}) &= (q_j(\lambda_{NI}), \dots, q_{j+k-1}(\lambda_{NI})), \\ \tilde{\boldsymbol{t}}(\lambda_{NI}) &= (t_j(\lambda_{NI}), \dots, t_{j+k-1}(\lambda_{NI})), \\ \boldsymbol{c}(\lambda_{NI}) &= (q_{j-1}(\lambda_{NI}), 0, \dots, 0, q_{j+k}(\lambda_{NI})), \\ \boldsymbol{d}(\lambda_{NI}) &= (t_{j-1}(\lambda_{NI}), 0, \dots, 0, t_{j+k}(\lambda_{NI})). \end{aligned}$$

Since A is invertible we have

$$\lambda_{NI}\tilde{\boldsymbol{q}}(\lambda_{NI}) + \lambda_F\tilde{\boldsymbol{t}}(\lambda_{NI}) = A^{-1}D\tilde{\boldsymbol{y}} + \lambda_{NI}A^{-1}\boldsymbol{c}(\lambda_{NI}) + \lambda_F A^{-1}\boldsymbol{d}(\lambda_{NI}),$$

and, since $\tilde{\boldsymbol{q}}(\lambda_{NI})$ and $\tilde{\boldsymbol{t}}(\lambda_{NI})$ provide the solution to the subgradient equations (A5), then

$$-\lambda_F \le \lambda_{NI}\tilde{\boldsymbol{q}}(\lambda_{NI}) + \lambda_F\tilde{\boldsymbol{t}}(\lambda_{NI}) \le \lambda_{NI} + \lambda_F. \tag{A7}$$

Next, as pointed out at [16] and [11]

$$(A^{-1})_{i,1} = (n-i+1)/(n+1) \quad \text{and} \quad (A^{-1})_{i,n} = i/(n+1),$$

then, one can show that

$$-\lambda_F \mathbf{1} \preceq \lambda_{NI} A^{-1}\boldsymbol{c}(\lambda_{NI}) + \lambda_F A^{-1}\boldsymbol{d}(\lambda_{NI}) \preceq \lambda_{NI}\mathbf{1} + \lambda_F \mathbf{1}. \tag{A8}$$

Further, let us consider the case of $\lambda_{NI}^* > \lambda_{NI}$. Then we have

$$\lambda_{NI}^*\tilde{\boldsymbol{q}}(\lambda_{NI}^*) + \lambda_F \tilde{\boldsymbol{t}}(\lambda_{NI}^*) = A^{-1}D\tilde{\boldsymbol{y}} + \lambda_{NI}^* A^{-1}\boldsymbol{c}(\lambda_{NI}^*) + \lambda_F A^{-1}\boldsymbol{d}(\lambda_{NI}^*).$$

Recall, above we set $\tilde{\boldsymbol{t}}(\lambda_{NI}^*) = \tilde{\boldsymbol{t}}(\lambda_{NI})$, and $q_{j-1}, q_{j+k}, t_{j-1}$ and t_{j+k} does not change their values until the merge, which means that $\boldsymbol{c}(\lambda_{NI}^*) = \boldsymbol{c}(\lambda_{NI})$, and $\boldsymbol{d}(\lambda_{NI}^*) = \boldsymbol{d}(\lambda_{NI})$.

Therefore, the subgradient equations for λ_{NI}^* can be written as

$$\lambda_{NI}^*\tilde{\boldsymbol{q}}(\lambda_{NI}^*) + \lambda_F \tilde{\boldsymbol{t}}(\lambda_{NI}) = A^{-1}D\tilde{\boldsymbol{y}} + \lambda_{NI}^* A^{-1}\boldsymbol{c}(\lambda_{NI}) + \lambda_F A^{-1}\boldsymbol{d}(\lambda_{NI}).$$

Next, since the term $A^{-1}D\tilde{\boldsymbol{y}}$ is not changed, $-\lambda_F \leq \lambda_F \tilde{\boldsymbol{t}}(\lambda_{NI}) \leq \lambda_F$, and

$$-\lambda_F \mathbf{1} \preceq \lambda_{NI}^* A^{-1}\boldsymbol{c}(\lambda_{NI}) + \lambda_F A^{-1}\boldsymbol{d}(\lambda_{NI}) \preceq \lambda_{NI}^*\mathbf{1} + \lambda_F \mathbf{1},$$

then we have

$$\mathbf{0} \preceq \tilde{\boldsymbol{q}}(\lambda_{NI}^*) \preceq \mathbf{1}.$$

Therefore we proved that $\hat{\beta}_i^{FNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*) = \hat{\beta}_{i+1}^{FNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*)$. Since $\hat{\beta}_i^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*)$ is given by soft thresholding of $\hat{\beta}_i^{FNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*)$, then $\hat{\beta}_i^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*) = \hat{\beta}_{i+1}^{FLNI}(\boldsymbol{y}, \boldsymbol{\lambda}^*)$ for $i \in [j, k]$.

Case 3: λ_{NI} and λ_L are fixed and $\lambda_F^* > \lambda_F$. The proof for this case is virtually identical to the proof for the Case 2. In this case we assume that $q_i(\lambda_F^*) = q_i(\lambda_2)$ for any $i \in [j, j+k-1]$. Next, $q_{j-1}, q_{j+k}, t_{j-1}$ and t_{j+k} do not change their values until the merge, which, again, means that $\boldsymbol{c}(\lambda_F^*) = \boldsymbol{c}(\lambda_F)$, and $\boldsymbol{d}(\lambda_F^*) = \boldsymbol{d}(\lambda_F)$. Finally, we can show that

$$-\mathbf{1} \preceq \tilde{\boldsymbol{t}}(\lambda_F^*) \preceq \mathbf{1}.$$

□

Proof of Theorem 5. For some fixed λ_F and λ_{NI} let us write subgradient equations for the fussed lasso nearly-isotonic approximator:

$$g_i = -(y_i - \beta_i) + \lambda_{NI}(q_i - q_{i-1}) + \lambda_F(t_i - t_{i-1}) = 0,$$

for $i = 1, \ldots, n$, where q_i and t_i, with $i = 1, \ldots, n-1$, are given by

$$q_i : \begin{cases} = 1, & \text{if } \beta_i - \beta_{i+1} > 0, \\ = 0, & \text{if } \beta_i - \beta_{i+1} < 0, \\ \in [0, 1], & \text{if } \beta_i = \beta_{i+1}, \end{cases} \qquad t_i : \begin{cases} = 1, & \text{if } \beta_i - \beta_{i+1} > 0, \\ = -1, & \text{if } \beta_i - \beta_{i+1} < 0, \\ \in [-1, 1], & \text{if } \beta_i = \beta_{i+1}, \end{cases} \tag{A9}$$

and $q_0 = q_n = t_0 = t_n = 0$.

Second, assume that in the solution $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$ there are K distinct constant regions $\mathcal{A}(\lambda_F, \lambda_{NI}) = \{A_1, \dots, A_K\}$, and f_j and l_j are the first and last indices, respectively, in the region A_j. Therefore, using the telescoping sums, for $k \in A_j$ the solution $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$ can be written as

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \frac{\sum_{i=f_j}^{l_j} y_i}{|A_j|} - \lambda_{NI}\frac{q_{f_{j+1}} - q_{l_j}}{|A_j|} - \lambda_F\frac{t_{f_{j+1}} - t_{l_j}}{|A_j|},$$

with $|A_j| = \#\{j : y_j \in A_j\}$.

We, first, prove that

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}).$$

Let us fix some λ_F, and take $\lambda_{NI}^* > \lambda_{NI}$ such that $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^*)$ has the same constant regions as $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$. Therefore, analogously to the case of one dimensional nearly-isotonic regression in [11], for a fixed λ_{NI} the solution is linear function of λ_{NI} in between the values of λ_{NI} (which are called knots) where some constant regions merge.

Assume now that $\lambda_{NI} = 0$. Next, assume that in the solution $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, 0)$ there are $K(0)$ distinct constant regions $\mathcal{A}(\lambda_F, 0) = \{A_1, \dots, A_K\}$, and f_j and l_j are the first and last indices, respectively, in those region A_j.

Next, we increase the value of $\lambda_{NI}^* > \lambda_{NI}$ and assume that we still have the same constant regions as for λ_F and λ_{NI}, i.e.

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^*) = \frac{\sum_{i=f_j}^{l_j} y_i}{|A_j|} - \lambda_{NI}^*\frac{q_{f_{j+1}} - q_{l_j}}{|A_j|} - \lambda_F\frac{t_{f_{j+1}} - t_{l_j}}{|A_j|},$$

i.e. at the value λ_{NI}^* not merge has happened, which means that

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^*) \neq \hat{\beta}_{k'}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^*)$$

if k and k' are not in the same $A_j \in \mathcal{A}(\lambda_F, 0)$. Next, recall that for any $k \in A_j$ we have

$$\hat{\beta}_k^{F}(\boldsymbol{y}, \lambda_F) = \hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, 0) = \frac{\sum_{i=f_j}^{l_j} y_i}{|A_j|} - \lambda_F\frac{t_{f_{j+1}} - t_{l_j}}{|A_j|}. \tag{A10}$$

Therefore, $\hat{\beta}_k^{F}(\boldsymbol{y}, \lambda_F)$ has the same constant regions as $\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, 0)$.

Then, recall that

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{NI}(\hat{\boldsymbol{\beta}}^{F}(\boldsymbol{y}, \lambda_F), \lambda_{NI})$$

Next, let us choose $\lambda'_{NI} < \lambda_{NI}^*$ such that, again, the constant regions of $\hat{\boldsymbol{\beta}}^{NI}(\hat{\beta}_k^{F}(\boldsymbol{y}, \lambda_F), \lambda'_{NI})$ are the same as for $\hat{\beta}_k^{F}(\boldsymbol{y}, \lambda_F)$ and $\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$. Then, for

$k \in A_j$ the solution is given by

$$\hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda'_{NI}) = \frac{\sum_{i=f_j}^{l_j} \hat{\beta}_i^F}{|A_j|} - \lambda'_{NI} \frac{q_{f_{j+1}} - q_{l_j}}{|A_j|},$$

and using (A10) we get

$$\hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda'_{NI}) = \frac{\sum_{i=f_j}^{l_j} y_i}{|A_j|} - \lambda'_{NI} \frac{q_{f_{j+1}} - q_{l_j}}{|A_j|} - \lambda_F \frac{t_{f_{j+1}} - t_{l_j}}{|A_j|},$$

which means that the solution is linear function of λ'_{NI} until some constant regions merge.

Note now

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda'_{NI}) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda'_{NI})$$

and, obviously, this equality holds at least until constant regions merge. Let $\lambda_{NI}^{(1)}$ be the first value of λ_{NI} when the first merge happens. At the value $\lambda_{NI}^{(1)}$ the equality

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)})$$

holds, since $\hat{\boldsymbol{\beta}}^{NI}$ is continuous in λ_{NI}.

Assume for simplicity of notation that at $\lambda_{NI} = \lambda_{NI}^{(1)}$ the constant region A_j merges with constant region A_{j+1}. Therefore, for $k \in A_j \cup A_{j+1}$ we have

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}) = \frac{\sum_{i=f_j}^{l_{j+1}} y_i}{|A_j| + |A_{j+1}|} - \lambda_{NI}^{(1)} \frac{q_{f_{j+2}} - q_{l_j}}{|A_j| + |A_{j+1}|} - \lambda_F \frac{t_{f_{j+2}} - t_{l_j}}{|A_j| + |A_{j+1}|},$$

and for $k \in A_m \neq A_j \cup A_{j+1}$:

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}) = \frac{\sum_{i=f_m}^{l_m} y_i}{|A_m|} - \lambda_{NI}^{(1)} \frac{q_{f_{m+1}} - q_{l_m}}{|A_m|} - \lambda_F \frac{t_{f_{m+1}} - t_{l_m}}{|A_m|}.$$

Further, for $\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)})$ for $k \in A_j \cup A_{j+1}$ we have

$$\hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}) = \frac{\sum_{i=f_j}^{l_{j+1}} \hat{\beta}_i^F}{|A_j| + |A_{j+1}|} - \lambda_{NI}^{(1)} \frac{q_{f_{j+2}} - q_{l_j}}{|A_j| + |A_{j+1}|} = \hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}),$$

and for $k \in A_m \neq A_j \cup A_{j+1}$:

$$\hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}) = \frac{\sum_{i=f_m}^{l_m} \hat{\beta}_i^F}{|A_m|} - \lambda_{NI}^{(1)} \frac{q_{f_{m+1}} - q_{l_m}}{|A_m|} = \hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)}).$$

Next, let us increase $\lambda_{NI}^{(1)}$ by $\delta\lambda$ so that no merge in $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}+\delta\lambda)$ happens. Then, for $k \in A_j \cup A_{j+1}$ we have

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda) = \frac{\sum_{i=f_j}^{l_{j+1}} y_i}{|A_j| + |A_{j+1}|} - (\lambda_{NI}^{(1)} + \delta\lambda)\frac{q_{f_{j+2}} - q_{l_j}}{|A_j| + |A_{j+1}|} - \lambda_F \frac{t_{f_{j+2}} - t_{l_j}}{|A_j| + |A_{j+1}|},$$

and for $k \in A_m \neq A_j \cup A_{j+1}$:

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda) = \frac{\sum_{i=f_m}^{l_m} y_i}{|A_m|} - (\lambda_{NI}^{(1)} + \delta\lambda)\frac{q_{f_{m+1}} - q_{l_m}}{|A_m|} - \lambda_F \frac{t_{f_{m+1}} - t_{l_m}}{|A_m|}.$$

Further, in the case of $\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)})$ we increase λ by $\delta\lambda' < \delta\lambda$ and, therefore, we have for $k \in A_j \cup A_{j+1}$:

$$\hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda') = \frac{\sum_{i=f_j}^{l_{j+1}} \hat{\beta}_i^F}{|A_j| + |A_{j+1}|} - (\lambda_{NI}^{(1)} + \delta\lambda')\frac{q_{f_{j+2}} - q_{l_j}}{|A_j| + |A_{j+1}|},$$

and for $k \in A_m \neq A_j \cup A_{j+1}$:

$$\hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda') = \frac{\sum_{i=f_m}^{l_m} \hat{\beta}_i^F}{|A_m|} - (\lambda_{NI}^{(1)} + \delta\lambda')\frac{q_{f_{m+1}} - q_{l_m}}{|A_m|}.$$

Therefore, before the next merge happens we have the following relation between the estimators $\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$ and $\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda) = \hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda') + (\delta\lambda - \delta\lambda')\frac{q_{f_{j+2}} - q_{l_j}}{|A_j| + |A_{j+1}|},$$

if $k \in A_j \cup A_{j+1}$, and

$$\hat{\beta}_k^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda) = \hat{\beta}_k^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(1)} + \delta\lambda') + (\delta\lambda - \delta\lambda')\frac{q_{f_{m+1}} - q_{l_m}}{|A_m|},$$

for $k \in A_m \neq A_j \cup A_{j+1}$.

We have proved that before the second merge we have

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI})$$

and at the value of $\lambda_{NI}^{(2)}$ when the second merge of some constant regions happens we have

$$\hat{\boldsymbol{\beta}}^{F\to NI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(2)}) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}^{(2)})$$

by the continuity.

We can continue this process until the last knot point in the path. Therefore we proved the equality of the estimators. The proof of

$$\hat{\boldsymbol{\beta}}^{NI\to F}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}).$$

is virtually the same with q_j suitably changed to t_j and λ_{NI} to λ_F and using the propertis of fused lasso from [17]. □

Proof of Theorem 6. First, for the fused estimator $\hat{\boldsymbol{\beta}}^F(\boldsymbol{y}, \lambda_F)$ let

$$K^F(\boldsymbol{y}, \lambda_F) = \#\{\text{fused groups in } \hat{\boldsymbol{\beta}}^F(\boldsymbol{y}, \lambda_F)\}.$$

Then, as it follows from [7], for $\hat{\boldsymbol{\beta}}^F(\boldsymbol{y}, \lambda_F)$ we have

$$\mathbb{E}[K^F(\boldsymbol{Y}, \lambda_F)] = df(\hat{\boldsymbol{\beta}}^F(\boldsymbol{Y}, \lambda_F)).$$

Next, from Proposition 3, it follows

$$\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{y}, \lambda_F, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^F(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F).$$

Therefore, using the property of covariance we have

$$\begin{aligned} df(\hat{\boldsymbol{\beta}}^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI})) = \sum_{i=1}^n \mathrm{Cov}[\hat{\beta}_i^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI}), Y_i] = \\ \sum_{i=1}^n \mathrm{Cov}[\hat{\beta}_i^F(\boldsymbol{Y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F), Y_i] = \\ \sum_{i=1}^n \mathrm{Cov}[\hat{\beta}_i^F(\boldsymbol{Y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F), Y_i - \frac{\lambda_{NI}}{2}[D^T\mathbf{1}]_i] = \\ \mathbb{E}[K^F(\boldsymbol{Y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F)] \equiv \mathbb{E}[K^{FNI}(\boldsymbol{Y}, \lambda_F, \lambda_{NI})], \end{aligned}$$

where $[\boldsymbol{a}]_i$ denotes i-th element in the vector $\boldsymbol{a} \in \mathbb{R}^n$.

Next, we prove the result for the fused lasso nearly-isotonic approximator. From Proposition 3 we have

$$\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{y}, \lambda_F, \lambda_L, \lambda_{NI}) = \hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F, \lambda_L).$$

Next, for the fused lasso $\hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y}, \lambda_F, \lambda_L)$ defined in (2) let

$$K^{FL}(\boldsymbol{y}, \lambda_F, \lambda_L) = \#\{\text{non-zero fused groups in } \hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{y}, \lambda_F, \lambda_L)\},$$

and from [7] it follows

$$\mathbb{E}[K^{FL}(\boldsymbol{Y}, \lambda_F, \lambda_L)] = df(\hat{\boldsymbol{\beta}}^{FL}(\boldsymbol{Y}, \lambda_F, \lambda_L)).$$

Further, again, using the property of the covariance, we have

$$\begin{aligned}
df(\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI})) &= \sum_{i=1}^{n} \mathrm{Cov}[\hat{\beta}_i^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI}), Y_i] \\
&= \sum_{i=1}^{n} \mathrm{Cov}[\hat{\beta}_i^{FL}(\boldsymbol{Y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F, \lambda_L), Y_i] \\
&= \sum_{i=1}^{n} \mathrm{Cov}[\hat{\beta}_i^{FL}(\boldsymbol{Y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F, \lambda_L), Y_i - \frac{\lambda_{NI}}{2}[D^T \mathbf{1}]_i] \\
&= \mathbb{E}[K^{FL}(\boldsymbol{Y} - \frac{\lambda_{NI}}{2} D^T \mathbf{1}, \frac{1}{2}\lambda_{NI} + \lambda_F, \lambda_L)] \\
&\equiv \mathbb{E}[K^{FLNI}(\boldsymbol{Y}, \lambda_F, \lambda_L, \lambda_{NI})].
\end{aligned}$$

Lastly, we note that the proof for the unbiased estimator of the degrees of freedom for nearly-isotonic regression, given in [11], can be done in the same way as in the current proof, using the relation (19) and, again, the result of the paper [7] for the fusion estimator $\hat{\boldsymbol{\beta}}^{FLNI}(\boldsymbol{Y}, \lambda_F)$. □

References

[1] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)

[2] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

[3] Phillips, D.L.: A technique for the numerical solution of certain integral equations of the first kind. Journal of the ACM (JACM) **9**(1), 84–97 (1962)

[4] Tikhonov, A.N., Goncharsky, A., Stepanov, V., Yagola, A.G.: Numerical Methods for the Solution of Ill-posed Problems vol. 328. Springer, ??? (1995)

[5] Rinaldo, A.: Properties and refinements of the fused lasso. The Annals of Statistics **37**(5B), 2922–2952 (2009)

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

[7] Tibshirani, R.J., Taylor, J.: The solution path of the generalized lasso. The Annals of Statistics **39**(3), 1335–1371 (2011)

[8] Best, M.J., Chakravarti, N.: Active set algorithms for isotonic regression; a unifying framework. Mathematical Programming **47**(1), 425–439 (1990)

[9] Stout, Q.F.: Isotonic regression via partitioning. Algorithmica **66**(1), 93–112 (2013)

[10] Robertson, T., Wright, F.T., Dykstra, R.L.: Order Restricted Statistical Inference. Wiley, ??? (1988)

[11] Tibshirani, R.J., Hoefling, H., Tibshirani, R.: Nearly-isotonic regression. Technometrics **53**(1), 54–61 (2011)

[12] Minami, K.: Estimating piecewise monotone signals. Electronic Journal of Statistics **14**(1), 1508–1576 (2020)

[13] Yu, Z., Chen, X., Li, X.: A dynamic programming approach for generalized nearly isotonic optimization. Mathematical Programming Computation, 1–31 (2022)

[14] Gómez, A., He, Z., Pang, J.-S.: Linear-step solvability of some folded concave and singly-parametric sparse optimization problems. Mathematical Programming, 1–42 (2022)

[15] Gaines, B.R., Kim, J., Zhou, H.: Algorithms for fitting the constrained lasso. Journal of Computational and Graphical Statistics **27**(4), 861–871 (2018)

[16] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. The Annals of Applied Statistics **1**(2), 302–332 (2007)

[17] Hoefling, H.: A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics **19**(4), 984–1006 (2010)

[18] Beran, R., Dümbgen, L.: Least squares and shrinkage estimation under bimonotonicity constraints. Statistics and computing **20**(2), 177–189 (2010)

[19] Deng, H., Zhang, C.-H.: Isotonic regression in multi-dimensional spaces and graphs. The Annals of Statistics **48**(6), 3672–3698 (2020)

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

[21] Han, Q., Zhang, C.-H.: Limit distribution theory for block estimators in multiple isotonic regression. The Annals of Statistics **48**(6), 3251–3282 (2020)

[22] Wang, Y.-X., Sharpnack, J., Smola, A., Tibshirani, R.: Trend filtering on graphs. In: Artificial Intelligence and Statistics, pp. 1042–1050 (2015). PMLR

[23] Gao, C., Han, F., Zhang, C.-H.: On estimation of isotonic piecewise constant signals. The Annals of Statistics **48**(2), 629–654 (2020)

[24] Efron, B.: How biased is the apparent error rate of a prediction rule? Journal of the American statistical Association **81**(394), 461–470 (1986)

[25] Meyer, M., Woodroofe, M.: On the degrees of freedom in shape-restricted regression. The annals of Statistics **28**(4), 1083–1104 (2000)

[26] Bento, J., Furmaniak, R., Ray, S.: On the complexity of the weighted fused lasso. IEEE Signal Processing Letters **25**(10), 1595–1599 (2018)

[27] Stellato, B., Banjac, G., Goulart, P., Bemporad, A., Boyd, S.: Osqp: An operator splitting solver for quadratic programs. Mathematical Programming Computation **12**(4), 637–672 (2020)

[28] Becker, B., Kohavi, R.: UCI Machine Learning Repository (1996). http://archive.ics.uci.edu/ml

[29] Wang, X., Ying, J., Cardoso, J.V.d.M., Palomar, D.P.: Efficient algorithms for general isotone optimization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 8575–8583 (2022)

[30] Kim, S.-J., Koh, K., Boyd, S., Gorinevsky, D.: ℓ_1 trend filtering. SIAM Review **51**(2), 339–360 (2009)