A sharp oracle inequality for Graph-Slope

Pierre C. Bellec, Joseph Salmon, Samuel Vaiter

A sharp oracle inequality for Graph-Slope

Pierre C Bellec

Department of Statistics & Biostatistics,
Rutgers, The State University of New Jersey,
501 Hill Center, Busch Campus,
110 Frelinghuysen Road, Piscataway, NJ 08854,
e-mail: pcb71@stat.rutgers.edu
url: http://www.stat.rutgers.edu/home/PCB71/

and

Joseph Salmon*

LTCI, CNRS, Télécom ParisTech,
Université Paris-Saclay, 75013, Paris,
e-mail: first.lastname@telecom-paristech.fr
url: http://josephsalmon.eu

and

Samuel Vaiter*

CNRS & IMB, Université de Bourgogne,
9 avenue Alain Savary, 21000, Dijon,
e-mail: samuel.vaiter@u-bourgogne.fr
url: http://samuelvaiter.com

Abstract:
Following recent success on the analysis of the Slope estimator, we provide a sharp oracle inequality in term of prediction error for Graph-Slope, a generalization of Slope to signals observed over a graph. In addition to improving upon best results obtained so far for the Total Variation denoiser (also referred to as Graph-Lasso or Generalized Lasso), we propose an efficient algorithm to compute Graph-Slope. The proposed algorithm is obtained by applying the forward-backward method to the dual formulation of the Graph-Slope optimization problem. We also provide experiments showing the interest of the method.

MSC 2010 subject classifications: Primary 62G08; Secondary 62J07.
Keywords and phrases: denoising, graph signal regularization, oracle inequality, convex optimization.

Contents

*J. Salmon and S. Vaiter were supported by the CNRS PE1 "ABS" grant.

1. Introduction

Many inference problems of interest involve signals defined on discrete graphs. This includes for instance two-dimensional imaging but also more advanced hyper-spectral imaging scenarios where the signal lives on a regular grid. Two types of structure arise naturally in such examples: The first type of structures comes from regularity or smoothness of the signal, which led to the development of wavelet methods. The second type of structure involves signals with few sharp discontinuities. For instance in one dimension, piecewise constant signals appear when transition states are present, the graph being a 1D path. In imaging, where the underlying graph is a regular 2D grid, occlusions create piece-wise smooth signals rather that smooth ones.

This paper studies regularizers for signals with sharp discontinuities. A popular choice in imaging is the Total Variation (TV) regularization [24]. For 1D signals, TV regularization has also long been used in statistics [18]. If an additional ℓ_1 regularization is added, this is sometimes referred to as the fused Lasso [30, 32, 13].

A natural extension of such methods to arbitrary graphs relies on ℓ_1 analysis penalties [15] which involve the incidence matrix of the underlying graph, see for instance [25] or the Edge Lasso of [26]. Such penalties have the form

$$\text{pen} \ : \beta \to \lambda \|\mathbf{D}^\top \beta\|_1 \ ,$$

where $\lambda > 0$ is a tuning parameter and $\mathbf{D}^\top$ is the (edge-vertex) incidence matrix of the graph defined below, and β represents the signal to be recovered. This approach is notably different from contributions in machine learning where ℓ_2 penalties, *i.e.,* Laplacian regularization, have been considered for spectral clustering [27, 20] (see also [33] for a review). Theoretical results in favor of the ℓ_1 norm instead of the squared ℓ_2 norm are studied in [25].

Penalties based on ℓ_0 regularization with the graph incidence matrix have recently been analyzed [16]. They are of interest as they do not suffer from the (shrinkage) bias created by the convex ℓ_1 norm. However, such methods present the difficulty that in the general case they lead to non-convex problems. Note that the 1D path is an exception since the associated optimization problem can be solve using dynamic programming [1]. Concerning the bias reduction though, simpler remedies could be used, including least-squares refitting on the model space associated, applying for instance the CLEAR method [14].

Following the introduction of the Slope regularization in the context of high dimensional regression [10], we propose Graph-Slope, its generalization to contexts where the signal is supported on a graph. In linear regression, Slope [10] is defined as follows. Given p tuning parameters $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ with at least one strict inequality, define the ordered ℓ_1 norm by

$$\|\theta\|_{[\lambda]} = \sum_{j=1}^{p} \lambda_j |\theta|_j^{\downarrow} \ , \tag{1.1}$$

where for any $\theta \in \mathbb{R}^p$, we use the notation[1] $(|\theta|_1^{\downarrow}, \ldots, |\theta|_p^{\downarrow})$ for the non-increasing rearrangement of its amplitudes $(|\theta_1|, \ldots, |\theta_p|)$. Then, given a design matrix $X \in \mathbb{R}^{n \times p}$ and a response vector $y \in \mathbb{R}^n$, the Slope estimator is defined as a solution of the minimization problem

$$\min_{b \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|^2 + \|\beta\|_{[\lambda]} \ .$$

If the tuning parameters $\lambda_1, \ldots, \lambda_p$ are all equal, then Slope is equal to the Lasso with tuning parameter λ_1.

Slope presents several advantages compared to Lasso in sparse linear regression. First, Slope provably controls the False Discovery Rate (FDR) for orthogonal design matrices [10] and experiments show that this property is also satisfied for some non-orthogonal design matrices [10]. Second, it appears that Slope has more power than Lasso in the sense that Slope will discover more nonzero coefficients of the unknown target vector [10]. An interpretation of this phenomenon is that Lasso shrinks too heavily small coefficients and may thus miss the smallest nonzero coefficients of the target vector. On the other hand, the Slope penalty induces less shrinkage on small coefficients, leading to more power. Third, while Lasso with the universal parameter is known to achieve the rate of estimation of order $(s/n)\log(p)$ (where s is the sparsity of the unknown target vector, n the number of measurements and p the number of covariates), Slope achieves the optimal rate of estimation of order $(s/n)\log(p/s)$ [28, 4]. However, Slope requires p tuning parameters while Lasso only requires one.

We propose a theoretical and experimental analysis of Graph-Slope, the counterpart estimator of Slope for signals defined on graphs. Graph-Slope is defined in the next section. Our theoretical contribution for Graph-Slope borrows some technical details recently introduced in [17] to control the Mean Square Error (MSE) for the Generalized Lasso.

Last but not least, we provide an efficient solver to compute the Graph-Slope estimator. It relies on accelerated proximal gradient descent to solve the dual formulation [3, 12, 22]. To obtain an efficient solver, we leverage the seminal contribution made in [36] showing the link between ordered ℓ_1 norm (1.1) and isotonic regression. Hence, we can use fast implementations of the PAVA algorithm (for Pool Adjacent Violators Algorithm, see for instance [6]), available for

[1] following the notation considered in [7]

instance in `scikit-learn` [23] for this purpose. Numerical experiments illustrate the benefit of Graph-Slope, in particular in terms of True Discovery Rate (TDR) performance.

A high level interpretation of our simulation results is as follows. In the model considered in this paper, a sharp discontinuity of the signal corresponds to an edge of the graph with nonzero coefficient. Since Graph-Lasso uses an ℓ_1-penalty, the penalty level is uniform across all edges of the graph. Edges with small coefficients are too heavily penalized with Graph-Lasso. Using Graph-Slope lets us reduce the penalty level on the edges with small coefficients. This leads to the discovery of more discontinuities of the true signal as compared to Graph-Lasso.

1.1. Model and notation

Let $\mathcal{G} = (V, E)$ be an undirected and connected graph on n vertices, $V = [n]$, and p edges, $E = [p]$. This graph can be represented by its edge-vertex incidence matrix $\mathbf{D}^\top = \mathbf{D}_\mathcal{G}^\top \in \mathbb{R}^{n\times p}$ (we drop the reference to $\mathcal{G}$ when no ambiguity is possible) defined as

$$(\mathbf{D}^\top)_{e,v} = \begin{cases} +1, & \text{if } v = \min(i,j) \\ -1, & \text{if } v = \max(i,j) \\ 0, & \text{otherwise} \end{cases} ,$$

where $e = \{i, j\}$. The matrix $\mathbf{L} = \mathbf{D}\mathbf{D}^\top$ is the so-called graph Laplacian of $\mathcal{G}$. The Laplacian $\mathbf{L}$ is invariant under a change of orientation of the graph.

For any $u \in \mathbb{R}^p$, we denote by $\|u\|_0$ the pseudo ℓ_0 norm of u : $\|u\|_0 = |\{j \in [p] : u_j \neq 0\}|$, and for any matrix $\mathbf{A}$, we denote by $\mathbf{A}^\dagger$ its Moore-Penrose pseudo-inverse. The canonical basis of $\mathbb{R}^p$ is denoted $(e_1, \dots, e_p)$.

For any norm $\|\cdot\|$ on $\mathbb{R}^n$, the associated *dual norm* $\|\cdot\|^*$ reads at $v \in \mathbb{R}^n$

$$\|v\|^* = \sup_{\|\beta\|\leqslant 1} \langle v, \beta\rangle \ .$$

As a consequence, for every $(\beta, v) \in \mathbb{R}^n \times \mathbb{R}^n$, one have $\langle \beta, v\rangle \leqslant \|\beta\|\|v\|^*$.

In this work, we consider the following denoising problem for a signal over a graph. Assume that each vertex $i \in [n]$ of the graph carries a signal $\beta_i^\star$. For each vertex $i \in [n]$ of the graph, one observes y_i, a noisy perturbation of $\beta_i^\star$. In vector form, one observes the vector $y \in \mathbb{R}^n$ and aims to estimate $\beta^\star \in \mathbb{R}^n$, *i.e.,*

$$y = \beta^\star + \varepsilon \ ,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$ is a noise vector. We will say that an edge $e = \{i, j\}$ of the graph carries the signal $(\mathbf{D}^\top\beta^\star)_e$. In particular, if two vertices i and j are neighbours and if they carry the same value of the signal, *i.e.,* $\beta_i^\star = \beta_j^\star$, then the corresponding edge $e = \{i, j\}$ carries the constant signal. The focus of the present paper is on signals $\beta^\star$ that have few discontinuities. A signal $\beta^\star \in \mathbb{R}^n$ has few discontinuities if $\mathbf{D}^\top\beta^\star$ has few nonzero coefficients, *i.e.,* $\|\mathbf{D}^\top\beta^\star\|_0$ is small, or equivalently if most edges of the graph carry the constant signal. In particular, if $\|\mathbf{D}^\top\beta^\star\|_0 = s$, we say that $\beta^\star$ is a vector of $\mathbf{D}^\top$-sparsity s.

1.2. The Graph-Slope estimator

We consider in this paper the so-called *Graph-Slope* variational scheme:

$$\hat{\beta} := \hat{\beta}^{\mathrm{GS}} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - \beta\|_n^2 + \|\mathbf{D}^\top \beta\|_{[\lambda]} \ , \tag{1.2}$$

where $\|\cdot\|_n^2 = \frac{1}{n}\|\cdot\|^2$ is the scaled Euclidean norm, and

$$\|\mathbf{D}^\top \beta\|_{[\lambda]} = \sum_{j=1}^p \lambda_j |\mathbf{D}^\top \beta|_j^\downarrow \ ,$$

with $\lambda = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$ satisfying $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geq 0$, and using for any vector $\theta \in \mathbb{R}^p$ the notation[2] $(|\theta|_1^\downarrow, \dots, |\theta|_p^\downarrow)$ for the non-increasing rearrangement of its amplitudes $(|\theta_1|, \dots, |\theta_p|)$. According to [10], $\|\cdot\|_{[\lambda]}$ is a norm over $\mathbb{R}^p$ if and only if $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geq 0$ with at least one strict inequality. This is a consequence of the observation that if $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geq 0$ then one can rewrite the Slope-norm of θ as the maximum over all $\tau \in \mathfrak{S}_p$ (the set of permutations over $[p]$), of the quantity $\sum_{i=1}^p \lambda_j |\theta_{\tau(j)}|$:

$$\|\theta\|_{[\lambda]} = \max_{\tau \in \mathfrak{S}_p} \sum_{j=1}^p \lambda_j |\theta_{\tau(j)}| = \sum_{j=1}^p \lambda_j |\theta|_j^\downarrow \ .$$

The Generalized Lasso (also sometimes referred to as TV denoiser) relies on ℓ_1 regularization. It was recently investigated in [17, 25], and can be defined as

$$\hat{\beta}^{\mathrm{GL}} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - \beta\|_n^2 + \lambda_1 \|\mathbf{D}^\top \beta\|_1 \ , \tag{1.3}$$

where $\|\cdot\|_1$ is the standard ℓ_1 norm, and $\lambda_1 > 0$ is a tuning parameter.

If $\lambda_1 = \lambda_2 = \cdots = \lambda_p$ then $\|\theta\|_{[\lambda]} = \lambda_1 \|\theta\|_1$ for all $\theta \in \mathbb{R}^p$, so that the minimization problems (1.3) and (1.2) are the same. On the other hand, if $\lambda_j > \lambda_{j+1}$ for some $j = 1, \dots, p-1$, then the optimization problems (1.3) and (1.2) differ. For instance, if $\lambda_1 > \lambda_2 > 0$, all coefficients of $\mathbf{D}^\top \beta$ are equally penalized in the Graph-Lasso (1.3), while coefficients of $\mathbf{D}^\top \beta$ are not uniformly penalized in in the Graph-Slope optimization problem (1.2). Indeed, in the Graph-Slope optimization problem (1.2), the largest coefficient of $\mathbf{D}^\top \beta$ is penalized as in (1.3) but smaller coefficients of $\mathbf{D}^\top \beta$ receive a smaller penalization. The Graph-Slope optimization problem (1.2) is more flexible than (1.3) as it allows the smaller coefficients of $\mathbf{D}^\top \beta$ to be less penalized than its larger coefficients. We will see in the next sections that this flexibility brings advantages to both the theoretical properties of $\hat{\beta}^{\mathrm{GS}}$ as well as its performance in simulations, as compared to $\hat{\beta}^{\mathrm{GL}}$.

[2] following the notation from [7].

2. Theoretical guarantees: sharp oracle inequality

We can now state the main theoretical result of the paper, a sharp oracle inequality for the Graph-Slope. For any integer s and weights $\lambda = (\lambda_1, \dots, \lambda_p)$, define

$$\Lambda(\lambda, s) = \Big(\sum_{j=1}^{s} \lambda_j^2 \Big)^{1/2}. \tag{2.1}$$

Theorem 2.1. *Assume that the Graph-Slope weights $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are such that the event*

$$\frac{1}{\sqrt{n}} \|\mathbf{D}^\dagger \varepsilon\|_{[\lambda]}^* \leq 1/2 \tag{2.2}$$

has probability at least $1/2$. Then, for any $\delta \in (0,1)$, we have with probability at least $1 - 2\delta$

$$\|\hat\beta - \beta^\star\|_n^2 \leq \inf_{s \in [p]} \left[\inf_{\substack{\beta \in \mathbb{R}^n \\ \|\mathbf{D}^\top \beta\|_0 \leq s}} \|\beta - \beta^\star\|_n^2 + \left(\frac{3\Lambda(\lambda, s)}{\kappa(s)} + \frac{\sigma + 2\sigma\sqrt{2\log(1/\delta)}}{\sqrt{n}} \right)^2 \right], \tag{2.3}$$

where $\Lambda(\cdot,\cdot)$ is defined in (2.1) and the compatibility factor $\kappa(s)$ is defined as

$$\kappa(s) \triangleq \inf_{v \in \mathbb{R}^n : 3\Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 > \sum_{j=s+1}^p \lambda_j |\mathbf{D}^\top v|_j^\downarrow} \left(\frac{\|v\|_n}{\|\mathbf{D}^\top v\|_2} \right). \tag{2.4}$$

Proof. Let β be a minimizer of the right hand side of (2.3) and let $s = \|\mathbf{D}^\top \beta\|_0$. Define the function $f(\cdot)$ by

$$f(e) = \sup_{v \in \mathbb{R}^n : \|v\|_n = 1} \left[\frac{e^\top v}{\sqrt{n}} + \Lambda(\lambda, s) \|\mathbf{D}^\top v\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^\downarrow \right].$$

Let also $w = (\hat\beta - \beta)/(\|\hat\beta - \beta\|_n)$. By Lemma A.3 and Lemma A.1 with $\alpha = 0$, for any $z \in \mathbb{R}^p$, we have almost surely

$$\begin{aligned}
\|\hat\beta - z\|_n^2 - \|\beta - z\|_n^2 \leq & \frac{2}{n} \varepsilon^\top (\hat\beta - \beta) + \|\mathbf{D}^\top \beta\|_{[\lambda]} - \|\mathbf{D}^\top \hat\beta\|_{[\lambda]} - \|\hat\beta - \beta\|_n^2, \\
\leq & \frac{2}{n} \varepsilon^\top (\hat\beta - \beta) + \Lambda(\lambda, s) \|\mathbf{D}^\top (\hat\beta - \beta)\|_2 \\
& - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top (\hat\beta - \beta)|_j^\downarrow - \|\hat\beta - \beta\|_n^2, \\
= & 2\|\hat\beta - \hat\beta\|_n \Big(\frac{\varepsilon^\top w}{\sqrt{n}} + \Lambda(\lambda, s) \|\mathbf{D}^\top w\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top w|_j^\downarrow \Big) \\
& - \|\hat\beta - \beta\|_n^2, \\
\leq & 2\|\hat\beta - \hat\beta\|_n f(\varepsilon) - \|\hat\beta - \beta\|_n^2 \leq f(\varepsilon)^2,
\end{aligned}$$

where for the last inequality we used the elementary inequality $2ab - a^2 \le b^2$.

By Lemma A.4 we have $\mathrm{Id}_n = (\mathrm{Id}_n - \Pi) + \Pi$ with $\mathrm{Id}_n - \Pi = (\mathbf{D}^\top)^\dagger \mathbf{D}^\top$ and where Π is the orthogonal projection onto $\ker(\mathbf{D}^\top)$. Furthermore, $\ker(\mathbf{D}^\top)$ has dimension 1 so that $\|\Pi\varepsilon\|_2^2/\sigma^2$ is a χ^2 random variable with 1 degree of freedom. Thus for any $v \in \mathbb{R}^n$ with $\|v\|_n = 1$ we have

$$\frac{1}{\sqrt{n}}\varepsilon^\top v = \frac{1}{\sqrt{n}}\varepsilon^\top \Pi v + \frac{1}{\sqrt{n}}\varepsilon^\top(\mathrm{Id}_n - \Pi)v \le \|\Pi\varepsilon\|_2 + \frac{1}{\sqrt{n}}\varepsilon^\top(\mathrm{Id}_n - \Pi)v. \quad (2.5)$$

Let us define the function $g(\cdot)$ by

$$g(e) = \sup_{v\in\mathbb{R}^n : \|v\|_n = 1} \left[\frac{1}{\sqrt{n}} e^\top (\mathrm{Id}_n - \Pi)v + \Lambda(\lambda, s)\|\mathbf{D}^\top v\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^\downarrow \right].$$

Then, by the definition of f, g and (2.5), we have almost surely $f(\varepsilon) \le \|\Pi\varepsilon\|_2 + g(\varepsilon)$. By a standard bound on χ^2 random variable with 1 degree of freedom, we have $\mathbb{P}(\|\Pi\varepsilon\|_2 \le \sigma + \sigma\sqrt{2\log(1/\delta)}) \ge 1 - \delta$. Furthermore, the function g is is 1-Lipschitz and $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id}_n)$. By the Gaussian concentration theorem [11, Theorem 10.17], we have

$$\mathbb{P}\left(g(\varepsilon) \le \mathrm{Med}[g(\varepsilon)] + \sigma\sqrt{2\log(1/\delta)} \right) \ge 1 - \delta \ ,$$

where $\mathrm{Med}[g(\varepsilon)]$ is the median of the random variable $g(\varepsilon)$. Combining these two probability bounds with the union bound, we obtain $f(\varepsilon) \le \mathrm{Med}[g(\varepsilon)] + \sigma + 2\sigma\sqrt{2\log(1/\delta)}$ with probability at least $1 - 2\delta$.

To complete the proof, it remains to show that

$$\mathrm{Med}[g(\varepsilon)] \le 3\Lambda(\lambda, s)/\kappa(s) \ .$$

By cdefinition of the median, it is enough to show that $\mathbb{P}(g(\varepsilon) \le 3\Lambda(\lambda, s)/\kappa(s)) \ge 1/2$. By Lemma A.4 and the fact that $\mathrm{Id}_n - \Pi = (\mathbf{D}^\top)^\dagger \mathbf{D}^\top$, we obtain that for all v,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\varepsilon^\top(\mathrm{Id}_n - \Pi)v &= \frac{1}{\sqrt{n}}\varepsilon^\top(\mathbf{D}^\top)^\dagger \mathbf{D}^\top v \\
&\le \frac{1}{\sqrt{n}}\|((\mathbf{D}^\top)^\dagger)^\top \varepsilon\|_{[\lambda]}^* \|\mathbf{D}^\top v\|_{[\lambda]} \\
&= \frac{1}{\sqrt{n}}\|\mathbf{D}^\dagger \varepsilon\|_{[\lambda]}^* \|\mathbf{D}^\top v\|_{[\lambda]} \ ,
\end{aligned}$$

where we used the duality between $\|\cdot\|_{[\lambda]}^*$ and $\|\cdot\|_{[\lambda]}$ for the second term and the fact that the transpose and the Moore-Penrose pseudo-inverse commute, which implies $(\mathbf{D}^\dagger)^\top = \mathbf{D}^{\top\dagger}$.

We now bound $g(\varepsilon)$ from above on the event (2.2). On the event (2.2),

$$\begin{aligned} g(\varepsilon) &\le \sup_{v\in\mathbb{R}^n:\|v\|_n=1} \left[\frac{1}{2}\|\mathbf{D}^\top v\|_{[\lambda]} + \Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow} \right] \\ &\le \sup_{v\in\mathbb{R}^n:\|v\|_n=1} \left[\frac{1}{2}\sum_{j=1}^{s} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow} + \Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 - \frac{1}{2}\sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow} \right] \\ &\le \frac{1}{2} \sup_{v\in\mathbb{R}^n:\|v\|_n=1} \left[3\Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow} \right] . \end{aligned}$$

Consider $v \in \mathbb{R}^n$ such that $\|v\|_n = 1$ and $3\Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 > \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow}$. Then, by the definition of $\kappa(s)$ given in defined in (2.4) we have

$$3\Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow} \le 3\Lambda(\lambda,s)\|v\|_n/\kappa(s) = 3\Lambda(\lambda,s)/\kappa(s).$$

Consider $v \in \mathbb{R}^n$ such that $\|v\|_n = 1$ and $3\Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 \le \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow}$, then we have trivially

$$3\Lambda(\lambda,s)\|\mathbf{D}^\top v\|_2 - \sum_{j=s+1}^{p} \lambda_j |\mathbf{D}^\top v|_j^{\downarrow} \le 0 \le 3\Lambda(\lambda,s)/\kappa(s) \ .$$

Thus, we have proved that on the event (2.2) that has probability at least $1/2$, we have $g(\varepsilon) \le 3\Lambda(\lambda,s)/\kappa(s)$. This implies that $\mathrm{Med}[g(\varepsilon)] \le 3\Lambda(\lambda,s)/\kappa(s)$ and the proof is complete. □

The constant $\kappa(s)$ is sometimes referred to as the compatibility factor of $\mathbf{D}^\top$. Bounds on the compatibility factor are obtained for a large class of random and deterministic graphs [17]. For instance, for graphs with bounded degree, the compatibility factor is bounded from below (see for instance [17, Lemma 3]). In linear regression, constants that measure the correlations of the design matrix have been proposed to study the Lasso and the Dantzig selector: [8] defined the Restricted Eigenvalue constant, [31] defined the Compatibility constant, [35] defined the Cone Invertibility factors and [13] defined the Compatibility factor, to name a few. The Weighted Restricted Eigenvalue constant was also defined in [4] to study the Slope estimator. These constants are the linear regression analogs of $\kappa(s)$ defined in (2.4).

Theorem 2.1 does not provide an explicit choice for the weights $\lambda_1 \ge \cdots \ge \lambda_p$. These weights should be large enough so that the event (2.2) has probability at least $1/2$. These weights should also be as small as possible in order to minimize the right hand side of (2.3). Define $g_1, \ldots, g_p$ by

$$g_j = e_j^\top \mathbf{D}^\dagger \varepsilon/\sqrt{n}, \qquad \text{for all } j = 1, \ldots, p \ . \tag{2.6}$$

and let $|g|_1^\downarrow \geq \cdots \geq |g|_p^\downarrow$ be a nondecreasing rearrangement of $(|g_1|, \ldots, |g_p|)$. Inequality (2.8) in the appendix reads

$$(1/\sqrt{n})\|\mathbf{D}^\dagger \varepsilon\|_{[\lambda]}^* \leq \max_{j=1,\ldots,p} \left(|g|_j^\downarrow \,/\, \lambda_j \right) \ .$$

Thus, if the event

$$\max_{j=1,\ldots,p} \left(|g|_j^\downarrow \,/\, \lambda_j \right) \leq 1/2 \ ,$$

has probability greater than $1/2$, then the event (2.2) has probability greater than $1/2$ as well, and the conclusion of Theorem 2.1 holds. This observation can be used to the following heuristics for the choice of the tuning parameters $\lambda_1 \geq \cdots \geq \lambda_p$. This heuristics can be implemented provided that the Moore-Penrose pseudo-inverse $\mathbf{D}^\dagger$ and the probability distribution of the noise random vector ε are both known. This heuristics goes as follows. Assume that one has generated many independent copies of the random vector ε, and denote by $\tilde{\mathbb{P}}$ the empirical probability distribution with respect to these independent copies of ε. Next, define λ_j as the $(1 - 1/(3p))$th quantile of $2|g|_j^\downarrow$, so that

$$\tilde{\mathbb{P}}(2|g|_j^\downarrow \leq \lambda_j) \geq 1 - 1/3p \ .$$

By the union bound over $j = 1, \ldots, p$,

$$\tilde{\mathbb{P}} \left[\max_{j=1,\ldots,p} \left(|g|_j^\downarrow \,/\, \lambda_j \right) \leq 1/2 \right] \geq 2/3 \ .$$

Now assume that many independent copies of ε have been generated and $\tilde{\mathbb{P}}$ is close to the probability distribution $\mathbb{P}$ of ε in total variation, say $\|\mathbb{P} - \tilde{\mathbb{P}}\|_{TV} \leq \gamma$ for some small constant $\gamma \in (0, 1/6)$. Then the event (2.2) has probability probability greater than $2/3 - \gamma$ with respect to the probability distribution $\mathbb{P}$ of ε, and the conclusion of Theorem 2.1 holds. This simple scheme provides a computational heuristics to choose the weights $\lambda_1, \ldots, \lambda_p$.

The following corollaries propose a theoretical choice for the weights. To state these corollaries, let us write

$$\rho(\mathcal{G}) = \max_{j \in [p]} \|(\mathbf{D}^\top)^\dagger e_j\|_n \ ,$$

following the notation in [17].

Corollary 1. *Assume that the Graph-Slope weights $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ satisfy for any $j \in [p]$*

$$\lambda_j \geq 8\sigma\rho(\mathcal{G}) \sqrt{\frac{\log(2p/j)}{n}} \ . \tag{2.7}$$

Then, for any $\delta \in (0, 1)$, the oracle inequality (2.3) holds with probability at least $1 - 2\delta$.

Proof. It is enough to show that if the weights $\lambda_1, \dots, \lambda_p$ satisfy (2.7) then the event (2.2) has probability at least $1/2$.

Define $g_1, \dots, g_p$ by (2.6) and let $|g|_1^\downarrow \geq \cdots \geq |g|_p^\downarrow$ be a nondecreasing rearrangement of $(|g_1|, \dots, |g_p|)$. For each j, the random variable g_j is a centered Gaussian random variable with variance at most $\|\mathbf{D}^\dagger e_j\|_n^2$. By definition of the dual norm, and [7, Corrolary II.4.3], we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|\mathbf{D}^\dagger \varepsilon\|_{[\lambda]}^* &= \sup_{a\in\mathbb{R}^p:\|a\|_{[\lambda]}=1} \frac{a^\top \mathbf{D}^\dagger \varepsilon}{\sqrt{n}} \\
&\leq \sup_{a\in\mathbb{R}^p:\|a\|_{[\lambda]}=1} \sum_{j=1}^p \lambda_j |a|_j^\downarrow \cdot \frac{|g|_j^\downarrow}{\lambda_j} \\
&\leq \max_{j=1,\dots,p} \frac{|g|_j^\downarrow}{\lambda_j} \qquad (2.8) \\
&\leq \max_{j=1,\dots,p} \frac{|g|_j^\downarrow}{8\sqrt{V\log(2p/j)}} \ ,
\end{aligned}$$

where $V = \sigma^2 \max_{j=1,\dots,p} \|\mathbf{D}^{\top\dagger} e_j\|_n^2$. Thus, by Lemma A.5 below, the event (2.2) has probability at least $1/2$. □

Under an explicit choice of tuning parameters, Corollary 1 yields the following result.

Corollary 2. *Under the same hypothesis as Theorem 2.1 but with the special choice* $\lambda_j = 8\sigma\rho(\mathcal{G})\sqrt{\frac{\log(2p/j)}{n}}$ *for any* $j \in [p]$*, then for any* $\delta \in (0,1)$*, we have with probability at least* $1-2\delta$

$$\|\hat\beta - \beta^\star\|_n^2 \leq \inf_{\substack{s\in[p],\beta\in\mathbb{R}^n \\ \|\mathbf{D}^\top\beta\|_0\leq s}} \left[\|\beta-\beta^\star\|_n^2 + \tfrac{\sigma^2}{n}\frac{48\rho^2(\mathcal{G})s}{\kappa^2(s)}\log\left(\tfrac{2ep}{s}\right)\right] + \tfrac{\sigma^2}{n}(2+16\log\left(\tfrac{1}{\delta}\right)) \ .$$

Proof. We apply Lemma A.2 with the choice $C = 8\sigma\rho(\mathcal{G})$ □

When the true signal satisfies $\|\mathbf{D}^\top\beta^\star\|_0 = s^\star$, the previous bound reduces to

$$\|\hat\beta - \beta^\star\|_n^2 \leq \frac{\sigma^2}{n}\left(\frac{48\rho(\mathcal{G})^2 s^\star}{\kappa(s^\star)^2}\log\left(\frac{2ep}{s^\star}\right) + 2 + 16\log\left(\frac{1}{\delta}\right)\right) \ .$$

Corollary 2 is an improvement w.r.t. the bound provided in [17, Theorem 2] for the TV denoiser (also sometimes referred to as the Generalized Lasso) relying on ℓ_1 regularization defined in Eq. (1.3).

Indeed, the contribution of the second term in Corollary 2 is reduced from $\log(ep/\delta)$ (in [17, Theorem 2]) to $\log(2ep/s)$. Thus the dependence of the right hand side of the oracle inequality in the confidence level δ is significantly reduced compared to the result of [17, Theorem 2].

A similar bound as in Corollary 2 could be obtained for ℓ_1 regularization adapting the proof from [4, Theorem 4.3]. However such a better bound would be obtained for a choice of regularization parameter relying on the $\mathbf{D}^\top$-sparsity of the signal. The Graph-Slope does not rely on such quantity, and thus Graph-Slope is adaptive to the unknown $\mathbf{D}^\top$-sparsity of the signal.

Remark 1. The optimal theoretical choice of parameter requires the knowledge of the noise level σ from the practitioner. Whenever the noise level σ is not known, the practitioner can use the corresponding Concomitant estimator to alleviate this issue [21, 5, 29], see also [19] for efficient algorithms to compute such scale-free estimators.

3. Numerical experiments

3.1. Algorithm for Graph-Slope

In this section, we propose an algorithm to compute a solution of the highly structured optimization problem (1.2). The data fidelity term $f : \beta \mapsto \|y - \beta\|_2^2/2$ is a convex smooth function with 1-Lipschitz gradient, and the map $\beta \mapsto \|\mathbf{D}^\top \beta\|_{[\lambda]}$ is the pre-composition by a linear operator of the norm $\|\cdot\|_{[\lambda]}$ whose proximal operator can be easily computed [36, 10]. Thus, the use of a dual or primal-dual proximal scheme can be advocated.

Problem (1.2) can be rewritten as

$$\min_{\beta \in \mathbb{R}^n} f(\beta) + g(\mathbf{D}^\top \beta) \ ,$$

where f is a smooth, 1-Lipschitz strictly convex function and $g = \|\cdot\|_{[\lambda]}$ is a convex, proper, lower semicontinuous function (see for instance [2, p. 275]). Its dual problem reads

$$\min_{\theta \in \mathbb{R}^p} f^\star(\mathbf{D}\theta) + g^\star(-\theta) \ ,$$

where $f^\star$ is the convex conjugate of f, *i.e.,* for any $x \in \mathbb{R}^n$

$$f^\star(x) = \sup_z \langle x, z\rangle - f(z) \ .$$

Classical computations leads to the following dual problem

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{D}\theta - y\|_2^2 - \frac{1}{2}\|y\|_2^2 \quad \text{subject to} \quad \|\theta\|_{[\lambda]}^* \leqslant 1 \ . \tag{3.1}$$

The dual formulation (3.1) can be rewritten as an unconstrained problem, using for any set $\mathcal{C} \subset \mathbb{R}^n$, and any $\theta \in \mathbb{R}^n$, the notation

$$\iota_{\mathcal{C}}(\theta) = \begin{cases} 0, & \text{if } \theta \in \mathcal{C} \\ +\infty, & \text{otherwise} \end{cases} .$$

Algorithm 1 FISTA on dual formulation

Require: (β^0, θ^0) initial guess, $L = \|\mathbf{D}\|^2$, $t_0 = 1$, ε duality gap tolerance
  $k \leftarrow 0$
  while $\Delta(\beta^k, \theta^k) > \varepsilon$ **do**
    $\theta^{k+1} \leftarrow \Pi_{\frac{1}{L}B_*}\left(\bar{\theta}^k - \frac{1}{L}(\mathbf{D}^\top(\mathbf{D}\bar{\theta}^k - y))\right)$ ▷ forward-backward step
    $\beta^{k+1} \leftarrow y - \mathbf{D}\theta^{k+1}$ ▷ necessary to compute Δ
    $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$ ▷ FISTA rule
    $\bar{\theta}^{k+1} \leftarrow \theta^{k+1} + \frac{t_k - 1}{t_{k+1}}(\theta^{k+1} - \theta^k)$ ▷ non-convex over-relaxation
    $k \leftarrow k + 1$
  end while

The quadratic term in y is constant and can be dropped. Thus the optimization problem (3.1) is equivalent to

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{D}\theta - y\|_2^2 + \iota_{\{\|\cdot\|_{[\lambda]}^* \leqslant 1\}}(\theta). \tag{3.2}$$

The formulation in (3.2) is now well suited to apply an accelerated version of the forward-backward algorithm such as FISTA [3]. As a stopping criteria, we use a duality gap criterion: $\Delta(\beta, \theta) \leqslant \varepsilon$, where

$$\Delta(\beta, \theta) = \frac{1}{2}\|y - \beta\|_2^2 + \|\mathbf{D}^\top \beta\|_{[\lambda]} + \frac{1}{2}\|\mathbf{D}\theta - y\|_2^2 - \frac{1}{2}\|y\|_2^2 \ ,$$

for a feasible pair (β, θ) and by $\Delta(\beta, \theta) = +\infty$ for an unfeasible pair. In practice we set $\varepsilon = 10^{-2}$ as a default value. Algorithm 1 summarizes the dual FISTA algorithm applied to the Graph-Slope minimization problem.

We recall that the proximity operator of a convex, proper, lower semicontinuous function f is given as the unique solution of the optimization problem

$$\mathrm{Prox}_{\lambda f}(\beta) = \operatorname*{argmin}_{z \in \mathbb{R}^n} \frac{1}{2}\|\beta - z\|_2^2 + \lambda f(z) \ .$$

To compute the proximity operator of $\iota_{\{\|\cdot\|_{[\lambda]}^* \leqslant 1\}}$, we use the Moreau's decomposition [22, p. 65] which links it to the proximity operator of the dual Slope-norm,

$$\begin{aligned}\theta &= \mathrm{Prox}_{\tau \iota_{\|\cdot\|_{[\lambda]}^* \leqslant 1}}(\theta) + \tau\, \mathrm{Prox}_{\frac{1}{\tau}\|\cdot\|_{[\lambda]}}\left(\frac{\theta}{\tau}\right) \\ &= \Pi_{\frac{1}{\tau}B_*}(\theta) + \tau\, \mathrm{Prox}_{\frac{1}{\tau}\|\cdot\|_{[\lambda]}}\left(\frac{\theta}{\tau}\right) \ ,\end{aligned}$$

where $\Pi_{\frac{1}{\tau}B_*}$ is the projection onto the unit ball B_* associated to the dual norm $\|\cdot\|^*$ scaled by a factor $1/\tau$. The proximity operator of $\|\cdot\|_{[\lambda]}$ can be obtained obtained in several ways [36, 10]. In our numerical experiments, we use the connection between this operator and the isotonic regression following [36], which can be computed in linear time. Under the assumption that the quantity

$(u_i - \lambda_i)$ is positive, non-increasing, computing $\mathrm{Prox}_{\|\cdot\|_{[\lambda]}}(u)$ is equivalent to solving the problem

$$\operatorname*{argmin}_{\theta\in\mathbb{R}^p} \frac{1}{2}\|u-\lambda-\theta\|_2^2 \quad \text{subject to} \quad \theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_p \geqslant 0 \ .$$

We have relied on the fast implementation implementation of the PAVA algorithm (for Pool Adjacent Violators Algorithm, see for instance [6]), available in `scikit-learn` [23] to solve this inner problem.

The source code used for our numerical experiments is freely available at `http://github.com/svaiter/gslope_oracle_inequality`.

3.2. Synthetic experiments

To illustrate the behavior of Graph-Slope, we first propose two synthetic experiments in moderate dimension. The first one is concerned with the so-called "Caveman" graph and the second one with the 1D path graph.

For these two scenarios, we analyze the performance following the same protocol. For a given noise level σ, we use the bounds derived in Theorem 2.1 (we dropped the constant term 8) and in [17], *i.e.,*

$$\lambda_{\mathrm{GL}} = \rho(\mathcal{G})\sigma\sqrt{\frac{2\log(p)}{n}} \quad \text{and} \quad (\lambda_{\mathrm{GS}})_j = \rho(\mathcal{G})\sigma\sqrt{\frac{2\log(p/j)}{n}} \quad \forall j \in [p] \ . \quad (3.3)$$

For every n_0 between 0 and p, we generate 1000 signals as follows. We draw J uniformly at random among all the subsets of $[p]$ of size n_0. Then, we let Π_J to be the projection onto $\operatorname{Ker} \mathbf{D}_J^\top$ and generate a vector $g \sim \mathcal{N}(0, \mathrm{Id}_n)$. We then construct $\beta^\star = c(\mathrm{Id} - \Pi_J)g$ where c is a given constant (here $c = 8$). This constrains the signal $\beta^\star$ to be of $\mathbf{D}^\top$-sparsity at most $p - n_0$.

We corrupt the signals by adding a zero mean Gaussian noise with variance σ^2, and run both the Graph-Lasso estimator and the Graph-Slope estimator. We then compute the mean of the mean-square error (MSE), the false detection rate (FDR) and the true detection rate (TDR). To clarify our vocabulary, given an estimator $\hat\beta$ and a ground truth $\beta^\star$, the MSE reads $\|\beta^\star - \hat\beta\|_n^2$, while the FDR and TDR read, respectively,

$$\mathrm{FDR}(\hat\beta, \beta^\star) = \begin{cases} \frac{|\{j\in[p] \,:\, j\in\mathrm{supp}(\mathbf{D}^\top\hat\beta) \text{ and } j\notin\mathrm{supp}(\mathbf{D}^\top\beta^\star)\}|}{|\,\mathrm{supp}(\mathbf{D}^\top\hat\beta)|} & \text{if } \mathbf{D}^\top\hat\beta \neq 0 \\ 0 & \text{if } \mathbf{D}^\top\hat\beta = 0, \end{cases}$$

and

$$\mathrm{TDR}(\hat\beta, \beta^\star) = \begin{cases} \frac{|\{j\in[p] \,:\, j\in\mathrm{supp}(\mathbf{D}^\top\hat\beta) \text{ and } j\in\mathrm{supp}(\mathbf{D}^\top\beta^\star)\}|}{|\,\mathrm{supp}(\mathbf{D}^\top\beta^\star)|}, & \text{if } \mathbf{D}^\top\beta^\star \neq 0, \\ 0, & \text{if } \mathbf{D}^\top\beta^\star = 0, \end{cases}$$

where for any $z \in \mathbb{R}^p$, $\mathrm{supp}(z) = \{j \in [p] \ : \ z_j \neq 0\}$.

Example on Caveman The caveman model was introduced in [34] to model small-world phenomenon in sociology. Here we consider its relaxed version, which is a graph formed by l cliques of size k (hence $n = lk$), such that with probability $q \in [0, 1]$, an edge of a clique is linked to a different clique. In our experiment, we set $l = 4$, $k = 10$ ($n = 40$) and $q = 0.1$. We provide a visualisation of such a graph in Figure 1a. For this realization, we have $p = 180$. The rewired edges are indicated in blue in Figure 1a whereas the edges similar to the complete graph on 10 nodes are in black. The signals are generated as random vectors of given $\mathbf{D}^\top$-sparsity with a noise level of $\sigma = 0.2$. Figure 1b shows the weights decay.

Figures 1c–1e represent the evolution of the MSE and TDR in function of the level of $\mathbf{D}^\top$-sparsity. We observe that while the MSE is close between the Graph-Lasso and the Graph-Slope estimator at low level of sparsity, the TDR is vastly improved in the case of Graph-Slope, with a small price concerning the FDR (a bit more for the Monte Carlo choice of the weights). Hence empirically, Graph-Slope will make more discoveries than Graph-Lasso without impacting the overall FDR/MSE, and even improving it.

Example on a path: 1D–Total Variation The classical 1D–Total Variation corresponds to the Graph-Lasso estimator $\hat{\beta}^{\mathrm{GL}}$ when $\mathcal{G}$ is the path graph over n vertices, hence with $p = n - 1$ edges. In our experiments, we take $n = 40$ and $\sigma = 0.6$. We display a typical realization of such signal in Figure 2a. Figure 2b shows the weights decay. Note that in this case, the Monte–Carlo weights shape differs from the one in the previous experiment. Indeed, they are adapated to the underlying graph, contrary to the theoretical weights λ_{GS} which depend only on the size of the graph. Figures 2c–2e represent the evolution of the MSE and TDR in function of the level of $\mathbf{D}^\top$-sparsity. Here, Graph-Slope does not improve the MSE significantly. However, as for the caveman experiments, Graph-Slope is more likely to make more discoveries than Graph-Lasso for a small price concerning the FDR.

3.3. Example on real data: Paris roads network

To conclude our numerical experiments, we present our result on a real-life graph, the road network of Paris, France. Thanks to the Python module `osmnx` [9], which downloads and simplifies OpenStreetMap data, we run our experiments on $p = 20108$ streets (edges) and $n = 10205$ intersections (vertices).

The ground truth signal is constructed as in [16] as follows. Starting from 30 infection sources, each infected intersections has a probability 0.75 to infect each of its neighbors. We let the infection process runs for 8 iterations. The resulting graph signal $\beta^\star$ is represented in Figure 3a with $\mathbf{D}^\top$-sparsity 1586. We then corrupt this signal by a zero mean Gaussian noise with standard-deviation $\sigma = 0.8$, leading to the observations y represented in Figure 3b.

Instead of using the parameters given in (3.4), we have computed the oracle parameters for the Graph-Lasso and Graph-Slope estimators by evaluating for

100 parameters of the form

$$\lambda_{\mathrm{GL}} = \alpha\sigma\sqrt{\frac{2\log(p)}{n}} \quad \text{and} \quad (\lambda_{\mathrm{GS}})_j = \alpha\sigma\sqrt{\frac{2\log(p/j)}{n}} \quad \forall j \in [p] \ , \tag{3.4}$$

where α lives on a geometric grid inside $[10^{-5}, 10^{1.5}]$. The best one in term of MSE (*i.e.,* in term of $\|\hat\beta - \beta^\star\|_n^2$) is refered to as the oracle parameter. The results are illustrated in Figure 3c for Graph-Lasso and in Figure 3d for Graph-Slope. We can see the benefit of Graph-Slope, for instance in the center of Paris where the sources of infections are better identified as shown in the close-up, see Figures 3e–3g.

Appendix A: Preliminary lemmas

Lemma A.1. *Let $s \in [p]$. For any two $\beta, \hat\beta \in \mathbb{R}^n$ such that $\|\mathbf{D}^\top\beta\|_0 \leqslant s$ we have*

$$\|\mathbf{D}^\top\beta\|_{[\lambda]} - \|\mathbf{D}^\top\hat\beta\|_{[\lambda]} \leqslant \sum_{j=1}^{s}\lambda_j|u|_j^\downarrow - \sum_{j=s+1}^{p}\lambda_j|u|_j^\downarrow, \leq \Lambda(\lambda,s)\|u\|_2 - \sum_{j=s+1}^{p}\lambda_j|u|_j^\downarrow,$$

where $u = \mathbf{D}^\top(\hat\beta - \beta)$ and $\Lambda(\lambda,s) = \left(\sum_{j=1}^s \lambda_j^2\right)^{1/2}$.

Proof. This is a consequence of [4, Lemma A.1]. We provide a proof here for completeness. The second inequality is simple consequence of the Cauchy-Schwarz inequality. Indeed, $(\sum_{j=1}^s \lambda_j|u|_j^\downarrow)^2 \leq (\sum_{j=1}^s \lambda_j^2)(\sum_{j=1}^s(|u|_j^\downarrow)^2) \leq \|u\|_2^2(\sum_{j=1}^s\lambda_j^2)$.

Let $v = \mathbf{D}^\top\beta$ and $w = \mathbf{D}^\top\hat\beta$ and note that $u = w - v$. Let φ be a permutation of $[p]$ such that $\|v\|_{[\lambda]} = \sum_{j=1}^s \lambda_j|v|_{\varphi(j)}$. The first inequality can be rewritten as

$$\|v\|_{[\lambda]} - \|w\|_{[\lambda]} = \|v\|_{[\lambda]} - \sup_\tau \sum_{j=1}^{p}\lambda_j|w|_{\tau(j)} \leqslant \sum_{j=1}^{s}\lambda_j|u|_j^\downarrow - \sum_{j=s+1}^{p}\lambda_j|u|_j^\downarrow, \tag{A.1}$$

where the supremum is taken over all permutations τ of $[p]$. We now prove (A.1). Let τ be a permutation of $[p]$ such that $\varphi(j) = \tau(j)$ for all $j = 1, \dots, s$. Then by the triangle inequality, we have

$$|v|_{\varphi(j)} - |w|_{\varphi(j)} = |v|_{\varphi(j)} - |w|_{\tau(j)} \leq |u|_{\tau(j)}$$

for each $j = 1, \dots, s$ since $u = w - v$. Furthermore, for each $j > s$, it holds that $v_{\tau(j)} = 0$ so that $w_{\tau(j)} = u_{\tau(j)}$. Thus,

$$\|v\|_{[\lambda]} - \|w\|_{[\lambda]} \leq \sum_{j=1}^{s}\lambda_j|v|_{\varphi(j)} - \sum_{j=1}^{p}\lambda_j|w|_{\tau(j)} \leq \sum_{j=1}^{s}\lambda_j|u|_{\tau(j)} - \sum_{j=s+1}^{p}\lambda_j|u|_{\tau(j)}.$$

It is clear that $\sum_{j=1}^s \lambda_j|u|_{\tau(j)} \leq \sum_{j=1}^s\lambda_j|u|_j^\downarrow$. Finally, notice that it is always possible to find a permutation τ such that $(|u|_{\tau(j)})_{j>s}$ is non-decreasing. For such choice of τ we have $-\sum_{j=s+1}^p\lambda_j|u|_{\tau(j)} \leq -\sum_{j=s+1}^p\lambda_j|u|_j^\downarrow$. □

Lemma A.2. *For the choice of weights:* $\forall j \in [p], \lambda_j = C\sqrt{\log(2p/j)/n}$, *the following inequalities hold true*

$$C\sqrt{\frac{s\log(2p/s)}{n}} \leqslant \Lambda(\lambda, s) \leqslant C\sqrt{\frac{s\log(2ep/s)}{n}} \ .$$

Proof. Reminding Stirling's formula $s\log(s/e) \leqslant \log(s!) \leqslant s\log(s)$, one can check that

$$s\log\left(\frac{2p}{s}\right) \leqslant \sum_{j=1}^{s} \log\left(\frac{2p}{j}\right) \leqslant s\log(2p) - \log(s!) = \log\left(\frac{2ep}{s}\right) \ .$$

The lemma holds true multiplying by $C/\sqrt{n}$ both sides of the previous display. □

Lemma A.3. *Let* $z, \varepsilon \in \mathbb{R}^n$, $y = z + \varepsilon$, *and* $\hat{\beta}$ *a solution of Problem* (1.2). *Then, for all* $\beta \in \mathbb{R}^n$,

$$\|\hat{\beta} - z\|_n^2 - \|\beta - z\|_n^2 \leqslant \frac{2}{n}\varepsilon^\top(\hat{\beta} - \beta) + \|\mathbf{D}^\top\beta\|_{[\lambda]} - \|\mathbf{D}^\top\hat{\beta}\|_{[\lambda]} - \|\hat{\beta} - \beta\|_n^2.$$

Proof. The objective function of the minimization problem (1.2) is the sum of two convex functions. The first term, *i.e.,* the function $\beta \to \frac{1}{2}\|y - \beta\|_n^2$, is 1-strongly convex with respect to the norm $\|\cdot\|_n$. The sum of a 1-strongly convex function and a convex function is 1-strongly convex, and thus we have

$$\frac{1}{2}\|\hat{\beta} - y\|_n^2 + \|\mathbf{D}^\top\hat{\beta}\|_{[\lambda]} \leq \mathbf{d}^T(\hat{\beta} - \beta) + \frac{1}{2}\|\beta - y\|_n^2 + \|\mathbf{D}^\top\beta\|_{[\lambda]} - \frac{1}{2}\|\beta - \hat{\beta}\|_n^2$$

for all $\beta \in \mathbb{R}^n$ and for any $\mathbf{d}$ in the subdifferential of the objective function (1.2) at $\hat{\beta}$. Since $\hat{\beta}$ is a minimizer of (1.2), we can choose $\mathbf{d} = 0$ in the above display. For $\mathbf{d} = 0$, the previous display is equivalent to the claim of the Lemma. □

Lemma A.4. *Let us suppose that the graph* $\mathcal{G}$ *has* K *connected components* $C_1, \ldots, C_K$. *Then,*

$$\ker(\mathbf{D}^\top) = \operatorname{span}(\mathbf{1}_{C_1}) \oplus \cdots \oplus \operatorname{span}(\mathbf{1}_{C_K}),$$

where for any $k \in [K]$, *the vectors* $\mathbf{1}_{C_k} \in \mathbb{R}^n$ *are defined by*

$$(\mathbf{1}_{C_k})_i = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases} \ , \ \text{for } i = 1, \cdots, |V| \ .$$

Moreover, the orthogonal projection over $\ker(\mathbf{D}^\top)$ *is denoted by* Π *and is the component-wise averaging given by*

$$(\Pi(\beta))_i = \frac{1}{|C_k|}\sum_{i \in C_k} \beta_i, \ \text{where } k \text{ is such that } C_k \ni i, \ \text{for } i = 1, \ldots, n.$$

Furthermore, if $\mathcal{G}$ *is a connected graph then* $\ker(\Pi) = \operatorname{span}(\mathbf{1}_n)$.

Proof. The proof can be done for the simple case of a connected graph (*i.e.,* $K = 1$), and the result can be generalized by tensorization of graph for $K > 1$ components. Hence, we assume that $K = 1$. For any $\beta \in \ker(\mathbf{D}^\top)$, the definition of the incidence matrix yields that for all $(i, j) \in E, \beta_i = \beta_j$. Since all vertices are connected, by recursion all the β_j's are identical, and $\beta \in \operatorname{span}(\mathbf{1}_n) = \operatorname{span}(\mathbf{1}_{C_1})$. The converse is proved in the same way. □

Lemma A.5 (Proposition E.2 in [4])**.** *Let $g_1, \dots, g_p$ be centered Gaussian random variables (not necessarily independent) with variance at most $V > 0$. Then,*

$$\mathbb{P}\left(\max_{j=1,\dots,p} \frac{|g|_j^{\downarrow}}{\sqrt{V \log(2p/j)}} \le 4 \right) \ge 1/2 \ .$$

References

[1] I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1):39–54, 1989.

[2] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] P. C Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016.

[5] A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[6] M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Math. Program.*, 47(1-3):425–439, 1990.

[7] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009.

[9] G. Boeing. OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. *ArXiv e-prints ArXiv:1611.01890*, 2016.

[10] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE-adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103, 2015.

[11] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[12] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

[13] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

[14] C.-A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter. CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration. *SIAM J. Imaging Sci.*, 10(1):243–284, 2017.

[15] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947–968, 2007.

[16] Z. Fan and L. Guan. ℓ_0-estimation of piecewise-constant signals on graphs. *ArXiv e-prints ArXiv:1703.01421*, 2017.

[17] J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. *ArXiv e-prints ArXiv:1603.09388*, 2016.

[18] E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997.

[19] E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. In *NCMIP*, 2017.

[20] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, volume 14, pages 849–856, 2001.

[21] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

[22] N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein. Proximal algorithms. *Foundations and Trends in Machine Learning*, 1(3):1–108, 2013.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[24] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[25] V. Sadhanala, Y.-X. Wang, and R. J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *NIPS*, pages 3513–3521, 2016.

[26] J. Sharpnack, A. Singh, and A. Rinaldo. Sparsistency of the edge lasso over graphs. In *AISTATS*, volume 22, pages 1028–1036, 2012.

[27] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[28] W. Su and E. J. Candès. Slope is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016.

[29] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[30] R. Tibshirani, M. A. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused LASSO. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[31] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. 3:1360–1392, 2009.

[32] V. Viallon, S. Lambert-Lacroix, H. Hoefling, and F. Picard. On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, 26(1-2):285–301, 2016.

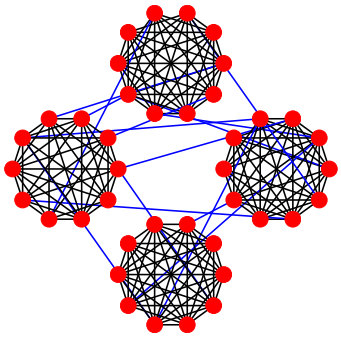
[33] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*,

17(4):395–416, 2007.

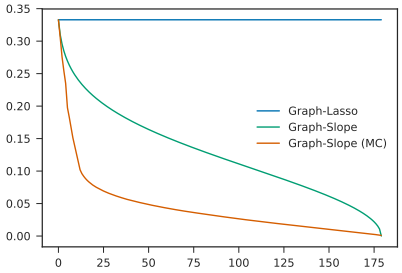
[34] D. J. Watts. Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, 105(2):493–527, 1999.

[35] F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *J. Mach. Learn. Res.*, 11(Dec):3519–3540, 2010.

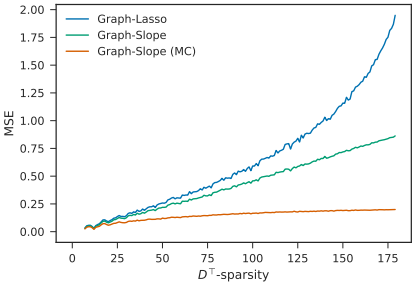
[36] X. Zeng and M. A. T. Figueiredo. The Ordered Weighted ℓ_1 Norm: Atomic Formulation, Projections, and Algorithms. *ArXiv e-prints arXiv:1409.4271*, 2014.



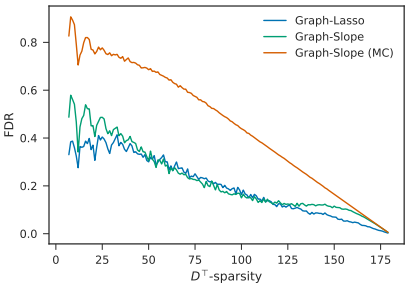
(a) Realization of a caveman graph



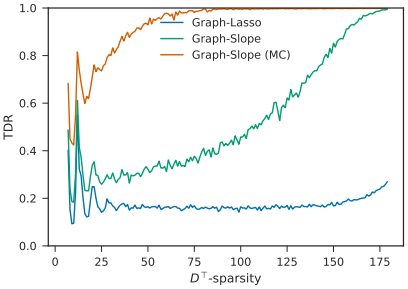
(b) Weights



(c) Mean-square error (MSE)

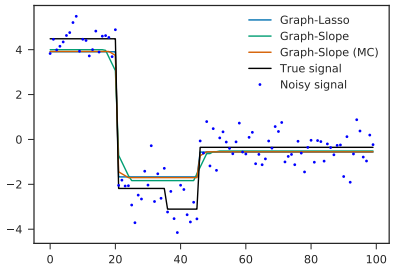


(d) False Detection Rate (FDR)

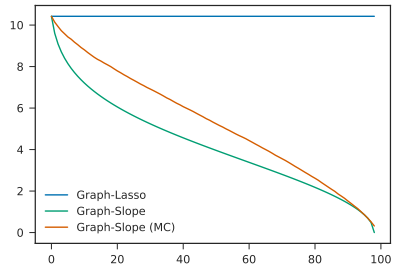


(e) True Detection Rate (TDR)

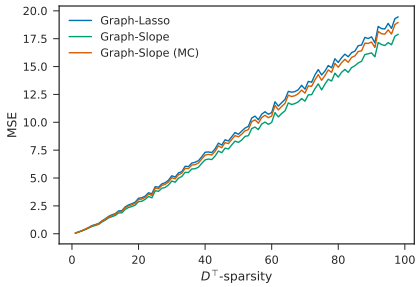
Fig 1: Relaxed caveman denoising



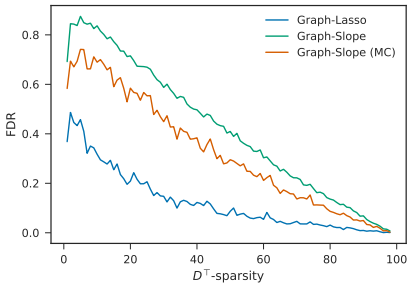
(a) Example of signal



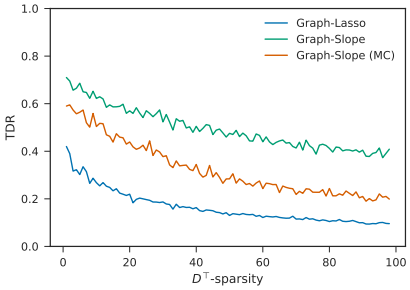
(b) Weights



(c) Mean-square error (MSE)

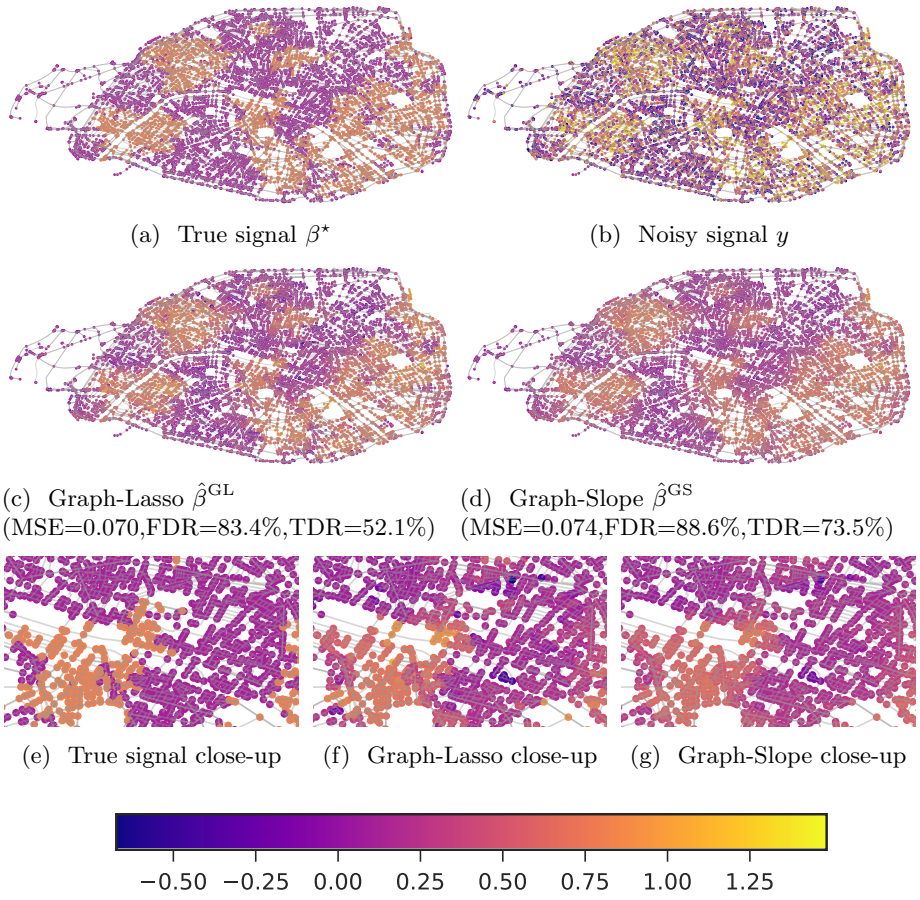


(d) False Detection Rate (FDR)



(e) True Detection Rate (TDR)

Fig 2: TV1D



(a) True signal $\beta^\star$ (b) Noisy signal y

(c) Graph-Lasso $\hat{\beta}^{\mathrm{GL}}$ (MSE=0.070,FDR=83.4%,TDR=52.1%)

(d) Graph-Slope $\hat{\beta}^{\mathrm{GS}}$ (MSE=0.074,FDR=88.6%,TDR=73.5%)

(e) True signal close-up (f) Graph-Lasso close-up (g) Graph-Slope close-up

Fig 3: Paris road network. Comparaison of oracle choice for the tuning parameter between Graph-Lasso and Graph-Slope.