Optimal rates for total variation denoising

Jan-Christian Hütter and Philippe Rigollet*

Massachusetts Institute of Technology

Abstract. Motivated by its practical success, we show that the two-dimensional total variation denoiser satisfies a sharp oracle inequality that leads to near optimal rates of estimation for a large class of image models such as bi-isotonic, Hölder smooth and cartoons. Our analysis hinges on properties of the unnormalized Laplacian of the two-dimensional grid such as eigenvector delocalization and spectral decay. We also present extensions to more than two dimensions as well as several other graphs.

AMS 2000 subject classifications: Primary 62G08; secondary 62C20, 62G05, 62H35.
Key words and phrases: Total variation regularization, TV denoising, sharp oracle inequalities, image denoising, edge Lasso, trend filtering, nonparametric regression, shape constrained regression, minimax.

1. INTRODUCTION

Total variation image denoising has known a spectacular practical success since its introduction by [ROF92] more than two decades ago. Surprisingly, little is known about its statistical performance. In this paper, we close this gap between theory and practice by providing a novel analysis for this estimator in a Gaussian white noise model. In this model, we observe a vector $y \in \mathbb{R}^n$ defined as

$$y = \theta^* + \varepsilon\,, \tag{1.1}$$

where $\theta^* \in \mathbb{R}^n$ is the unknown parameter of interest and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is a Gaussian random vector. In practice, θ^* corresponds to a vectorization of an image and we observe it corrupted by the noise ε. The goal of image denoising is to estimate θ^* as accurately as possible. In this paper, we follow the standard employed in the image denoising literature and measure the performance of an estimator $\hat{\theta}$ by its *mean squared error*. It is defined by

$$\mathsf{MSE}(\hat{\theta}) := \frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2.$$

Note that a lot of the work concerning the fused Lasso in the context of graphs has been focused on sparsistency results, *i.e.*, conditions under which we can expect to recover the set of edges along which the signal has a jump [QJ12, SSR12, OV15, VLLHP16], which is a different objective than controlling the MSE.

*Supported in part by NSF grants DMS-1317308 and CAREER-DMS-1053987.

The total variation (TV) denoiser $\hat{\theta}$ is defined as follows. Let $G = (V, E)$ be an undirected connected graph with vertex set V and edge set E such that $|V| = n, |E| = m$. The graph G traditionally employed in image denoising is the two-dimensional (2D) grid graph defined as follows. The vertex set is $V = [N]^2$ and the edge set $E \subset [N]^2 \times [N]^2$ contains edge $e = \big([i,j],[k,l]\big)$ if and only if $[k,l] - [i,j] \in \{[1,0],[0,1]\}$. Nevertheless, our results remain valid for other graphs as discussed in Section 4 and we work with a general graph G unless otherwise mentioned.

Throughout this paper it will be convenient to represent a graph G by its edge-vertex incidence matrix $D = D(G) \in \{-1, 0, 1\}^{m \times n}$. Without loss of generality, identify V to $[n]$ and E to $[m]$ whenever convenient. To each edge $e = (i, j) \in E$ corresponds a row $D_{e,:}$ of D with entries given as follows. The kth entry $D_{e,k}$ of $D_{e,:}$ is given by

$$D_{e,k} = \begin{cases} 1 & \text{if } k = \min(i,j) \\ -1 & \text{if } k = \max(i,j) \\ 0 & \text{otherwise.} \end{cases}$$

Note that the matrix $L = D^\top D$ is the *unnormalized Laplacian* of the graph G [Chu97]. It can be represented as $L = \mathsf{diag}(A\mathbb{1}_n) - A$, where A is the adjacency matrix of G and $\mathsf{diag}(A\mathbf{1}_n)$ is the diagonal matrix with jth diagonal element given by the degree of vertex j.

The TV denoiser $\hat{\theta}$ associated to G is then given by any solution to the following minimization problem

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{n}\|\theta - y\|_2^2 + \lambda \|D\theta\|_1 \,, \tag{1.2}$$

where $\lambda > 0$ is a regularization parameter to be chosen carefully. Our results below give a precise choice for this parameter. Note that (1.2) is a convex problem that may be solved efficiently (see [AT16] and references therein).

Akin to the sparse case, the TV penalty in (1.2) is a convex relaxation for the number of times θ changes values along the edges of G. Intuitively, this is a good idea if θ^* takes small number of values for example. In this paper, we favor an analysis where θ^* is not of such form but may be well approximated by a piecewise constant vector. Our main result, Theorem 2, is a sharp oracle inequality that trades off approximation error against estimation error. In Section 5, we present several examples where approximation error can be explicitly controlled: Hölder functions, Isotonic matrices and cartoon images. In each case, our results are near optimal in a minimax sense.

Our analysis partially leverages insight gained from recent results for the one-dimensional case where G is the path graph by [DHL14]. In this case, the TV denoiser is often referred to as a fused (or fusion) Lasso [TSR+05, Rin09]. Moreover, the TV denoiser $\hat{\theta}$ defined in (1.2) is often called to *generalized fused Lasso*. The analysis provided in [DHL14] is specific to the path graph and does not extend to more general graphs. We extend these results to other graphs, with particular emphasis on the 2D grid. Critically, our analysis can be extended to graphs with specific spectral properties, such as random graphs with bounded degree. It is worth mentioning that our techniques, unfortunately do not recover the results of [DHL14] for the path graph.

1.1 Notation

For two integers $n, m \in \mathbb{N}$, we write $[n] = \{1, \dots, n\}$, $[\![n, m[\![= \{n, n+1, \dots, m-1\}$ and $[\![n, m]\!] = \{n, n+1, \dots, m\}$. Moreover, for two real numbers a, b, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

We reserve bold-face letters like $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ for multi-indices whose elements are written in regular font, *e.g.*, $\boldsymbol{i} = (i_1, \dots, i_d)$.

We denote by $\mathbf{1}_d$ the all-ones vector of $\mathbb{R}^d$.

We write $\mathbb{1}(\cdot)$ for the indicator function.

For any two sets $A, B \subset \mathbb{R}^d$ we define their Minkowski sum as $A + B = \{a + b\,, a \in A, b \in B\}$. Moreover, for any $\eta \geq 0$, we denote by $\mathcal{B}(\eta) = \{x \in \mathbb{R}^d\,, \|x\| \leq \eta\}$ the Euclidean ball of radius η.

For any vector $x \in \mathbb{R}^d$, $T \subset [d]$, we define $x_T \in \mathbb{R}^d$ to be the vector with j coordinate given by $(x_T)_j = x_j \mathbb{1}(j \in T)$.

We denote by $A^\dagger$ the Moore-Penrose pseudo-inverse of a matrix A. Moreover, $\otimes$ denotes the Kronecker product between matrices, $(A \otimes B)_{p(r-1)+v,q(s-1)+w} = A_{r,s}B_{v,w}$.

The notation $\lesssim$ means that the left-hand side is bounded by the right-hand side up to numerical constant that might change from line to line. Similarly, the constants C, c are generic as well and are allowed to change.

1.2 Previous work

Despite an overwhelming practical success, theoretical results for the TV denoiser have been very limited. Recently, several advances were made in [NW13] and [WSST15]. The latter paper is perhaps the closest to our contribution and we propose a more detailed discussion here.

First and foremost, [WSST15] studies the more general framework of *trend filtering* where instead of applying the difference operator D in the penalty, one may apply D^{k+1} (with appropriate corrections due to the shrinking dimension of the image space). In this paper, we focus on the case where $k = 0$.

Second, while our paper focuses on fast rates (of the order $1/n$), [WSST15] also studies graphs that lead to slower rates. A prime example is the path graph that is omitted from the present work and where [WSST15] recover the optimal rate $n^{2/3}$ for signals θ^* such that $\|D\theta^*\|_1 \leq C$. This rate was previously known to be optimal [DJ95] for such signals, using comparison with Besov spaces. Remarkably, if θ^* is piecewise constant with large enough pieces, [DHL14] proved that this rate can be improved to a rate of order $1/n$ using a rather delicate argument. Moreover, their result is also valid in a oracle sense, allowing for model misspecification and leading to adaptive estimation of smooth functions on the real line. Part of our results extend this application to higher dimensions.

Third, it was also noticed by [WSST15] that fast rates arise as soon as the underlying graph satisfies two conditions: (i) bounded degree and (ii) constant spectral gap. We obtain similar results for these graphs but our techniques lead to slightly tighter bounds that are strictly better for graphs with large maximum degree.

Perhaps the most important point of comparison is that our analysis leads to an optimal fast rate of order $1/n$ for the 2D grid, unlike the rate $n^{-4/5}$ that was obtained by [WSST15]. This is achieved by a careful analysis of the pseudo inverse $D^\dagger$ of D. In particular, our argument bypasses truncation of the spectrum altogether.

Our fast rates lead to optimal rates of estimation in several models of interest for the signal θ^* that were not obtained in [WSST15]. These applications are detailed in section 5.

Finally, our results are also valid in the case of misspecified models and are presented in terms of sharp oracle inequalities. Moreover, in the spirit of [DHL14], we allow for a tradeoff between rates that depend on the ℓ_1-norm of the discrete gradient and the number of active edges (ℓ_0-norm of the discrete gradient). The latter result allows for scale-independent rates in the case that the jumps across active edges are large. The scale-free results are key in obtaining optimal rates for cartoon images in subsection 5.2. Both the oracle part and the scale-free results are novel compared to [WSST15].

2. SHARP ORACLE INEQUALITY

We start by defining two quantities involved in estimating the performance of the Lasso:

DEFINITION 1 (Compatibility factor, inverse scaling actor). *Let $D \in \{-1,0,1\}^{m\times n}$ be an incidence matrix and write $S := D^\dagger = [s_1,\dots,s_m]$. The* compatibility factor *of D for a set $T \subseteq [m]$ is defined as*

$$\kappa_\emptyset := 1, \quad \kappa_T = \kappa_T(D) := \inf_{\theta\in\mathbb{R}^n} \frac{\sqrt{|T|}\|\theta\|_2}{\|(D\theta)_T\|_1} \quad \text{for } T \neq \emptyset\,.$$

If we omit the subscript, then we mean the worst possible value of the constant, i.e., *$\kappa = \inf_{T\subseteq[m]} \kappa_T$.*

Moreover, the inverse scaling factor *of D is defined as*

$$\rho = \rho(D) := \max_{j\in[m]} \|s_j\|_2\,.$$

We prove the following main result.

THEOREM 2 (Sharp oracle inequality for TV denoising). *Fix $\delta \in (0,1)$, $T \subset [m]$ and let D being the incidence matrix of a connected graph G. Define the regularization parameter*

$$\lambda := \frac{1}{n}\sigma\rho\sqrt{2\log\big(\frac{em}{\delta}\big)},$$

With this choice of λ, the TV denoiser $\hat\theta$ defined in (1.2) *satisfies*

$$\frac{1}{n}\|\hat\theta - \theta^*\|^2 \le \inf_{\bar\theta\in\mathbb{R}^n}\left\{\frac{1}{n}\|\bar\theta-\theta^*\|^2 + 4\lambda\|(D\bar\theta)_{T^c}\|_1\right\} + \frac{8\sigma^2}{n}\left(\frac{|T|\rho^2}{\kappa_T^2}\log\big(\frac{em}{\delta}\big) + \log\big(\frac{e}{\delta}\big)\right). \tag{2.3}$$

on the estimation error with probability at least $1-2\delta$.

We delay the proof to the Appendix, Subsection B.1.

The sharp oracle inequality (2.3) allows trading off $|T|$ with $\|(D\bar\theta)_{T^c}\|_1$. For $T = \operatorname{supp}(D\bar\theta)$, we recover the ℓ_0 rate $\sigma^2\kappa_T^{-2}\rho^2\log(m/\delta)|T|/n$, while setting $T = \emptyset$ yields the ℓ_1 rate $\sigma\rho\sqrt{\log(m/\delta)}\|D\bar\theta\|_1/n$. We will see in Section 5 that both rates are essential to get minimax rates for certain complexity classes

In order to evaluate the performance of the TV denoiser $\hat\theta$ on any graph G and in particular on the 2D grid, we need estimates on ρ and κ.

It turns out that bounding the compatibility factor is rather easy for all bounded degree graphs.

LEMMA 3. *Let D be the incidence matrix of a graph G with maximal degree d and $\emptyset \neq T \subseteq E$. Then,*

$$\kappa_T = \inf_{\theta\in\mathbb{R}^n} \frac{\sqrt{|T|}\|\theta\|}{\|(D\theta)_T\|_1} \ge \frac{1}{2\min\{\sqrt{d},\sqrt{|T|}\}}.$$

PROOF. Let D be the incidence matrix of a graph $G = (V,E)$, $\theta \in \mathbb{R}^n$, and let $T \subset E = [m]$. Moreover, denote by $d_i = \#\{j \in [n] : (i,j) \in E\}$ the degree of vertex i and by $d = \max_{i\in[n]} d_i$ the maximum degree of the graph.

Then, by triangle inequality,

$$\|(D\theta)_T\|_1 = \sum_{(i,j)\in T} |\theta_i - \theta_j| \le \sqrt{|T|}\sqrt{\sum_{(i,j)\in T} |\theta_i - \theta_j|^2} \le 2\sqrt{|T|}\min\{\sqrt{|T|}, \sqrt{d}\}\|\theta\|_2$$

□

3. TOTAL VARIATION REGULARIZATION ON THE GRID

3.1 TV regularization in 2D

In this section, we show that $\rho \lesssim \sqrt{\log n}$. Note that this is different from the 1D case: if we consider the incidence matrix $\widetilde{D}$ of the path graph and for simplification add an additional row penalizing the absolute value of the first entry, *i.e.*,

$$(\widetilde{D}\theta)_1 = \theta_1, \quad (\widetilde{D}\theta)_i = \theta_i - \theta_{i-1}, \quad i = 2, \dots, n,$$

then one can show that $(D^\dagger)_{i,j} = (D^{-1})_{i,j} = \mathbb{1}(i \ge j)$. Hence, in this case $\rho = \sqrt{n}$. Moreover, the inverse scaling factor ρ remains of the order $\sqrt{n}$ even if we close the path into a cycle. The analyses of [WSST15] and [DHL14] are geared towards refining the estimates used in the proof of Theorem 2 in order to recover rates faster than $n^{-1/2}$. Rather, we focus on extending results to the central example of the two dimensional grid, which is paramount in image processing.

We proceed to estimate ρ in the case of the total variation regularization on the $N \times N$ 2D grid. Let $n = N^2$ and write $D_1 \in \mathbb{R}^{(N-1)\times N}$ for the incidence matrix of the path graph on N vertices, $D_1 x = x_{j+1} - x_j$, $j = 1, \dots, N-1$ for $x \in \mathbb{R}^N$.

Reshaping a signal θ on the $N \times N$ square in column major form as a vector $\theta \in \mathbb{R}^n$, we can write the incidence matrix of the grid as

$$D_2 = \begin{bmatrix} D_1 \otimes I \\ I \otimes D_1 \end{bmatrix}.$$

Proposition 4. *The incidence matrix D_2 of the 2D grid on n vertices has inverse scaling factor $\rho \lesssim \sqrt{\log n}$.*

We delay the proof to the Appendix, Subsection B.2. By combining the estimates from Lemma 3 and Proposition 4 with Theorem 2, we get the following rate for TV regularization on a regular grid in 2D.

Corollary 5. *Fix $\delta \in (0,1)$ and let D denote the incidence matrix of the 2D grid. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat{\theta}$ defined in (1.2) with $\lambda = c\sigma\sqrt{(\log n)\log(en/\delta)/n}$ satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \inf_{\substack{\bar{\theta}\in\mathbb{R}^n \\ T\subseteq[m]}} \left\{ \frac{1}{n}\|\bar{\theta} - \theta^*\|^2 + 4\lambda\|(D\bar{\theta})_{T^c}\|_1 \right\} + \frac{C\sigma^2}{n}\left(|T|(\log n)\log(en/\delta) + \log(e/\delta)\right), \quad (3.4)$$

with probability at least $1 - 2\delta$. In particular, it yields

$$\mathsf{MSE}(\hat{\theta}) \lesssim \frac{\sigma\|D\theta^*\|_1 \wedge \sigma^2\|D\theta^*\|_0}{n}\log^2(en/\delta)$$

where $\|D\theta^*\|_0$ denotes the number of nonzero components of $D\theta^*$.

3.2 TV regularization in higher dimensions

Akin to the 2D case, in d dimensions, we have $n = N^d$ and we can write

$$D_d = \begin{bmatrix} D_1 \otimes I \otimes \cdots \otimes I \\ I \otimes D_1 \otimes \cdots \otimes I \\ \vdots \\ I \otimes I \otimes \cdots \otimes D_1 \end{bmatrix}.$$

Using similar calculations as in the 2D case, we can show that the inverse scaling factor ρ is now bounded by a constant, uniformly in N.

PROPOSITION 6. *For the incidence matrix of the regular grid on N^d nodes in d dimensions, $\rho \le C(d)$, for some $C(d) > 0$.*

We delay the proof to the Appendix, subsection B.3. It readily yields the following rate for TV regularization on a regular grid in $d \ge 3$ dimensions:

COROLLARY 7. *Fix $\delta \in (0,1)$, an integer $d \ge 3$ and let D_d denote the incidence matrix of the d-dimensional grid. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat\theta$ defined in* (1.2) *with $\lambda = c\sigma\sqrt{\log(en/\delta)}/n$ satisfies*

$$\frac{1}{n}\|\hat\theta - \theta^*\|^2 \le \inf_{\substack{\bar\theta \in \mathbb{R}^n \\ T \subseteq [m]}} \left\{ \frac{1}{n}\|\bar\theta - \theta^*\|^2 + 4\lambda \|(D_d\bar\theta)_{T^c}\|_1 \right\} + \frac{C\sigma^2}{n}\left(|T|\log(en/\delta) + \log(e/\delta)\right),$$

with probability at least $1 - 2\delta$. In particular, it yields

$$\mathsf{MSE}(\hat\theta) \lesssim \frac{\sigma\|D_d\theta^*\|_1 \wedge \sigma^2\|D_d\theta^*\|_0}{n}\log(en/\delta).$$

3.3 The hypercube

We note that in the case $N = 2$, the grid becomes the d-dimensional hypercube. In this case, we can refine our analysis in this case to get the same result as in Proposition 6 without dependence on the dimension.

PROPOSITION 8. *For any $d \ge 1$, the inverse scaling factor associated to the d-dimensional hypercube satisfies $\rho \le 1$.*

PROOF. We use the same analysis as in the proof of Proposition 6, Subsection B.3, noting that the eigenvectors of the 1-dimensional hypercube are given by $v_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top/\sqrt{2}$ and $v_2 = \begin{bmatrix} 1 & -1 \end{bmatrix}^\top/\sqrt{2}$ with associated eigenvalues 0 and 2, respectively. Using the same notation as before, we have

$$\langle v_{k_j}, e_{i_j}\rangle^2 \le 1/2, \quad \text{for } \boldsymbol{k}, \boldsymbol{i} \in \{0,1\}^d, j \in [d],$$

$$\langle v_{k_1}, d_{i_1}\rangle^2 \le 2.$$

This gives

$$\|s_{\boldsymbol{i}}^{(1)}\|_2^2 = \sum_{\boldsymbol{k}\in\{0,1\}^d\setminus\{0\}} \left(\sum_{j=1}^d \lambda_{k_j}\right)^{-2} \langle v_{k_1}, d_{i_1}\rangle^2 \prod_{j=2}^d \langle v_{k_j}, e_{i_j}\rangle^2$$
$$\leq 2^{2-d} \sum_{\boldsymbol{k}\in\{0,1\}^d\setminus\{0\}} \left(\sum_{j=1}^d 2k_j\right)^{-2} \leq 1.$$

□

COROLLARY 9. *Fix $\delta \in (0,1)$, an integer $d \geq 1$ and let $D_\square$ denote the incidence matrix of the d-dimensional hypercube. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat{\theta}$ defined in (1.2) with $n = 2^d$ and $\lambda = c\sigma\sqrt{\log(en/\delta)}/n$ satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \leq \inf_{\substack{\bar{\theta}\in\mathbb{R}^n \\ T\subseteq[m]}} \left\{\frac{1}{n}\|\bar{\theta} - \theta^*\|^2 + 4\lambda\|(D_\square\bar{\theta})_{T^c}\|_1\right\} + \frac{C\sigma^2}{n}\left(d|T|\log(en/\delta) + \log(e/\delta)\right),$$

with probability at least $1 - 2\delta$. In particular, it yields

$$\mathsf{MSE}(\hat{\theta}) \lesssim \frac{\sigma\|D_\square\theta^*\|_1 \wedge \sigma^2 d\|D_\square\theta^*\|_0}{n}\log(en/\delta).$$

4. OTHER GRAPHS

4.1 Complete graph

Considering jumps along the complete graph has been proposed as a way to regularize when there is no actual structural prior information available; see [She10] where it has been studied under the name *clustered Lasso*.

PROPOSITION 10. *For the complete graph K_n, we have $\kappa \gtrsim 1/\sqrt{n}$ and $\rho \lesssim 1/n$.*

PROOF. The bound on κ follows from Lemma 3. To bound ρ, note that we can write the pseudoinverse of the incidence matrix as

$$S = D^\dagger = (D^\top D)^\dagger D^\top.$$

The matrix $D^\top D$ is the graph Laplacian of the complete graph which has the form $nI - \mathbf{1}\mathbf{1}^\top$ from which we can read off its eigenvalues as $\lambda_1 = 0$, $\lambda_i = n$, for $i = 2, \ldots, n$. Choose an eigenbasis $\{v_i\}_{i=1,\ldots,n}$ for $D^\top D$. Then,

$$\|s_j\|_2^2 = \sum_{k=2}^n \frac{1}{\lambda_k^2}\langle v_k, d_j\rangle^2 = \frac{1}{n^2}\sum_{k=2}^n \langle v_k, d_j\rangle^2 \leq \frac{1}{n^2}\|d_j\|_2^2 \leq \frac{2}{n^2}.$$

for all j. □

It yields the following corollary.

COROLLARY 11. *Fix $\delta \in (0,1)$, and let D_{K_n} denote the incidence matrix of the complete graph on n vertices. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat{\theta}$ defined in (1.2) with $\lambda = c\sigma\sqrt{\log(en/\delta)}/n^2$ satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \inf_{\substack{\bar{\theta}\in\mathbb{R}^n \\ T\subseteq[m]}} \left\{ \frac{1}{n}\|\bar{\theta} - \theta^*\|^2 + 4\lambda\|(D_{K_n}\bar{\theta})_{T^c}\|_1 \right\} + \frac{C\sigma^2}{n^2}\left(|T|\log(en/\delta) + \log(e/\delta)\right),$$

with probability at least $1 - 2\delta$. In particular, it yields

$$\mathsf{MSE}(\hat{\theta}) \lesssim \frac{\sigma\|D_{K_n}\theta^*\|_1 \wedge \sigma^2\|D_{K_n}\theta^*\|_0}{n^2}\log(en/\delta)\,.$$

This implies that up to log factors, one performance bound on the TV denoiser for the clique is of the order $|T|/n^2$, where $|T|$ is the number of edges with a jump in the ground truth θ^*. In the case of a signal that takes on $k \ll n$ different values, with $k-1$ of them attained on small islands of size $l \ll n$, this leads to a rate of kl/n, the same we would get for the Lasso if the background value on the complement of the islands was zero.

On the other hand, if there are two large components with different values, $|T|$ will be of the order of n^2, so the result is not informative in this case.

4.2 Star graph

Denote by S_n the star graph on n nodes, having one center node that is connected to $n-1$ leaves. Note that the the question of sparsistency of TV denoising for this graph, together with related ones, has been studied in [OV15] as a way to regularize stratified data.

PROPOSITION 12. *For the star graph S_n, we have $\kappa_T \gtrsim 1/\sqrt{|T|}$ and $\rho \le 1$.*

PROOF. The estimate on κ follows directly from Lemma 3. To compute ρ, observe that

$$D = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{bmatrix}, \quad D^\dagger = \begin{bmatrix} 1/n & 1/n & \dots & 1/n \\ -(n-1)/n & 1/n & \dots & 1/n \\ 1/n & -(n-1)/n & 1/n & \dots 1/n \\ \vdots & & \ddots & \vdots \\ 1/n & 1/n & \dots & -(n-1)/n, \end{bmatrix}$$

so that

$$d_{i,j} = \begin{cases} 1, & j = 1, \\ -1, & i = j-1 \ge 2, \\ 0, & \text{otherwise.} \end{cases} \qquad s_{i,j} = \begin{cases} -\dfrac{n-1}{n}, & i = j+1 \\ \dfrac{1}{n}, & \text{otherwise.} \end{cases},$$

whence the properties of the pseudoinverse can be verified by direct calculation. From this, we can estimate the norm of the columns of S by

$$\|s_j\|^2 = \sum_{i=1}^{n-1}\frac{1}{n^2} + \left(\frac{n-1}{n}\right)^2 = \frac{n^2-n}{n^2} \le 1.$$

□

The following corollary immediately follows.

COROLLARY 13. *Fix $\delta \in (0,1)$, and let $D_\star$ denote the incidence matrix of the star graph on n vertices. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat{\theta}$ defined in (1.2) with $\lambda = c\sigma\sqrt{\log(en/\delta)}/n$ satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \inf_{\substack{\bar{\theta}\in\mathbb{R}^n \\ T\subseteq[m]}} \left\{ \frac{1}{n}\|\bar{\theta} - \theta^*\|^2 + 4\lambda\|(D_\star\bar{\theta})_{T^c}\|_1 \right\} + \frac{C\sigma^2}{n}\left(|T|^2\log(en/\delta) + \log(e/\delta)\right),$$

with probability at least $1 - 2\delta$. In particular, it yields

$$\mathsf{MSE}(\hat{\theta}) \lesssim \frac{\sigma\|D_\star\theta^*\|_1 \wedge \sigma^2\|D_\star\theta^*\|_0^2}{n}\log(en/\delta).$$

The star graph leads to a useful regularization when most of the outer nodes take the same value as the central node and only a few outer nodes take a different one. Specifically, let 1 denote the central vertex and consider the set $\Theta^\star(s) \subset \mathbb{R}_n$ defined for any integer $s \in [n-1]$ by

$$\Theta^\star(s) = \left\{\theta \in \mathbb{R}^n \,:\, \sum_{j=2}^n \mathbb{I}(\theta_j \neq \theta_1) \le s\right\}.$$

Then it holds

$$\sup_{\theta^*\in\Theta^\star(s)} \frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \lesssim \frac{\sigma^2 s^2}{n}\log(en/\delta)$$

with probability at least $1 - 2\delta$.

4.3 Random graphs

In the case of random graphs, it was noted in [WSST15] that one can bound ρ if one has bounds on the second smallest eigenvalue of the Laplacian of the graph. We can slightly improve on their estimation of ρ.

PROPOSITION 14. *Suppose G is a connected graph whose Laplacian admits an eigenvalue decomposition $D^\top D = V\Lambda V^\top$, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, $V = [v_1, \dots, v_n]$, $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$.*

If the graph Laplacian has a spectral gap, i.e., *there exists a constant $c_1 > 0$ such that $\lambda_2 \ge c_1$, then $\rho \le \sqrt{2}/c_1$.*

PROOF. Writing $D^\dagger = [s_1, \dots, s_m] = (D^\top D)^\dagger D^\top$, $D^\top = [d_1, \dots, d_m]$, we note that the columns d_j have 2-norm $\|d_j\|_2 = \sqrt{2}$ because D is the incidence matrix of a graph, so

$$\|s_j\|_2^2 = \sum_{k=2}^n \frac{1}{\lambda_k^2}\langle v_k, d_j\rangle^2 \le \frac{1}{\lambda_2^2}\sum_{k=2}^n \langle v_k, d_j\rangle^2 \le \frac{1}{\lambda_2^2}\|d_j\|_2^2 \le \frac{2}{c_1^2}.$$

□

We can combine this with bounds on the spectral gap of two families of random graphs, Erdős-Rényi random graphs $G(n,p)$ and random regular graphs. Both of these models exhibit a spectral gap of the order $O(d)$ in a regime where the degree increases logarithmically with the number of vertices, see [KOV14] and [Fri04], respectively. Together with the bound on κ from Lemma 3, we get the following rate.

COROLLARY 15. *Fix $\delta \in (0,1)$ and let D denote the incidence matrix of either a random d-regular graph with $d_n = d_0(\log n)^\beta$ or Erdős-Rényi random graph with $G(n,p)$ with $p_n = d_n/n$ for some constant $\beta > 0$ and $d_0 > 1$. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat{\theta}$ defined in* (1.2) *with $\lambda = c\sigma\sqrt{\log(ed_n n/\delta)}/(d_n n)$ satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \inf_{\substack{\bar{\theta}\in\mathbb{R}^n \\ T\subseteq[m]}} \left\{ \frac{1}{n}\|\bar{\theta} - \theta^*\|^2 + 4\lambda\|(D\bar{\theta})_{T^c}\|_1 \right\} + \frac{C\sigma^2}{d_n n}\left(|T|\log(en/\delta) + \log(e/\delta)\right),$$

with probability at least $1 - 2\delta$ over ε and with high probability over the realizations of the graph. In particular, it yields

$$\mathsf{MSE}(\hat{\theta}) \lesssim \frac{\sigma^2\|D\theta^*\|_0}{d_n n}\log(en/\delta)$$

where $\|D\theta^*\|_0$ denotes the number of nonzero components of $D\theta^*$.

In the context of TV denoising, Erdős-Rényi random graphs with expected degree d_n can be considered a sparsification of the complete graph considered in Section 4.1. In the same model considered in Subsection 4.1 of k islands with l nodes each, we would get a performance rate of kl/n, the same as before. On the other hand, the underlying graph is much sparser, so we could possibly get a computational benefit from choosing it instead of the complete graph. The behavior of random graphs is compared to that of the complete graph in Section A in the appendix. They indicate that the computational saving occur at a negligible statistical cost.

4.4 Power graph of the cycle

In practice, nearest neighbor graphs often arise in the context of spatial regularization. The grid is one such example and as an extension, we consider the kth power of the cycle graph as a toy example to study the effect of increasing the connectivity of the graph.

Define the cycle graph C_n to be the graph on n vertices with $i \sim j$ if and only if $i - j \equiv \pm 1 \bmod n$ and its kth power graph C_n^k as the graph with the same vertex set but with $i \sim j$ if and only if there is a path of length at most k from i to j in C_n.

PROPOSITION 16. *For $G = C_n^k$ where $k \le n/2$, $\rho \lesssim \sqrt{n}/k^3 + 1$ and $\kappa \gtrsim 1/\sqrt{k}$.*

PROOF. The bound on κ follows from Lemma 3 and the fact that the degree of C_n^k is bounded by $2k$.

To bound ρ, write $D^\dagger = [s_1, \dots, s_m] = (D^\top D)^\dagger D^\top$ and $D^\top = [d_1, \dots, d_m]$ and use the same technique and notation as in the proof of Proposition 4 in Subsection B.2. The Laplacian of C_n^k has the form of a circulant matrix whose first row is

$$a = \big[2k \underbrace{\ -1 \ \dots \ -1}_{k \text{ times}} \ 0 \ \dots \ 0 \underbrace{\ -1 \ \dots \ -1}_{k \text{ times}}\big].$$

Hence, we can choose the discrete Fourier basis $(v_m)_j = \exp(2\pi \mathrm{i} mj/n)$, $m, j \in [\![0, n[\![$ as an eigenbasis. The eigenvalues are given by

$$\lambda_m = \sum_{l=0}^{n-1} e^{2\pi \mathrm{i} ml/n} a_l = 2k - \sum_{l=1}^{k}\left(e^{2\pi \mathrm{i} lm/n} + e^{-2\pi \mathrm{i} lm/n}\right) = 2\sum_{l=1}^{k}\left(1 - \cos\left(\frac{2\pi lm}{n}\right)\right).$$

By the formula for the sums of squares,

$$\sum_{l=1}^{k} l^2 = \frac{1}{6}k(k+1)(2k+1) \geq \frac{1}{3}k^3\,,$$

and using the same estimates for the cosine as in Subsection B.2, $2-2\cos x \geq x^2/2$ for $x \in [0,1/2]$, and $2-2\cos x \geq 0.1$ for $x \in [1/2,\pi]$, we see that for $2\pi lm/n \geq 1/2$,

$$2\sum_{l=1}^{k}\left(1-\cos\left(\frac{2\pi lm}{n}\right)\right) \geq \frac{1}{2}\sum_{l=1}^{k}\left(\frac{2\pi lm}{n}\right)^2 \geq \frac{k^3}{6}\left(\frac{2\pi m}{n}\right)^2 .$$

Moreover, by the Lipschitz continuity of the exponential,

$$|\langle v_m, d_j\rangle|^2 = \frac{1}{n}\left|e^{2\pi \mathrm{i} m(j+1)/n} - e^{2\pi \mathrm{i} mj/n}\right|^2 \leq \frac{4m^2\pi^2}{n^3}\,.$$

By expressing the norm of the columns of $D^\dagger$ in terms of the eigendecomposition and combining pairs eigenvalues with the same value which have the same eigenvectors up to a sign in the exponential, we finally get

$$\begin{aligned}
\|s_j\|_2^2 &= \sum_{m=1}^{n-1}\frac{1}{\lambda_m^2}\langle v_m, d_j\rangle^2 = \sum_{m=1}^{n-1}\frac{1}{\lambda_m^2}\langle v_m, d_j\rangle^2 \\
&\leq 8\frac{\pi^2}{n^3}\sum_{m=1}^{\lceil n-1/2\rceil} m^2\left(2\sum_{l=1}^{k}\left(1-\cos\left(\frac{2\pi lm}{n}\right)\right)\right)^{-2}\left(\mathbb{1}(2\pi km/n \leq 1/2) + \mathbb{1}(2\pi km/n > 1/2)\right) \\
&\lesssim n\sum_{m=1}^{\lceil n-1/(8\pi k)\rceil}\frac{1}{m^2k^6} + \frac{1}{n^3}\sum_{m=1}^{n} m^2 \lesssim \frac{n}{k^3} + 1 \lesssim \frac{n}{k^6} + 1\,.
\end{aligned}$$

□

COROLLARY 17. *Fix $\delta \in (0,1)$ and let D denote the incidence matrix of C_n^k. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat\theta$ defined in (1.2) with $\lambda = c\sigma\sqrt{\log(en/\delta)}/(\sqrt{n}k^3 \wedge n)$ satisfies*

$$\frac{1}{n}\|\hat\theta - \theta^*\|^2 \leq \inf_{\substack{\bar\theta\in\mathbb{R}^n \\ T\subseteq[m]}}\left\{\frac{1}{n}\|\bar\theta-\theta^*\|^2 + 4\lambda\|(D\bar\theta)_{T^c}\|_1\right\} + C\sigma^2\left(\frac{1}{k^5}\vee\frac{k}{n}\right)\left(|T|\log(en/\delta) + \log(e/\delta)\right),$$

with probability at least $1-2\delta$. In particular, it yields

$$\mathsf{MSE}(\hat\theta) \lesssim \sigma^2\|D\theta^*\|_0\left(\frac{1}{k^5}\vee\frac{k}{n}\right)\log(en/\delta)$$

where $\|D\theta^*\|_0$ denotes the number of nonzero components of $D\theta^*$.

5. APPLICATIONS

The rate for the grid obtained in Corollaries 5 and 7 can be used to derive rates for function estimation in dimension $d \geq 2$. This allows us to generalize the results [DHL14, Proposition 6] for

the adaptive estimation of Hölder functions and [CGS15, Bel15] for the estimation of bi-isotonic matrices. Moreover, we can also generalize to piecewise Hölder functions, called "cartoon images".

In the first two subsections, we are interested in real valued functions on $[0,1]^d$. To relate function estimation to our problem, consider the vectors θ to be a discretization of a continuous signal $f\colon [0,1]^d \to \mathbb{R}$ on the regular grid $\mathcal{X}_N^d := \{x_{\boldsymbol{i}} := \boldsymbol{i}/N : \boldsymbol{i} \in [N]^d\}$, so $\theta_{\boldsymbol{i}} = f(x_{\boldsymbol{i}}) = f(i_1/N, \dots, i_d/N)$, $\boldsymbol{i} \in [N]^d$. Furthermore, for any function, $f\colon [0,1]^d \to \mathbb{R}$, define the pseudo-norm $\|f\|_n$ by

$$\|f\|_n^2 = \frac{1}{n} \sum_{\boldsymbol{i} \in [N]^d} f(x_{\boldsymbol{i}})^2 \,.$$

5.1 Hölder functions

In [DHL14, Proposition 7], the authors showed that the TV denoiser in one dimension achieves the minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$ for estimating Hölder continuous functions (with parameter $\alpha \in (0,1]$) on the line up to logarithmic factors. Here, we show that the TV denoiser achieves a rate of $n^{-\frac{2\alpha}{d\alpha+d}}$, again up to logarithmic factors, which means it is near minimax for two-dimensional observations as well. Unlike the one-dimensional result of [DHL14], the TV denoiser in two dimensions is adaptive to the unknown parameter α.

Definition 18 (Hölder function). *For $\alpha \in (0,1]$, $L > 0$, we say that a function $f\colon [0,1]^d \to \mathbb{R}$ is* (α, L)-Hölder continuous *if it satisfies*

$$|f(y) - f(x)| \le L\|x - y\|_\infty^\alpha \quad \textit{for all } x, y \in [0,1]^d.$$

For such an f, we write $f \in H(\alpha, L)$.

Note that we picked the ℓ_∞-norm for convenience here. By the equivalence of norms in finite dimensions, the ℓ_2-norm would yield the same definition up to a dimension-dependent constant. Moreover, for samples of a Hölder continuous function on a grid, Definition 18 implies

$$|\theta_{\boldsymbol{i}} - \theta_{\boldsymbol{j}}| \le LN^{-\alpha}\|\boldsymbol{i} - \boldsymbol{j}\|_\infty^\alpha, \tag{5.5}$$

so we can directly work with the ℓ_∞-distance between the indices.

Proposition 19. *Fix $\delta \in (0,1)$, $d \ge 2$, $L > 0$, $N \ge 1$, $n = N^d$ and $\alpha \in (0,1]$ and let y be sampled according to the Gaussian sequence model (1.1), where $\theta^*_{\boldsymbol{i}} = f^*(x_{\boldsymbol{i}})$, $\boldsymbol{i} \in [N]^d$ for some unknown function $f^* : [0,1]^d \to \mathbb{R}$. There exist positive constants c, C and $C' = C'(\sigma, L, d)$ such that the following holds. Let $\hat\theta$ be the TV denoiser defined in (1.2) for the N^d grid with incidence matrix D_d and tuning parameter $\lambda = c\sigma\sqrt{r_d(n)\log(en/\delta)}/n, c > 0$ where $r_2(n) = \log n$ and $r_d(n) = 1$ for $d \ge 3$. Moreover, let $\hat f : [0,1]^d \to \mathbb{R}$ be defined by $\hat f(x_{\boldsymbol{i}}) = \hat\theta_{\boldsymbol{i}}$ for $\boldsymbol{i} \in [N]^d$ and arbitrarily elsewhere on the unit hypercube $[0,1]^d$.*

Further, assume that $N \ge C'(L, \sigma, d)\sqrt{r_d(n), \log(en/\delta)}$. Then,

$$\|\widehat f - f^*\|_n^2 \le \inf_{\bar f \in H(\alpha, L)} \big\{\|\bar f - f^*\|_n^2\big\} + C\frac{\big(L^2(\sigma\sqrt{r_d(n)\log(en/\delta)})^{2\alpha}\big)^{\frac{1}{\alpha+1}}}{n^{\frac{2\alpha}{d\alpha+d}}} + C\frac{\sigma^2}{n}\log(e/\delta)\,,$$

with probability at least $1 - 2\delta$. In particular, for $d = 2$, it yields the near optimal rate

$$\|\widehat f - f^*\|_n^2 \le \inf_{\bar f \in H(\alpha, L)} \big\{\|\bar f - f^*\|_n^2\big\} + C\big(L^2(\sigma\log(en/\delta))^{2\alpha}\big)^{\frac{1}{\alpha+1}} n^{-\frac{2\alpha}{2\alpha+2}} + C\frac{\sigma^2}{n}\log(e/\delta)\,,$$

PROOF. Throughout this proof, it will be convenient to identify a function g to the vector $(g(x_{\boldsymbol{i}}), \boldsymbol{i} \in [N]^d)$. We use (3.4) to get that for any vector $\bar{f} \in \mathbb{R}^{N^d}$, it holds

$$\frac{1}{n}\|\widehat{f} - f^*\|^2 \le \frac{1}{n}\|\bar{f} - f^*\|_2^2 + 4\lambda\|D\bar{f}\|_1 + C\frac{\sigma^2}{n}\log(e/\delta)\,. \tag{5.6}$$

Denote by $\Theta(\alpha, L)$ the set of vectors on the grid that satisfy (5.5) and observe that it is a closed convex set so that $f_{\mathrm{proj}} = \mathrm{argmin}_{\theta\in\Theta(\alpha,L)}\|\theta - f^*\|^2$ is uniquely defined. Moreover,

$$\|\bar{f} - f^*\|^2 \le \|f_{\mathrm{proj}} - \bar{f}\|^2 + \|f_{\mathrm{proj}} - f^*\|^2,$$

which plugged back into (5.6) yields

$$\frac{1}{n}\|\widehat{f} - f^*\|^2 \le \frac{1}{n}\inf_{f\in H(\alpha,L)}\|f - f^*\|^2 + \left\{\frac{1}{n}\|\bar{f} - f_{\mathrm{proj}}\|^2 + 4\lambda\|D\bar{f}\|_1\right\} + C\frac{\sigma^2}{n}\log(e/\delta)\,,$$

The remainder of the proof consists in choosing $\bar{f}$ to balance the approximation error and the stochastic error.

Fix an integer k to be determined later and for any $\boldsymbol{i} \in [N]^d$, define $a_{\boldsymbol{i}} = k\lfloor \boldsymbol{i}/k\rfloor$. Next, define a piecewise constant approximation $\bar{f}$ to f_{proj} by $\bar{f}_{\boldsymbol{i}} = (f_{\mathrm{proj}})_{a_{\boldsymbol{i}}}$ for $\boldsymbol{i} \in [N]^d$.

We first control the approximation error for all $\boldsymbol{i} \in [N]^d$ as follows:

$$|f_{\mathrm{proj}}(\boldsymbol{i}) - \bar{f}(\boldsymbol{i})| = |f_{\mathrm{proj}}(\boldsymbol{i}) - f_{\mathrm{proj}}(a_{\boldsymbol{i}})| \le LN^{-\alpha}\|\boldsymbol{i} - a_{\boldsymbol{i}}\|_\infty^\alpha \le L(k/N)^\alpha\,.$$

It yields

$$\frac{1}{n}\|\bar{f} - f_{\mathrm{proj}}\|_2^2 \le L^2(k/N)^{2\alpha}\,.$$

Next, we control the term $\|D\bar{f}\|_1$. To that end, observe that if $\boldsymbol{i}$ and $\boldsymbol{i'}$ are neighbors in the grid, then

$$|\bar{f}_{\boldsymbol{i}} - \bar{f}_{\boldsymbol{i'}}| \le LN^{-\alpha}\mathbb{1}(a_{\boldsymbol{i}} \ne a_{\boldsymbol{i'}})$$

Therefore

$$\|D\bar{f}\|_1 \le L(k/N)^\alpha \sum_{\boldsymbol{i}\sim\boldsymbol{i'}} \mathbb{1}(a_{\boldsymbol{i}} \ne a_{\boldsymbol{i'}}) \le L(k/N)^\alpha 2dk^{d-1}\Big(\frac{N}{k}\Big)^d = 2dL\frac{N^d}{k^{1-\alpha}N^\alpha}\,.$$

Hence

$$\lambda\|D\bar{f}\|_1 \lesssim \frac{L\sigma}{k^{1-\alpha}N^\alpha}\sqrt{r_d(n)\log(en/\delta)}$$

Choosing now

$$M = \left(\frac{\sigma N^\alpha\sqrt{r_d(n)\log(en/\delta)}}{L}\right)^{\frac{1}{\alpha+1}}, \quad k = \lceil M\rceil\,,$$

yields the desired result, taking into account that $M \in [1, N]$ if we assume

$$N \ge \left(\frac{L}{\sigma\sqrt{r_d(n)\log(en/\delta)}}\right)^{1/\alpha} \vee \frac{\sigma\sqrt{r_d(n)\log(en/\delta)}}{L}\,.$$

□

Unlike [DHL14, Proposition 7], this result does not require the knowledge of L or α to compute the tuning parameter λ, but only the noise level σ. As a result, the estimator is therefore adaptive to the smoothness of the underlying function. This effect comes from better estimates on ρ than in the 1D case.

[Mv97] have shown that asymptotically, TV regularization together with spline regression achieves the minimax rate for the estimation of k-times differentiable functions in 2D. Our result however holds for finite sample size and fractional smoothness, albeit only for $\alpha \in (0, 1]$.

It is not surprising that our results are suboptimal for $d \geq 3$. Indeed, penalizing by the size of jumps is not appropriate for Hölder functions. One should rather penalize by the number of blocks. It is merely a coincidence that in two dimensions this method leads to optimal and adaptive rates for Hölder functions.

5.2 Piecewise constant and piecewise Hölder functions

Recall that Corollary 5 allows us to get scale free results, i.e., bounds that do not scale with jump height. It is therefore well suited to detect sharp boundaries, one of the features often associated with total variation regularization. To formalize this point, we analyze two models that involve a boundary, namely piecewise constant and piecewise smooth signals. The framework below largely builds upon [WNC05].

First, let us define the box-counting dimension of a set, which we will use as the measure of the complexity of the boundary.

Definition 20 (Box-counting dimension). *Let $B \subseteq [0,1]^d$ be a set and denote by $N(r)$ the minimum number of (Euclidean) balls of radius r required to cover B. The* box-counting dimension *of B is defined as*

$$\dim_{\text{box}}(B) := \limsup_{r \to 0} \frac{\log N(r)}{\log(1/r)}.$$

The box-counting dimension generalizes the notion of linear dimension. For instance, if B is a smooth d_0-dimensional manifold, then its box-counting dimension is equal to d_0.

Definition 21 (Piecewise constant functions). *For $\beta > 0$, we call a function $f \colon \mathbb{R}^d \to \mathbb{R}$* piecewise constant *and write $f \in PC(d, \beta)$ if there is an associated boundary set $B(f)$ such that:*

1. *The function f is locally constant on $[0,1]^d \setminus B(f)$,* i.e., *for all $x \in [0,1]^d \setminus B(f)$, there is an $\varepsilon > 0$ such that for all y with $\|y - x\| < \varepsilon$, $f(x) = f(y)$.*
2. *The boundary set $B(f)$ has covering number $N(r) \leq \beta r^{-(d-1)}$ for some $\beta > 0$. In particular, $\dim_{\text{box}}(B) \leq d - 1$.*

Definition 22 (Piecewise Hölder functions). *For $\alpha \in (0, 1]$ and $\beta, L > 0$, we call a function $f \colon \mathbb{R}^d \to \mathbb{R}$* piecewise Hölder *and write $f \in PH(d, \beta, \alpha, L)$ if there is an associated boundary set $B(f)$ such that:*

1. *f is locally L-Hölder on $[0,1]^d \setminus B(f)$,* i.e., *for all $x \in [0,1]^d \setminus B(f)$, there is an $\varepsilon > 0$ such that for all y with $\|y - x\| < \varepsilon$, $|f(x) - f(y)| \leq L\|x - y\|_\infty^\alpha$.*
2. *$B(f)$ has box-counting dimension at most $d - 1$, and its covering number $N(r)$ is bounded by $N(r) \leq \beta r^{-(d-1)}$.*

One intuition behind these definitions is to consider the signal as a "cartoon image" containing large patches that are constant or fairly smooth, split by sharp boundaries.

Using Corollary 5, we can now establish estimation rates for these classes of functions.

PROPOSITION 23. *Fix $\delta \in (0,1)$, $d \geq 2$, $N \geq 1$, $n = N^d$ and let y be sampled according to the Gaussian sequence model (1.1), where $\theta^*_{\boldsymbol{i}} = f^*(x_{\boldsymbol{i}})$, $\boldsymbol{i} \in [N]^d$ for some unknown function $f^* \in PC(d,\beta)$. There exist positive constants c and C such that the following holds. Let $\hat{\theta}$ be the TV denoiser defined in (1.2) for the d-dimensional grid with incidence matrix D_d and tuning parameter $\lambda = c\sigma\sqrt{r_d(n)\log(en/\delta)}/n$, where $r_2(n) = \log n$ and $r_d(n) = 1$ for $d \geq 3$. Moreover, let $\hat{f} : [0,1]^d \to \mathbb{R}$ be defined by $\hat{f}(x_{\boldsymbol{i}}) = \hat{\theta}_{\boldsymbol{i}}$ for $\boldsymbol{i} \in [N]^d$ and arbitrarily elsewhere on the unit hypercube $[0,1]^d$. Then,*

$$\|\widehat{f} - f^*\|_n^2 \lesssim \frac{\sigma^2\beta}{n^{1/d}} r_d(n)\log(en/\delta) + \frac{\sigma^2}{n}\log(e/\delta)\,,$$

with probability at least $1 - 2\delta$.

PROOF. For any $x \in \mathbb{R}^d$, and any closed set $B \subset \mathbb{R}^d$ define the distance from x to B by $d(x,B) = \min_{b\in B}\|x - b\|$. Next define

$$T := \{(\boldsymbol{i},\boldsymbol{j}) : \boldsymbol{i} \sim \boldsymbol{j} \text{ and } \min\{d(x_{\boldsymbol{i}}, B(f)), d(x_{\boldsymbol{j}}, B(f))\} \leq 4/N\}$$

to be the set of edges whose nodes are close to the boundary $B(f)$. It can be readily checked that the vector $\big(f(x_{\boldsymbol{i}}), \boldsymbol{i} \in V\big)$ is constant on the connected components of $(V, E \setminus T)$.

First, let us state a lemma that allows us to bound the number of grid points in a neighborhood of a set by the volume of said set.

LEMMA 24. *[ACSW12, Lemma 8.3] Let $B \subseteq [0,1]^d$, $A = [0,1]^d \cap \big(B + \mathcal{B}(\eta)\big)$, $4/N \leq \eta \leq 1$. Then, the number of grid points on a regular d-dimensional grid $\mathcal{X}_N^d$ intersecting A is bounded by*

$$8^{-d}N^d \operatorname{vol}(A) \leq |A \cap \mathcal{X}_N^d| \leq 4^d N^d \operatorname{vol}(A).$$

By Lemma 24, Definition 20 and the triangle inequality, we get

$$\begin{aligned} |T| &\leq \big|\mathcal{X}_N^d \cap \big(B(f) + \mathcal{B}(4/N)\big)\big| \leq 4^d N^d \operatorname{vol}(B(f) + \mathcal{B}(4/N)) \\ &\leq 4^d N^d \beta \operatorname{vol}(B(1))(8/N)^d (8/N)^{-(d-1)} \leq C(d)\beta N^{d-1}\,, \end{aligned} \tag{5.7}$$

where $C(d)$ is a dimension-dependent constant.

Since f^* is constant along all edges not included in T, $\|(Df^*)_{T^c}\|_1 = 0$. Taking into account $n = N^d$, Corollaries 5 and 7 readily yield the desired result. □

Combining the results for piecewise constant and Hölder smooth functions, we can get the following extension to piecewise smooth functions.

PROPOSITION 25. *Fix $\delta \in (0,1)$, $d \geq 2$, $N \geq 1$, $n = N^d$ and let y be sampled according to the Gaussian sequence model (1.1), where $\theta^*_{\boldsymbol{i}} = f^*(x_{\boldsymbol{i}})$, $\boldsymbol{i} \in [N]^d$ for some unknown function $f^* \in PH(d,\beta,\alpha,L)$, $\alpha \in (0,1]$, $L > 0$, $\beta > 0$. There exist positive constants c, C and $C' = C'(\sigma, L, d)$ such that the following holds. Let $\hat{\theta}$ be the TV denoiser defined in (1.2) for the d-dimensional grid with incidence matrix D_d and tuning parameter $\lambda = c\sigma\sqrt{r_d(n)\log(en/\delta)}/n$, where $r_2(n) = \log n$ and $r_d(n) = 1$ for $d \geq 3$. Moreover, let $\hat{f} : [0,1]^d \to \mathbb{R}$ be defined by $\hat{f}(x_{\boldsymbol{i}}) = \hat{\theta}_{\boldsymbol{i}}$ for $\boldsymbol{i} \in [N]^d$ and arbitrarily elsewhere on the unit hypercube $[0,1]^d$.*

If $N \geq C'(L, \sigma, d)\sqrt{r_d(n)\log(en/\delta)}$, *then*

$$\frac{1}{n}\|\widehat{f} - f^*\|^2 \lesssim \frac{\left(L^2(\sigma\sqrt{r_d(n)\log(en/\delta)})^{2\alpha}\right)^{\frac{1}{\alpha+1}}}{n^{\frac{2\alpha}{d\alpha+d}}} + \frac{\sigma^2\beta}{n^{1/d}} r_d(n)\log(en\delta) + \frac{\sigma^2}{n}\log(e/\delta)\,,$$

with probability at least $1 - 2\delta$.

Proof. As in the proof of Proposition 23, in Corollaries 5 and 7, set

$$T := \{(\boldsymbol{i}, \boldsymbol{j}) : \boldsymbol{i} \text{ neighbor of } \boldsymbol{j} \text{ and } d(x_{\boldsymbol{i}}, B(f)) \wedge d(x_{\boldsymbol{j}}, B(f)) \leq 4/N\},$$

and note that $|T| \leq C(d)\beta N^{d-1}$, using the same argument as in (5.7). Moreover, it can be readily checked that the vector $\big(f(x_{\boldsymbol{i}}), \boldsymbol{i} \in V\big)$ satisfies the Hölder condition (5.5) on the connected components of $(V, E \setminus T)$.

Next, we adopt the same discretization of as in Proposition 19, with a slight modification to take into account that f^* is only Hölder-continuous within connected components of the underlying grid. To that end, fix an integer k to be determined later and for any $\boldsymbol{i} \in [N]^d$, define indices $a_{\boldsymbol{i}}$ and corresponding boxes $A_{\boldsymbol{i}}$ by

$$(a_i)_j = \begin{cases} k\lfloor i_j/k \rfloor, & i_j \leq N, \\ N, & i_j = N+1, \end{cases} \qquad A_{\boldsymbol{i}} = [\![a_{i_1}, a_{i_1+1}[\![\times \cdots \times [\![a_{i_d}, a_{i_d+1}[\![\,.$$

For each of the boxes A and every connected component C of $(V, E \setminus T)$ within, pick a fixed representative $b(C)$ and write $C(\boldsymbol{i})$ for the connected component in $A_{\boldsymbol{i}}$ that $\boldsymbol{i}$ belongs to. Next, define a piecewise constant approximation $\bar{f}$ to f^* by $\bar{f}_{\boldsymbol{i}} = f^*_{b(C(\boldsymbol{i}))}$ for $\boldsymbol{i} \in [N]^d$.

Using the same arguments as in the proof of Proposition 19, we get first that $n^{-1}\|\bar{f} - f^*\|_2^2 \leq L^2(k/N)^{2\alpha}$ and second that

$$\|(D\bar{f})_{T^c}\|_1 \leq 2dL\frac{N^d}{k^{1-\alpha}N^\alpha}\,.$$

Choosing now

$$k = \left\lceil \left(\frac{\sigma N^\alpha \sqrt{r_d(n)\log(en/\delta)}}{L}\right)^{\frac{1}{\alpha+1}} \right\rceil$$

and applying Corollaries 5 and 7 yields the desired result. □

For a Lipschitz boundary in two dimensions, this matches the minimax bound $n^{-2\alpha/(2\alpha+2)} \vee n^{-1/2}$ for boundary fragments in [KT93, Theorem 5.1.2] up to logarithmic factors. However, unlike the framework [KT93], our techniques do not allow an improvement of the bound for smoother boundaries parametrization because $|T|$ will always be of the order $O(N^{d-1})$. On the other hand, unlike the algorithms in [KT93] and [ACSW12], our analysis allows for any jump sizes, so TV regularization automatically adapts to both $B(f)$ and α.

5.3 Bi-isotonic matrices

In our final example, we consider two-dimensional signals that increase in both directions, sometimes referred to *bi-isotonic*. The class of bi-isotonic matrices is defined as follows,

$$\mathcal{M} := \big\{\theta \in \mathbb{R}^{N\times N} : \theta_{j_1,j_2} \geq \theta_{i_1,i_2} \text{ if } j_1 \geq i_1 \text{ and } j_2 \geq i_2\big\}$$

Recently, [CGS15, Bel15] showed that the least squares estimator for $\mathcal{M}$ yields the near minimax rate $\sqrt{D(\theta^*)/n}(\log n)^4$, where $D(\theta^*) := (\theta^*_{N,N} - \theta^*_{1,1})^2$ denotes the square variation of the matrix.

In the following, we show that the 2D TV denoiser can match this rate and that it also improves on the exponent of the log factors.

PROPOSITION 26. *Let y be a sample of the Gaussian sequence model (1.1), $\delta \in (0,1)$ and denote by $\theta^\uparrow$ the projection of θ^* onto $\mathcal{M}$. Fix $\delta \in (0,1)$ and let $\hat{\theta}$ denote the TV denoiser on the 2D grid defined in (1.2) with $\lambda = c\sigma\sqrt{(\log n)\log(en/\delta)}/n$. Then there exists a constant $C > 0$ such that*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \frac{1}{n}\|\theta^\uparrow - \theta^*\|^2 + C\sigma\sqrt{\frac{(\log n)\log(n/\delta)}{n}}\sqrt{D(\theta^\uparrow)} + C\frac{\sigma^2}{n}\log(e/\delta),$$

with probability at least $1 - 2\delta$.

PROOF. We use the slow rate version of (3.4) for $\bar{\theta} = \theta^\uparrow$,

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \frac{1}{n}\|\theta^\uparrow - \theta^*\|^2 + 4\lambda\|D\theta^\uparrow\|_1 + \frac{C\sigma^2}{n}\log(e/\delta). \tag{5.8}$$

Because $\theta^\uparrow$ is bi-isotonic, summing along the rows yields

$$\sum_{i=1}^{N}\sum_{j=1}^{N-1}|\theta^\uparrow_{i,j+1} - \theta^\uparrow_{i,j}| \le \sum_{i=1}^{N}(\theta^\uparrow_{i,N} - \theta^\uparrow_{i,1}) \le N(\theta^\uparrow_{N,N} - \theta^\uparrow_{1,1}),$$

and similarly along columns, which combined gives us

$$\|D\theta^\uparrow\|_1 \le 2N(\theta^\uparrow_{N,N} - \theta^\uparrow_{1,1}) = 2\sqrt{nD(\theta^\uparrow)}$$

Plugging this into (5.8), together with inserting the value of λ, we have

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \frac{1}{n}\|\theta^\uparrow - \theta^*\|^2 + C\sigma\sqrt{\frac{(\log n)\log(n/\delta)}{n}}\sqrt{D(\theta^\uparrow)} + C\frac{\sigma^2}{n}\log(e/\delta),$$

for some $C > 0$. □

We recover the results of [CGS15, Bel15] with a smaller exponent in the logarithmic factor. On the other hand, the TV-denoiser requires an estimate for σ (or at least an upper bound), unlike the least squares estimator, which does not require any tuning.

Note further that our rate scales with σ rather than σ^2 in [CGS15, Bel15]. This is because we use a "slow rate" bound.

Unlike [CGS15, Bel15] we do not show that our estimator adapts to the number of rectangles on which the matrix is piecewise constant. In particular, they show that if the number of such rectangles is a constant, then the least squares estimator achieves a fast rate of order $\sigma^2(\log n)^8/n$. This is not the case in the present paper. Indeed, the TV denoiser is not the correct tool for that. Even in the case of two rectangles, the number of active edges on the 2D grid is already linear in N leading to rates that are slower than $\sigma^2/N \gg \sigma^2(\log n)^8/n$. Nevertheless, it is not hard to show that if θ^* is an $N \times N$ matrix with a triangular structure the form $\theta^*_{ij} = \mathbb{1}(i \ge j)$, then this matrix

is well approximated by N rectangles. In this case, the results of [CGS15, Bel15] yield a bound for the least squares estimator $\hat{\theta}^{\text{LS}}$ of the form

$$\frac{1}{n}\|\hat{\theta}^{\text{LS}} - \theta^*\|^2 \leq C\sigma^2 \frac{(\log n)^8}{\sqrt{n}}$$

and it is not hard to see that the TV denoiser yields

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \leq C(\sigma \wedge 1)^2 \frac{(\log n)^2}{\sqrt{n}}$$

where both results are stated with large but constant probability (say 99%). It is not excluded that the least squares estimator still achieves faster rates in this case but the currently available results do not lead to better rates.

Finally, note that unlike [DHL14, Proposition 6], λ does not have to depend on $D(\theta^{\uparrow})$ here because of the better behavior of ρ for the 2D grid.

Acknowledgments We would like to thank Vivian Viallon for bringing [SSR12] to our attention. We thank the participants in the workshop "Computationally and Statistically Efficient Inference for Complex Large-scale Data", that took place in Oberwolfach on March 6–12, 2016; in particular Axel Munk for pointers to the literature on the spectral decomposition of the Toeplitz matrix in (B.12) and Alessandro Rinaldo for interesting discussion. Finally, we thank Ryan Tibshirani for pointing us to the paper [WSST15] and discussing his results with us. A short version of this paper was already submitted when we became aware of the prior work [WSST15]. We are grateful to Ryan for being forthcoming in dealing with this delicate situation and being willing to look at a preliminary version of the present paper to help us assess the differences between the two works.

APPENDIX A: NUMERICAL EXPERIMENTS

In order to illustrate our findings in Subsections 4.1 and 4.3, we used the TV denoiser implementation from [XKWG14].

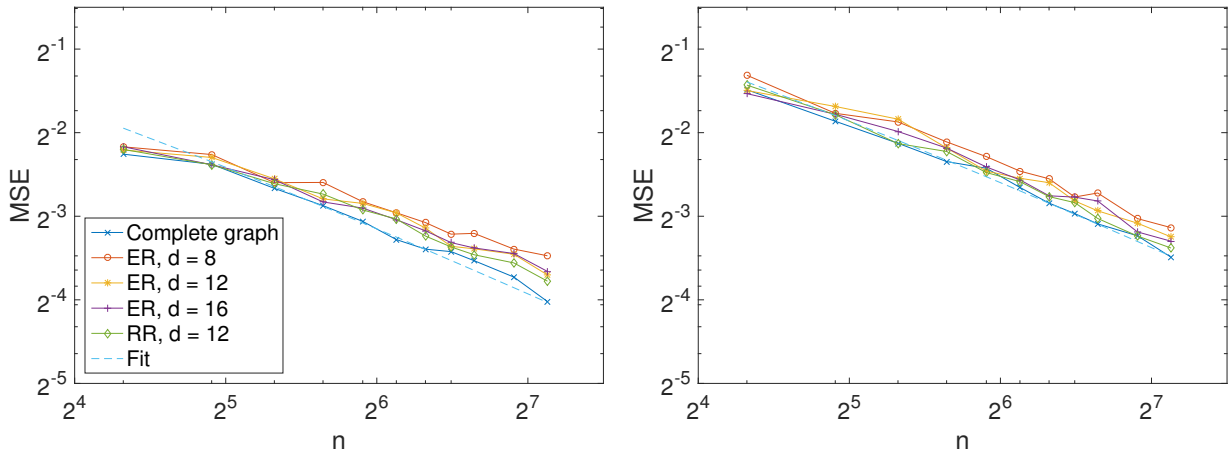
The Island model. Consider a partition of $[n]$ into k blocks $B_1, \dots, B_k$ of size $|B_j| = l, l \in [k]$ and a block B_0 of size $|B_0| = n - kl$. We focus on cases where $n \gg kl$ and we call block B_0, the *background component* and the blocks $B_j, j \in k$ are called *islands*. The unknown parameter θ^* has coordinates $\theta^*_i = 50 + 10j, i \in B_j, j \in k$, and $\theta^*_l = 50, i \in B_0$.

Graphs. We consider three types of graphs to determine our penalty structure: the complete graph, the Erdős-Rényi random graph with expected degree d and the random d-regular graph[1] for different values of d. Note that in the case of the random graphs, we refer to a realization from a given distribution as "the" random graph.

Choice of λ. We consider two choices for the regularization parameter λ: the fixed choice, denote by λ_{th} dictated by our theoretical results and an oracle choice λ_{or} on a geometric grid, obtained by $\lambda_{\text{or}} = 10\lambda_{\text{th}}\beta^{j^*}$ where j^* is the smallest $j \geq 1$ such that $\|\hat{\theta}(10\lambda_{\text{th}}\beta^{j^*+1}) - \theta^*\|_2 \geq \|\hat{\theta}(10\lambda_{\text{th}}\beta^{j^*}) - \theta^*\|_2$, and $\hat{\theta}(\lambda)$ is the solution to (1.2) and $\beta = 0.85$.

Throughout the simulations, we choose $\sigma = 0.5$. The plotted results are averaged over 50 realizations of the noise and, in the case of a random graph, over realizations of said random graph.

[1]To generate instances of the random regular graph, we employed the code from [Pun10], which implements the pairing algorithm by Bollobás, [Bol80].



(a) Oracle choice of λ

(b) Fixed choice of $\lambda = \sigma\rho\sqrt{\log(n)}$

Figure 1: MSE for the Island model with $k = l = 3$ for different choices of the graph and different choices of the regularization parameter λ. The dotted line the best fit of form $C\log(n)/n$ to the complete graph case.

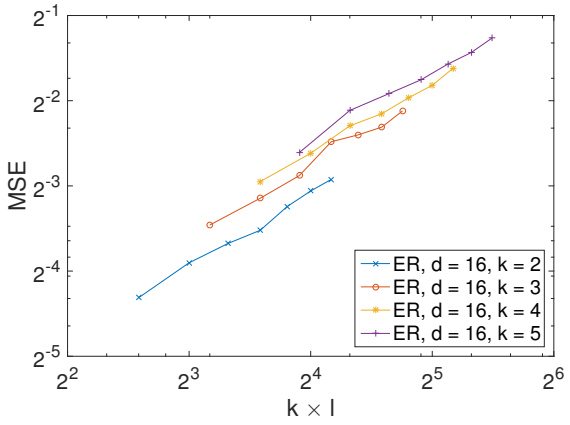


Figure 2: MSE for the Island model for different choices of $k \cdot l$ ($n = 100, \lambda = \sigma\rho\sqrt{\log(n)}$).

In Figure 1, we consider the Island model with $k = 3$ islands, each of size $l = 3$. We plot (on a log-log scale) the mean squared error of the TV denoiser as a function of n for both the oracle choice and the theoretical choice of λ for different graph models: the complete graph, the Erdős-Rényi random graphs with expected degree d for $d = 2, 12, 16$ and the random 12-regular graph. The dotted line indicates the best fit of the form $C\log(n)/n$ that our theoretical analysis predicts in the complete graph case. In all cases, we can see that the mean squared error essentially scales as $C(\log n)/n$ as predicted by our theory. Moreover, all graphs show similar performance, though the sparse ones lead to better computational performance.

The purpose of Figure 2 is to illustrate that the scaling kl/n for the model with islands obtained in subsection 4.3 is indeed the correct one. In this set of simulations we use the Erdős-Rényi graph with expected degree $d = 16$ and plot the mean squared error for different values of the pair (k, l). Specifically, we choose $(k, l) \in [2 : 5] \times [3 : 9]$ and indeed observe a linear dependence on the product kl.

APPENDIX B: PROOFS

B.1 Proof of the main theorem: a sharp oracle inequality for TV denoising

In this subsection, we prove Theorem 2 that we recall for convenience

THEOREM (Sharp oracle inequality for TV denoising). *Fix $\delta \in (0, 1)$, $T \subset [m]$ and let D being the incidence matrix of a connected graph G. Define the regularization parameter*

$$\lambda := \frac{1}{n}\sigma\rho\sqrt{2\log\big(\frac{em}{\delta}\big)},$$

With this choice of λ, the TV denoiser $\hat{\theta}$ defined in (1.2) *satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \inf_{\bar{\theta}\in\mathbb{R}^n}\left\{\frac{1}{n}\|\bar{\theta} - \theta^*\|^2 + 4\lambda\|(D\bar{\theta})_{T^c}\|_1\right\} + \frac{8\sigma^2}{n}\left(\frac{|T|\rho^2}{\kappa_T^2}\log\big(\frac{em}{\delta}\big) + \log\big(\frac{e}{\delta}\big)\right).$$

on the estimation error with probability at least $1 - 2\delta$.

Our proof is based on the sharp oracle inequality for the Lasso in [Gir14, Theorem 4.1, Corollary 4.3] and slightly stronger statements that appear in [DHL14, Theorems 3 and 4].

We start by considering the first order optimality conditions of the convex problem (1.2). By the chain rule for the subdifferential, [Roc70, Theorem 23.9], the subdifferential of the ℓ_1 term is

$$\partial\|D\theta\|_1 = D^\top \operatorname{sign}(D\theta),$$

where

$$\operatorname{sign}(x)_i = \begin{cases} 1 & \text{if } x_i > 0, \\ [-1, 1] & \text{if } x_i = 0, \\ -1 & \text{if } x_i < 0\,. \end{cases}$$

Therefore, for any $\bar{\theta} \in \mathbb{R}^n$, $z \in \operatorname{sign}(D\hat{\theta})$ we get

$$\frac{1}{n}\bar{\theta}^\top(y - \hat{\theta}) = \lambda\bar{\theta}^\top D^\top z = \lambda(D\bar{\theta})^\top z.$$

It yields

$$\frac{1}{n}\hat{\theta}^\top(y-\hat{\theta}) = \lambda\|D\hat{\theta}\|_1 \quad \text{and} \quad \frac{1}{n}\bar{\theta}^\top(y-\hat{\theta}) \le \lambda\|D\bar{\theta}\|_1,$$

In turn, subtracting the above two, we get

$$\frac{1}{n}(\bar{\theta}-\hat{\theta})^\top(\theta^*-\hat{\theta}) \le \frac{1}{n}\varepsilon^\top(\hat{\theta}-\bar{\theta}) + \lambda\|D\bar{\theta}\|_1 - \lambda\|D\hat{\theta}\|_1.$$

Next, using polarization, we can rewrite the above display as

$$\frac{1}{n}(\|\bar{\theta}-\hat{\theta}\|^2 + \|\theta^*-\hat{\theta}\|^2) \le \frac{1}{n}\|\bar{\theta}-\theta^*\|^2 + \frac{2}{n}\varepsilon^\top(\hat{\theta}-\bar{\theta}) + 2\lambda\|D\bar{\theta}\|_1 - 2\lambda\|D\hat{\theta}\|_1. \tag{B.9}$$

We first control the error term $\varepsilon^\top(\hat{\theta}-\bar{\theta})$ as follows. Let Π denote the projection matrix onto $\ker(D)$ and remember that $D^\dagger D = (I-\Pi)$, the projection on $\ker(D)^\perp$. Since $\ker(D) = \ker(D^\top D)$, the kernel of the graph Laplacian, and G is connected, we have $\ker(D) = \operatorname{span}(\mathbf{1}_n)$ [Chu97]; in particular, $\dim\ker(D) = 1$. It yields

$$\begin{aligned}
\varepsilon^\top(\hat{\theta}-\bar{\theta}) &= (\Pi\varepsilon)^\top(\hat{\theta}-\bar{\theta}) + ((I-\Pi)\varepsilon)^\top(\hat{\theta}-\bar{\theta}) \\
&= (\Pi\varepsilon)^\top(\hat{\theta}-\bar{\theta}) + \varepsilon^\top D^\dagger D(\hat{\theta}-\bar{\theta}) \\
&= (\Pi\varepsilon)^\top(\hat{\theta}-\bar{\theta}) + ((D^\dagger)^\top\varepsilon)^\top D(\hat{\theta}-\bar{\theta}) \\
&\le \|\Pi\varepsilon\|\|\widehat{\theta}-\bar{\theta}\| + \|(D^\dagger)^\top\varepsilon\|_\infty\|D(\hat{\theta}-\bar{\theta})\|_1\,,
\end{aligned} \tag{B.10}$$

where in (B.10), we use Hölder's inequality.

To bound the right-hand side in (B.10), we first use the maximal inequality for Gaussian random variables [BLM13, Corollary 2.6]: It yields the following two inequalities hold simultaneously on an even of probability $1-\delta$

$$\|(D^\dagger)^\top\varepsilon\|_\infty \le \sigma\rho\sqrt{2\log(em/\delta)} = \lambda n\,, \qquad \|\Pi\varepsilon\|_2 \le 2\sigma\sqrt{2\log(e/\delta)}\,.$$

Next, note that by the triangle inequality we have

$$\|D(\hat{\theta}-\bar{\theta})\|_1 + \|D\bar{\theta}\|_1 - \|D\hat{\theta}\|_1 \le 2\|(D(\hat{\theta}-\bar{\theta}))_T\|_1 + 2\|(D\bar{\theta})_{T^c}\|_1. \tag{B.11}$$

Moreover, $\|D(\hat{\theta}-\bar{\theta})_T\|_1 \le \kappa_T^{-1}\sqrt{|T|}\|\hat{\theta}-\bar{\theta}\|$. Together with (B.9)–(B.11), it yields

$$\frac{1}{n}(\|\bar{\theta}-\hat{\theta}\|^2 + \|\theta^*-\hat{\theta}\|^2) \le \frac{1}{n}\|\bar{\theta}-\theta^*\|^2 + 4\lambda\|(D\bar{\theta})_{T^c}\|_1 + \frac{4}{n}\|\hat{\theta}-\bar{\theta}\|\left(\sigma\sqrt{2\log\left(e/\delta\right)} + n\frac{\lambda}{\kappa_T}\sqrt{|T|}\right)$$

To conclude the proof, we apply Young's inequality to produce $\frac{1}{n}\|\hat{\theta}-\bar{\theta}\|^2$ which cancels out.

B.2 Control of the inverse scaling factor for the 2D grid

In this subsection, we prove Proposition 4 that we recall here for convenience.

PROPOSITION. *The incidence matrix D_2 of the 2D grid on n vertices has inverse scaling factor $\rho \lesssim \sqrt{\log n}$.*

PROOF. Note first that $S = D_2^\dagger = (D_2^\top D_2)^\dagger D_2^\top$. Moreover, the matrix $D_2^\top D_2$ can be expressed in terms of $D_1^\top D_1$ as

$$D_2^\top D_2 = \begin{bmatrix} D_1^\top \otimes I & I \otimes D_1^\top \end{bmatrix} \begin{bmatrix} D_1 \otimes I \\ I \otimes D_1 \end{bmatrix} = D_1^\top D_1 \otimes I + I \otimes D_1^\top D_1 .$$

It follows from [Str07, Chapter 1.5] that the unnormalized Laplacian $D_1^\top D_1$ of the path graph admits the following spectral decomposition

$$D_1^\top D_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & -1 & 2 & -1 \\ 0 & & \dots & 0 & -1 & 1 \end{bmatrix} = V_1 \Lambda_1 V_1^\top \tag{B.12}$$

where $\Lambda_1 = \mathrm{diag}(\lambda_0, \dots, \lambda_{N-1})$, with

$$\lambda_k = 2 - 2\cos\frac{k\pi}{N}, \quad k \in [\![0, N[\![,$$

and $V_1 = [v_0, \dots, v_{N-1}]$ is the discrete Fourier transform DCT-2 on $\mathbb{R}^N$ so that each eigenvector $v_k \in \mathbb{R}^N$ has coordinates

$$\begin{aligned} (v_0)_j &= \frac{1}{N}, \quad j \in [\![0, N[\![\\ (v_k)_j &= \sqrt{\frac{2}{N}} \cos\left(\frac{(j+1/2)k\pi}{N}\right), \quad j \in [\![0, N[\![, \, k \in [\![1, N[\![. \end{aligned}$$

Therefore, $D_2^\top D_2 = V_2 \Lambda_2 V_2^\top$, where $\Lambda_2 = \Lambda_1 \otimes I + I \otimes \Lambda_1$ and $V_2 = V_1 \otimes V_1$.

As a result, S has $2N(N-1)$ columns and can be written as

$$S = D_2^\dagger = V_2 \Lambda_2^\dagger V_2^\top \begin{bmatrix} D_1^\top \otimes I & I \otimes D_1^\top \end{bmatrix} = [(s^{(1)}_{i,j})_{i\in[N-1], j\in[N]}, (s^{(2)}_{i,j})_{i\in[N], j\in[N-1]}]$$

Write $D_1^\top = [d_1, \dots, d_{N-1}]$ and note that the columns of S have norm given for $\diamond \in \{1, 2\}$ by

$$\begin{aligned} \|s^{(\diamond)}_{i,j}\|_2^2 &= \sum_{\substack{k,l=0 \\ (k,l)\neq(0,0)}}^{N-1} \frac{1}{(\lambda_k + \lambda_l)^2} \langle v_k \otimes v_l, d_i \otimes e_j \rangle^2 \\ &= \sum_{\substack{k,l=0 \\ (k,l)\neq(0,0)}}^{N-1} \frac{1}{(4 - 2\cos\frac{k\pi}{N} - 2\cos\frac{l\pi}{N})^2} \langle v_k, d_i \rangle^2 \langle v_l, e_j \rangle^2 , \end{aligned}$$

where $e_0, \dots, e_{N-1}$ are the vectors of the canonical basis of $\mathbb{R}^N$. Next, note that

$$\langle v_l, d_i \rangle^2 = \frac{2}{N} \left(\cos\frac{l\pi(i+3/2)}{N} - \cos\frac{l\pi(i+1/2)}{N} \right)^2 \le \frac{2l^2\pi^2}{N^3} .$$

because $x \mapsto \cos x$ is 1-Lipschitz. Moreover, we have immediately that $\langle v_k, e_j \rangle^2 \le 2/N$.

It remains to bound the sum. To that end, observe that $2-2\cos x \geq x^2/2$ for any $x \in [0,1/2]$ and $2-2\cos x \geq 0.1$, for $x \in [1/2,\pi]$. Hence, we can split the sum into to parts to get

$$\begin{aligned}
\left\|s_{i,j}^{(\diamond)}\right\|_2^2 &\leq \frac{4\pi^2}{N^4} \sum_{\substack{k,l=0\\(k,l)\neq(0,0)}}^{N-1} \frac{l^2}{(4-2\cos\frac{k\pi}{N}-2\cos\frac{l\pi}{N})^2} \\
&\leq \frac{4\pi^2}{N^4} \sum_{\substack{k,l=0\\(k,l)\neq(0,0)}}^{N-1} \frac{l^2}{(4-2\cos\frac{k\pi}{N}-2\cos\frac{l\pi}{N})^2} \big[\mathbb{1}(\frac{2\pi}{N}(k\vee l)\leq 1)+\mathbb{1}(\frac{2\pi}{N}(k\vee l)>1)\big] \\
&\leq 16 \sum_{\substack{k,l=0\\(k,l)\neq(0,0)}}^{N-1} \frac{l^2}{(k^2+l^2)^2} + \frac{400\pi^2}{N^3}\sum_{k=0}^{N-1} k^2 \lesssim \sum_{k,l=1}^{N-1} \frac{l^2}{(k^2+l^2)^2} + 1\,.
\end{aligned}$$

Using a comparison between series and integral, noting that $x \to x^2/(k^2+x^2)^2$ is increasing on $[0,k^2]$ and decreasing on $[k^2,\infty)$, it is immediate that

$$\sum_{k,l=1}^{N-1} \frac{l^2}{(k^2+l^2)^2} \leq \sum_{k=1}^{N-1} \frac{1}{k}\int_0^\infty \frac{x^2}{(1+x^2)^2}\,\mathrm{d}x + \sum_{k=1}^{N-1}\frac{1}{4k^2} \lesssim \sum_{k=1}^{N}\frac{1}{k}+1 \lesssim \log N.$$

To conclude the proof, observe that $n=N^2$. □

B.3 Control of the inverse scaling factor for high-dimensional grids

In this subsection, we prove Proposition 6 that we recall here for convenience.

PROPOSITION. *For the incidence matrix of the regular grid on N^d nodes in d dimensions, $\rho \leq C(d)$, for some $C(d)>0$.*

PROOF. Similarly to the proof of Proposition 4, the eigendecomposition of $(D_d^\top D_d)^\dagger$ has the form $\Lambda_d = \Lambda_1\otimes I\otimes\cdots\otimes I+\cdots+I\otimes I\otimes\cdots\otimes\Lambda_1$, $V_d = V_1^{\otimes d}$. Keeping the same notation as in the preceding proof,

$$S = D_d^\dagger = [(s_{\boldsymbol{i}}^{(j)})_{i_j\in[N-1],\,i_k\in[N],}\ \text{for } k\neq j,\ j\in[d]],$$

we have

$$\begin{aligned}
\|s_{\boldsymbol{i}}^{(1)}\|_2^2 &= \sum_{\substack{k_l=0,\,\boldsymbol{k}\neq 0\\ l=1,\dots,d}}^{N-1} \left(\sum_{j=1}^d \lambda_{k_j}\right)^{-2} \langle v_{k_1}, d_{i_1}\rangle^2 \prod_{j=2}^d \langle v_{k_j}, e_{i_j}\rangle^2 \\
&= \sum_{\substack{k_l=0,\,\boldsymbol{k}\neq 0\\ l=1,\dots,d}}^{N-1} \left(\sum_{j=1}^d \left(2-2\cos\frac{k_j\pi}{N}\right)\right)^{-2} \langle v_{k_1}, d_{i_1}\rangle^2 \prod_{j=2}^d \langle v_{k_j}, e_{i_j}\rangle^2
\end{aligned}$$

and by symmetry, this case is enough to deduce the claim for an arbitrary $s_{\boldsymbol{i},j}$, $j\in[d]$. Observing again that

$$\|v_{k_j}\|_\infty \leq \sqrt{2/N}, \quad \langle v_{k_j}, e_{i_j}\rangle^2 \leq 2/N,$$

and

$$\langle v_{k_1}, d_{i_1}\rangle^2 \le \frac{2k_1^2}{N^3},$$

it remains to bound the sum above.

For this, use the same bounds on the cosine function to split it up into a part bounded by a constant and one that behaves like a square:

$$\begin{aligned}
\|s_{\boldsymbol{i}}^{(1)}\|^2 &\le \frac{2^d}{N^{d+2}} \sum_{\substack{k_l=0,\,\boldsymbol{k}\neq 0\\ l=1,\dots,d}}^{N-1} k_1^2 \left(2d - 2\sum_{j=1}^d \cos\frac{k_j\pi}{N}\right)^{-2} \\
&\le \frac{2^d}{N^{d+2}} \sum_{\substack{k_l=0\\ \boldsymbol{k}\neq 0}}^{N-1} k_1^2 \left(2d - 2\sum_{j=1}^d \cos\frac{k_j\pi}{N}\right)^{-2} (\mathbb{1}(\forall j:\ k_j\pi/N \le 1/2) + \mathbb{1}(\exists j:\ k_j\pi/N > 1/2)) \\
&\lesssim \frac{2^d}{N^{d-2}} \sum_{\substack{k_l=0\\ \boldsymbol{k}\neq 0}}^{N-1} k_1^2 \left(\sum_{j=1}^d k_j^2\right)^{-2} + 1
\end{aligned}$$

We again want to exclude all indices having a zero element. This amounts to finding a bound of the order $o(N^{d+2})$ for the same sum in one dimension less than we are considering here, times d for each coordinate that can be zero. In order to achieve this, we argue by induction: in $d = 3$ dimensions, the corresponding summation runs over two indices and has been shown to be of order $O(\log n) = o(N)$ in the proof of Proposition 4, so the base case is valid. The following analysis will show that the whole sum is $O(N^{d+2})$ for $d \ge 3$, which is the induction step. This means we can assume

$$\|s_{\boldsymbol{i}}^{(1)}\|^2 \le \frac{2^d}{N^{d-2}} \sum_{\substack{k_l=1\\ l=1,\dots,d}}^{N-1} k_1^2\Big(\sum_{j=1}^d k_j^2\Big)^{-2} + o(d) \lesssim \frac{2^d}{N^{d-2}} \sum_{\substack{k_l=1\\ l=1,\dots,d}}^{N-1} k_1^2\Big(\sum_{j=1}^d k_j^2\Big)^{-2} + 1$$

Next, observe that $\int_0^\infty x^2(1+x^2)^{-2}\,\mathrm{d}x \lesssim 1$. It yields

$$\begin{aligned}
\sum_{\substack{k_l=1\\ l=1,\dots,d}}^{N-1} k_1^2\Big(\sum_{j=1}^d k_j^2\Big)^{-2} &\le \frac{2^d}{N^{d-2}} \sum_{\substack{k_l=1\\ l=2,\dots,d}}^{N-1} \int_0^\infty x^2\Big(x^2 + \sum_{j=2}^d k_j^2\Big)^{-2}\,\mathrm{d}x + \frac{2^d}{N^{d-2}} \sum_{\substack{k_j=1\\ j=2,\dots,d}}^{N-1} \Big(\sum_{j=2}^d k_j^2\Big)^{-1} \\
&= \frac{2^d}{N^{d-2}} \sum_{\substack{k_j=1\\ j=2,\dots,d}}^{N-1} \Big(\sum_{j=2}^d k_j^2\Big)^{-1/2} \int_0^\infty \frac{y^2}{(y^2+1)^2}\,\mathrm{d}x + \frac{2^d}{N^{d-2}} \sum_{\substack{k_j=1\\ j=2,\dots,d}}^{N-1} \Big(\sum_{j=2}^d k_j^2\Big)^{-1} \\
&\lesssim \frac{2^d}{N^{d-2}} \sum_{\substack{k_l=1\\ l=2,\dots,d}}^{N-1} \Big(\sum_{j=2}^d k_j^2\Big)^{-1/2}
\end{aligned}$$

Next, bounded the series by an integral together with a change to polar coordinates, we get

$$\begin{aligned}\frac{2^d}{N^{d-2}} \sum_{\substack{k_l=1 \\ l=2,\ldots,d}}^{N-1} \Big(\sum_{j=2}^{d} k_j^2\Big)^{-1/2} &\leq \frac{2^d}{N^{d-2}} \int_{\{0\leq x_j\leq N,\, j=1,\ldots,d-1\}} \frac{1}{\|x\|_2}\,\mathrm{d}x \\ &\leq \frac{2^d}{N}\int_0^N \int_0^N \frac{1}{\sqrt{x^2+y^2}}\,\mathrm{d}x\,\mathrm{d}y \\ &= 2^d \log(3+2\sqrt{2}) \leq 2^d\end{aligned}$$

□

REFERENCES

[ACSW12] Ery Arias-Castro, Joseph Salmon, and Rebecca Willett, *Oracle inequalities and minimax rates for nonlocal means and related adaptive kernel-based methods*, SIAM Journal on Imaging Sciences **5** (2012), no. 3, 944–992.

[AT16] Taylor B. Arnold and Ryan J. Tibshirani, *Efficient implementations of the generalized lasso dual path algorithm*, Journal of Computational and Graphical Statistics **25** (2016), no. 1, 1–27.

[Bel15] Pierre C. Bellec, *Sharp oracle inequalities for Least Squares estimators in shape restricted regression*, arXiv preprint arXiv:1510.08029 (2015).

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, OUP Oxford, February 2013.

[Bol80] Béla Bollobás, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, European Journal of Combinatorics **1** (1980), no. 4, 311–316.

[CGS15] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen, *On matrix estimation under monotonicity constraints*, arXiv preprint arXiv:1506.03430 (2015).

[Chu97] Fan RK Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.

[DHL14] Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer, *On the prediction performance of the lasso*, to appear in Bernoulli, arXiv 1402.1700, February 2014.

[DJ95] David L. Donoho and Iain M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc. **90** (1995), no. 432, 1200–1224. MR1379464 (96k:62093)

[Fri04] J Friedman, *A proof of Alon's second eigenvalue conjecture and related problems*, Mem. Amer. Math. Soc **195** (2004), no. 910.

[Gir14] Christophe Giraud, *Introduction to high-dimensional statistics*, CRC Press, 2014.

[KOV14] Theodore Kolokolnikov, Braxton Osting, and James Von Brecht, *Algebraic connectivity of Erdös-Rényi graphs near the connectivity threshold*, Manuscript in preparation (2014).

[KT93] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*, Lecture Notes in Statistics, vol. 82, Springer New York, New York, NY, 1993.

[Mv97] Enno Mammen and Sara van de Geer, *Locally adaptive regression splines*, The Annals of Statistics **25** (1997), no. 1, 387–413.

[NW13] Deanna Needell and Rachel Ward, *Near-optimal compressed sensing guarantees for total variation minimization*, IEEE Transactions on Image Processing **22** (2013), no. 10, 3941–3949.

[OV15] Edouard Ollier and Vivian Viallon, *Regression modeling on stratified data: automatic and covariate-specific selection of the reference stratum with simple L_1-norm penalties*, arXiv:1508.05476 [math, stat] (2015).

[Pun10] Golan Pundak, *Random Regular Generator*, MATLAB Central File Exchange (2010).

[QJ12] Junyang Qian and Jinzhu Jia, *On pattern recovery of the fused Lasso*, arXiv:1211.5194 (2012).

[Rin09] A. Rinaldo, *Properties and refinements of the fused lasso*, The Annals of Statistics **37** (2009), no. 5B, 2922–2952.

[Roc70] Ralph Tyrell Rockafellar, *Convex analysis*, Princeton university press, 1970.

[ROF92] Leonid I. Rudin, Stanley Osher, and Emad Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1, 259–268.

[She10] Yiyuan She, *Sparse regression with exact clustering*, Electronic Journal of Statistics **4** (2010), 1055–1096.

[SSR12] James Sharpnack, Aarti Singh, and Alessandro Rinaldo, *Sparsistency of the edge lasso over graphs*, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12) (Neil D. Lawrence and Mark A. Girolami, eds.), vol. 22, 2012, pp. 1028–1036.

[Str07] Gilbert Strang, *Computational science and engineering*, vol. 1, Wellesley-Cambridge Press Wellesley, 2007.

[TSR+05] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 1, 91–108.

[VLLHP16] Vivian Viallon, Sophie Lambert-Lacroix, Hölger Hoefling, and Franck Picard, *On the robustness of the generalized fused lasso to prior specifications*, Statistics and Computing **26** (2016), no. 1-2, 285–301.

[WNC05] Rebecca Willett, Robert Nowak, and Rui M. Castro, *Faster rates in regression via active learning*, Advances in Neural Information Processing Systems, 2005, pp. 179–186.

[WSST15] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan Tibshirani, *Trend Filtering on Graphs*, Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, 2015, pp. 1042–1050.

[XKWG14] Bo Xin, Yoshinobu Kawahara, Yizhou Wang, and Wen Gao, *Efficient Generalized Fused Lasso and Its Application to the Diagnosis of Alzheimer's Disease*, Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 2163–2169.

Jan-Christian Hütter
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(huetter@math.mit.edu)

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(rigollet@math.mit.edu)