Journal of Machine Learning Research 21 (2020) 1-38 Submitted 3/20; Revised 10/20; Published 12/20

Adaptive Rates for Total Variation Image Denoising

Francesco Ortelli FORTELLI@ETHZ.CH
Seminar für Statistik, ETH Zürich
Rämistrasse 101
8092 Zürich, Schweiz

Sara van de Geer GEER@ETHZ.CH
Seminar für Statistik, ETH Zürich
Rämistrasse 101
8092 Zürich, Schweiz

Editor: Arnak Dalalyan

Abstract

We study the theoretical properties of image denoising via total variation penalized least-squares. We define the total vatiation in terms of the two-dimensional total discrete derivative of the image and show that it gives rise to denoised images that are piecewise constant on rectangular sets.

We prove that, if the true image is piecewise constant on just a few rectangular sets, the denoised image converges to the true image at a parametric rate, up to a log factor. More generally, we show that the denoised image enjoys oracle properties, that is, it is almost as good as if some aspects of the true image were known.

In other words, image denoising with total variation regularization leads to an adaptive reconstruction of the true image.

Keywords: total variation, image denoising, fused Lasso, oracle inequalities

1. Introduction

Image denoising is a broad and active field of research, where the aim is to reconstruct an image corrupted with noise (Dabov et al., 2007; Elad, 2010; Arias-Castro et al., 2012; Zhang et al., 2018; Goyal et al., 2020). Generally, some assumptions on the structure of the underlying image have to be made to favor denoised images showing such structure (Mammen and Tsybakov, 1995; Polzehl and Spokoiny, 2003). One of these assumptions is that the image to reconstruct is constant on few sets belonging to some specific class, as for instance the class of connected sets or the class of rectangular connected sets. Image denoising with total variation regularization is known to promote such piecewise-constant denoised images (Bach, 2011).

The use of total variation penalties for image denoising dates back to Rudin et al. (1992) and has been the subject of various studies (Mammen and van de Geer, 1997; Chambolle and Lions, 1997; Caselles et al., 2015; Chambolle et al., 2017). For an overview over some theoretical and practical aspects, see Chambolle et al. (2010). The theoretical study of total variation for image denoising has recently experienced a surge of interest (Sadhanala

et al., 2016; Wang et al., 2016; Hütter and Rigollet, 2016; Padilla et al., 2018; Chatterjee and Goswami, 2019; Fang et al., 2019).

1.1 Review of the Literature

Consider a continuous image $\phi(x,y), (x,y) \in [0,1]^2$ and a discrete or discretized image $f(j,k), (j,k) \in \{1,\dots,n_1\} \times \{1,\dots,n_2\}$. In the literature we encounter different definitions of (two-dimensional) total variation. Three of them are listed in what follows.

- In the seminal work by Rudin et al. (1992), total variation is defined in terms of partial derivatives as
$$\int_0^1 \int_0^1 \sqrt{\left(\frac{\partial}{\partial x}\phi(x,y)\right)^2 + \left(\frac{\partial}{\partial y}\phi(x,y)\right)^2}\, dx\ dy.$$
Different discretization procedures have been proposed for the total variation by Rudin et al. (1992): isotropic, anisotropic, upwind (Chambolle et al., 2011) and Shannon (Abergel and Moisan, 2017) total vatiation. For more details and a recently proposed discretization we refer to Condat (2017).

- Total variation in terms of partial derivatives can also be defined as
$$\int_0^1 \int_0^1 \left|\frac{\partial}{\partial x}\phi(x,y)\right| + \left|\frac{\partial}{\partial y}\phi(x,y)\right| dx\ dy.$$
Its discrete version
$$\sum_{j=2}^{n_1}\sum_{k=1}^{n_2}|f(j,k) - f(j-1,k)| + \sum_{j=1}^{n_1}\sum_{k=2}^{n_2}|f(j,k) - f(j,k-1)|$$
is considered in Sadhanala et al. (2016); Wang et al. (2016); Hütter and Rigollet (2016); Chatterjee and Goswami (2019) and corresponds, up to normalization, to summing up the edge differences of the discrete image across a two-dimensional grid graph. Used as a penalty for least squares, this definition results in denoised images which are piecewise constant on connected sets of any shape (Bach, 2011).

- Alternatively, total variation can be defined in terms of the total derivative as
$$\int_0^1 \int_0^1 \left|\frac{\partial}{\partial x}\frac{\partial}{\partial y}\phi(x,y)\right| dx\ dy$$
and in discretized form as
$$\sum_{j=2}^{n_1}\sum_{k=2}^{n_2}|f(j,k) - f(j-1,k) - f(j,k-1) + f(j-1,k-1)|.$$
This approach is adopted by Mammen and van de Geer (1997) and Fang et al. (2019) and will also be adopted in this paper. As a penalty for least squares, this definition will be shown to render denoised images which are piecewise constant on rectangular sets.

In the literature, the second definition of total variation in terms of partial derivatives is more popular than the one in terms of total derivatives. Least-squares estimators with a penalty on the discrete partial derivatives of the image are the subject of a vast statistical literature. Let n denote the number of pixels of the image. Sadhanala et al. (2016) derive minimax rates, which, for large n and under the canonical scaling, are of order $\sqrt{\log(n)/n}$. Later, Sadhanala et al. (2017) extend the minimax results to higher order differences and to higher dimensions. Hütter and Rigollet (2016) prove sharp oracle inequalities with the rate $\log n/\sqrt{n}$. Lastly, the very recent work by Chatterjee and Goswami (2019) focuses on the constrained optimization problem, solvable for example by Fadili and Peyre (2011), and a tuning-free version thereof. For a certain underlying image, a rate faster than the minimax rate is obtained. The approach by Chatterjee and Goswami (2019), as the one used for higher-order total variation regularization (Guntuboyina et al., 2020), is based on bounding Gaussian widths of tangent cones.

On the other side, Mammen and van de Geer (1997) define total variation in terms of total derivatives, as in this paper, and obtain the rate $n^{-3/5}$ for the estimation of the "interaction terms". The same definition of total variation is used by Fang et al. (2019), who study a constrained version of the estimator.

1.2 Contributions

We prove upper bounds on the mean squared error for image denoising with a total variation penalty promoting piecewise constant estimates on rectangular sets. These upper bounds are presented in the form of oracle inequalities, cf. Koltchinskii (2006); Lounici et al. (2011); Dalalyan and Salmon (2012); Stucky and van de Geer (2017); Bellec et al. (2017, 2018); Bellec (2018); Elsener and van de Geer (2019). Oracle inequalities are finite-sample theoretical guarantees on the performance of an estimator treating in a unified way both the cases of well-specified and misspecified models. In particular, we show that the mean squared error of the denoised image is upper bounded by the optimal tradeoff between "approximation error" and "estimation error". This optimal tradeoff depends on the true underlying image, which is unknown. Hence the term "oracle": the estimator is shown to perform as well as if it knew the aspects of the true image necessary to reach this optimal tradeoff.

We derive oracle inequalities with both fast and slow rates.

- In the case of fast rates, the estimation error is shown to be of order $(s^*)^{3/2}\log^2(n)/n$ for oracle images being constant on s^* rectangular sets of roughly the same size. The parametric rate is reached up to the log term and a factor $(s^*)^{1/2}$ due to the two-dimensionality of the problem. The general result with fast rates is Theorem 19, while a special case is exposed in Theorem 6. Theorem 19 is an adaptive result: the bound on the mean squared error of the denoised image depends on the structure in the underlying image. This dependence is mediated by a so-called oracle, which trades off the fidelity to the underlying image and the number of constant rectangular regions s^* to estimate.

- In the case of slow rates, the estimation error is shown to be of order $n^{-5/8}\log^{3/8} n$ under the assumption that the total variation of the image is bounded, cf. Theorem

32. This rate outperforms the rate $n^{-3/5}$ obtained by Mammen and van de Geer (1997).

These contributions build on previous research in the one-dimensional setting, where the classical example is the fused Lasso (Tibshirani et al., 2005; Dalalyan et al., 2017; Lin et al., 2017; Ortelli and van de Geer, 2018). The term "fused Lasso" often refers to the penalty on the total variation of the coefficients in a linear model. Generalizations of the fused Lasso to other graph structures than the chain graph and to penalties on higher-order discrete derivatives are known under the name of edge Lasso (Sharpnack et al., 2012; Hütter and Rigollet, 2016) and trend filtering (Tibshirani, 2014; Wang et al., 2016; Ortelli and van de Geer, 2019b; Guntuboyina et al., 2020), respectively.

1.3 Technical Tools

The skeleton of our proofs in Section 6 closely follows the proofs of oracle inequalities for similar estimation problems, cf. the proofs in Hütter and Rigollet (2016); Dalalyan et al. (2017); Ortelli and van de Geer (2019b, 2020). The more involved part is adapting to two dimensions the techniques previously applied in one dimension, in particular: the derivation of the synthesis form of the estimator, the bound used to control the noise also known as bound on the increments of the empirical process and the bound on the "effective sparsity".

- We define image denoising with total variation as an analysis estimation problem: the observations are approximated by a candidate estimator, some aspects of which are penalized. The penalized aspects are computed via a linear operator, the so-called analysis operator, which in our case corresponds to the two-dimensional discrete derivative operator. Elad et al. (2007) explain how to obtain a synthesis formulation of analysis estimators. In the synthesis formulation, the candidate estimator is synthesized by a linear combination of atoms. The atoms constitute the moral equivalent of basis vectors (or basis matrices in the case of image denoising). The collection of atoms is called dictionary. The penalty is then enforced on the convex relaxation of the number of atoms used to synthesize the estimator.

  As in our previous work in one dimension (Ortelli and van de Geer, 2018, 2019b), the first step is to reformulate total variation image denoising in synthesis form and show that the dictionary consists of a collection of indicator functions of half-intervals, see Section 5 and in particular Lemma 7 and 8. As a consequence, the estimator will be piecewise constant on rectangular regions. Moreover the insights from the synthesis formulation of the estimator will help us in the further analysis of its behavior.

- A central step in the derivation of oracle inequalities is to control the random part of the estimation problem consisting of the increments of an empirical process (van de Geer, 2009), whose increments need to be bounded. We apply to the case of image estimation a technique developed by Dalalyan et al. (2017). This technique involves the decomposition of the increments of the empirical process into two parts: a part projected onto a suitable linear space and a remainder, see Lemma 13 in Subsection 6.1. The projected part will usually be of low rank, while the remainder will contribute to the "effective sparsity".

The dictionary atoms of the synthesis formulation are strongly correlated. Thus, even when choosing a low-rank linear subspace spanned by only few dictionary atoms, the remainder will be small. As a crucial consequence, also the contribution to the effective sparsity will be small.

- The effective sparsity, see Definition 15 in Subsection 6.2 in vector form or Definition 20 in Subsection 7.2 in matrix form, gives an indication for the effective number of parameters we have in the model. Indeed, oracle inequalities with fast rates usually show an estimation error of the order "effective sparsity $\times$ $\log n/n$", and thus the effective sparsity can be interpreted as the effective degrees of freedom that are spent to estimate the model parameters. In this paper, because of the projection arguments used to bound the increments of the empirical process, the effective sparsity will be multiplied by a factor smaller than $\log n/n$ to obtain the fast rate. In the literature, the reciprocal of a stronger version of the effective sparsity is also known under the name of "compatibility constant", which is related to the restricted eigenvalue (van de Geer and Bühlmann, 2009; van de Geer, 2016, 2018). Also in this case, we extend a previously known one-dimensional bound (Ortelli and van de Geer, 2019b) based on interpolating polynomials to the two-dimensional case. The new bound is based on an interpolating matrix, which interpolates the active parameters and can be found in Lemma 22 in Subsection 7.2.

1.4 Organization of the Paper

In Section 2 we introduce the required notation. In Section 3 we define the model and the estimator and we show how an image can be decomposed into global mean, (centered) row and column means and interaction terms. This is a so-called ANOVA decomposition of an image. As a preview of the main result, we state in Section 4 the special case for a square image. In Section 5 we formulate the estimator for the interaction terms in synthesis form. In Section 6 we expose the standard techniques used to obtain oracle inequalities with fast and slow rates for general analysis problems. The derivation of bounds on the effective sparsity is given in Section 7, where we also present the details of our main result, which is an oracle inequality with fast rates. In Section 8 we prove the slow rate $n^{-5/8}\log^{3/8} n$. Section 9 concludes the paper.

2. Notation and Definitions

We expose the mathematical notation required and some basic definitions.

2.1 Matrix Notation

We model images as matrices of dimension $n_1 \times n_2$ with entries the real-valued pixel values. Let $n := n_1 n_2$ denote the total number of pixels of an image.

For two integers $i^* \leq i$, we use the notation $[i^* : i] = \{i^*, \dots, i\}$. If $i^* = 1$, we write $[i] := [1 : i]$. For a row index $j \in [n_1]$ and a column index $k \in [n_2]$, we refer to the corresponding entry of the matrix $f \in \mathbb{R}^{n_1 \times n_2}$ in two different ways: either by $f_{j,k}$ using subscripts or by $f(j,k)$ using arguments.

These two equivalent notations will be useful in different situations. For instance, the notation using arguments will come in handy in Section 5, when deriving the synthesis form of the estimator.

By $\|f\|_2$ we denote the Frobenius norm of $f \in \mathbb{R}^{n_1 \times n_2}$, that is

$$\|f\|_2 := \left(\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} f_{j,k}^2 \right)^{1/2}.$$

Moreover we define

$$\|f\|_1 := \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} |f_{j,k}|$$

as the sum of the absolute values of the entries of $f \in \mathbb{R}^{n_1 \times n_2}$.

2.2 Total Variation

Let $f \in \mathbb{R}^{n_1 \times n_2}$ be an image. Let $D_1 \in \mathbb{R}^{(n_1-1) \times n_1}$ and $D_2 \in \mathbb{R}^{(n_2-1) \times n_2}$ be discrete difference operators, that is, matrices of the form

$$\begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}.$$

Definition 1 (Two-dimensional discrete derivative operator) *The two-dimensional discrete derivative operator* $\Delta : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^{(n_1-1) \times (n_2-1)}$ *is defined as*

$$\Delta f = D_1 f D_2^T.$$

Note that Δ is a linear operator and

$$(\Delta f)_{j,k} := f_{j,k} - f_{j,k-1} - f_{j-1,k} + f_{j-1,k-1}, \ (j,k) \in [2 : n_1] \times [2 : n_2].$$

Definition 2 (Total variation) *The total variation* $\mathrm{TV}(f)$ *of an image* f *is defined as*

$$\mathrm{TV}(f) := \|\Delta f\|_1 = \sum_{j=2}^{n_1} \sum_{k=2}^{n_2} |f_{j,k} - f_{j,k-1} - f_{j-1,k} + f_{j-1,k-1}|.$$

2.3 Active Set

Fix some set $S \subseteq [3 : n_1 - 1] \times [3 : n_2 - 1]$. We can think of S as the subset of coefficients of the total derivative Δf which are active, that is, nonzero. The cardinality of S is denoted by $s := |S|$. We write $S := \{t_1, \ldots, t_s\}$. We refer to the elements of S as jump locations. The coordinates of a jump location t_m are denoted by $(t_{1,m}, t_{2,m})$, $m = 1, \ldots, s$. Note that we require that $2 < t_{1,m} < n_1$ and $2 < t_{2,m} < n_2$. This assumption ensures that we have no boundary effects when doing partial integration, see Lemma 23.

For two matrices $a = \{a_{j,k}\}_{(j,k) \in [2:n_1] \times [2:n_2]}$ and $b = \{b_{j,k}\}_{(j,k) \in [2:n_1] \times [2:n_2]}$ we use the symbol $\odot$ for entry-wise multiplication: $(a \odot b)_{j,k} := a_{j,k} b_{j,k}, \ (j,k) \in [2 : n_1] \times [2 : n_2]$.

Moreover we define $a_S := \{a_{j,k}, \ (j,k) \in S\}$ and $a_{-S} := \{a_{j,k}, \ (j,k) \notin S\}$. We will use the same notation $a_S \in \mathbb{R}^{(n_1-1) \times (n_2-1)}$ for the matrix which shares its entries with a for $(j,k) \in S$ and has all its other entries equal to zero. Similarly, $a_{-S} \in \mathbb{R}^{(n_1-1) \times (n_2-1)}$ shares its entries with a for $(j,k) \notin S$ and has its other entries equal to zero.

2.4 Linear Projections

For a linear space $\mathcal{W}$, let $\mathrm{P}_{\mathcal{W}}$ denote the projection operator on $\mathcal{W}$ and $\mathrm{A}_{\mathcal{W}} := \mathrm{I} - \mathrm{P}_{\mathcal{W}}$ the corresponding antiprojection operator. The antiprojection operator on $\mathcal{W}$ computes the residuals of the orthogonal projection on $\mathcal{W}$.

3. Preliminaries

We want to estimate the image $f^0 \in \mathbb{R}^{n_1 \times n_2}$ based on its noisy observation $Y = f^0 + \epsilon$, where $\epsilon \in \mathbb{R}^{n_1 \times n_2}$ is a noise matrix with i.i.d. Gaussian entries with known variance σ^2. For the case of unknown variance, Ortelli and van de Geer (2020) show how to simultaneously estimate the signal and the noise variance by extending the idea of the square-root Lasso (Belloni et al., 2011) to total variation penalized least-squares.

3.1 ANOVA Decomposition of an Image

In this subsection we introduce the ANOVA decomposition, which separates an image $f \in \mathbb{R}^{n_1 \times n_2}$ into four mutually orthogonal components: the global mean, the two matrices of main effects and the matrix of interaction terms.

Definition 3 (Global mean) *The global mean $f(\circ, \circ) \in \mathbb{R}$ is defined as*

$$f(\circ, \circ) := \frac{1}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} f(j,k).$$

Definition 4 (Main effects) *The main effects are defined as $f(\cdot, \circ) = \{f(j, \circ)\}_{(j,k) \in [n_1] \times [n_2]}$ and $f(\circ, \cdot) = \{f(\circ, k)\}_{(j,k) \in [n_1] \times [n_2]}$, where*

$$f(j, \circ) := \frac{1}{n_2} \sum_{k=1}^{n_2} f(j,k) - f(\circ, \circ), \ j \in [n_1]$$

and

$$f(\circ, k) := \frac{1}{n_1} \sum_{j=1}^{n_1} f(j,k) - f(\circ, \circ), \ k \in [n_2].$$

Note that $f(\cdot, \circ)$ has identical columns and $f(\circ, \cdot)$ has identical rows. We define the total variation of the main effects as

$$\mathrm{TV}_1(f) := \sum_{j=2}^{n_1} |f(j, \circ) - f(j-1, \circ)|$$

and

$$\mathrm{TV}_2(f) := \sum_{k=2}^{n_2} |f(\circ, k) - f(\circ, k-1)|.$$

Definition 5 (Interaction terms) *The interaction terms are defined as*

$$\tilde{f}(j,k) = f(j,k) - f(\circ, \circ) - f(j, \circ) - f(\circ, k), \ (j,k) \in [n_1] \times [n_2].$$

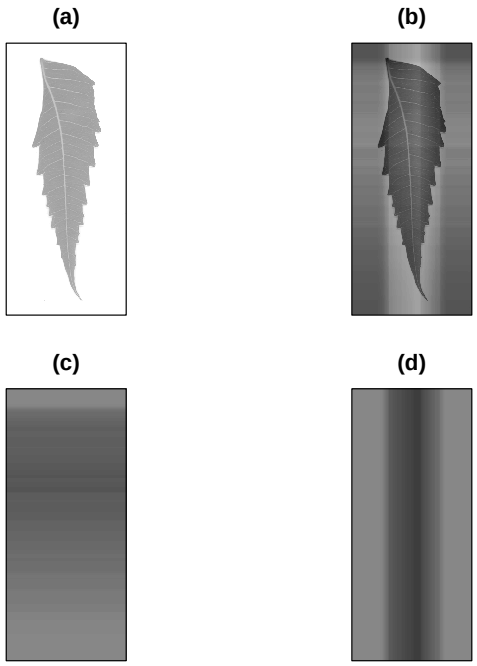

Figure 1: The ANOVA decomposition of the image lg1 from the Leaf Shapes Database by Waghmare. The image (a) is the original image f, (b) represents the interaction terms $\tilde{f}$ and (c) and (d) are the main effects $f(\cdot, \circ)$ and $f(\circ, \cdot)$, respectively.

Let $\psi^{1,1} = \{1\}^{n_1 \times n_2}$. The ANOVA decomposition of an image f is

$$f = f(\circ, \circ)\psi^{1,1} + f(\cdot, \circ) + f(\circ, \cdot) + \tilde{f}$$

and is illustrated in Figure 1 for an image from the Leaf Shape Database. Note that $f(\circ, \circ)\psi^{1,1}$, $f(\cdot, \circ)$, $f(\circ, \cdot)$ and $\tilde{f}$ are mutually orthogonal and thus we have that

$$\|f\|_2^2 = n_1 n_2 f^2(\circ, \circ) + \|f(\cdot, \circ)\|_2^2 + \|f(\circ, \cdot)\|_2^2 + \|\tilde{f}\|_2^2.$$

We now use the ANOVA decomposition to define the estimator for the interaction terms, which is the main object studied in this paper.

3.2 The Estimator

We consider the estimator

$$\hat{f} := \underset{\mathrm{f} \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left\{ \|Y - \mathrm{f}\|_2^2 / n + 2\lambda \mathrm{TV}(\mathrm{f}) + 2\lambda_1 \mathrm{TV}_1(\mathrm{f}) + 2\lambda_2 \mathrm{TV}_2(\mathrm{f}) \right\},$$

where $\lambda, \lambda_1, \lambda_2 > 0$ are positive tuning parameters. We call $\hat{f}$ the two-dimensional total variation regularized least squares estimator. This estimator has the form of an analysis estimator (Elad et al., 2007): it approximates the observations under a regularization penalty on the ℓ_1-norm of a linear operator of the signal f.

Since $Y(\circ,\circ)\psi^{1,1}$, $Y(\cdot,\circ)$, $Y(\circ,\cdot)$ and $\tilde{Y}$ are mutually orthogonal, we may decompose the estimator as

$$\hat{f} = \hat{f}(\circ,\circ)\psi^{1,1} + \hat{f}(\cdot,\circ) + \hat{f}(\circ,\cdot) + \hat{\tilde{f}},$$

where

$$\begin{aligned}
\hat{f}(\circ,\circ) &:= Y(\circ,\circ),\\
\hat{f}(\cdot,\circ) &:= \operatorname*{arg\,min}_{\mathrm{f}\in\mathbb{R}^{n_1\times n_2}} \left\{\|Y(\cdot,\circ)-\mathrm{f}\|_2^2/n + 2\lambda_1 \mathrm{TV}_1(\mathrm{f})\right\},\\
\hat{f}(\circ,\cdot) &:= \operatorname*{arg\,min}_{\mathrm{f}\in\mathbb{R}^{n_1\times n_2}} \left\{\|Y(\circ,\cdot)-\mathrm{f}\|_2^2/n + 2\lambda_2 \mathrm{TV}_2(\mathrm{f})\right\},\\
\hat{\tilde{f}} &:= \operatorname*{arg\,min}_{\mathrm{f}\in\mathbb{R}^{n_1\times n_2}} \left\{\|\tilde{Y}-\mathrm{f}\|_2^2/n + 2\lambda \mathrm{TV}(\mathrm{f})\right\}.
\end{aligned}$$

We can also apply the ANOVA decomposition to the underlying image f^0:

$$f^0 = f^0(\circ,\circ)\psi^{1,1} + f^0(\cdot,\circ) + f^0(\circ,\cdot) + \tilde{f}^0.$$

Then we can estimate $f^0(\circ,\circ)$ by $\hat{f}(\circ,\circ)$, $f^0(\cdot,\circ)$ by $\hat{f}(\cdot,\circ)$, $f^0(\circ,\cdot)$ by $\hat{f}(\circ,\cdot)$ and $\tilde{f}^0$ by $\hat{\tilde{f}}$.

Ordinary least squares is an appropriate method for estimating $f^0(\circ,\circ)$. Indeed the rate of convergence of $Y(\circ,\circ)$ to $f^0(\circ,\circ)$ is n^{-1}. The estimation of $f^0(\cdot,\circ)$ and $f^0(\circ,\cdot)$ by ordinary least squares would lead to rates of convergence of order n_1/n and n_2/n, respectively. In this paper we show that both the fast and the slow rate of estimation of $\tilde{f}^0$ by $\hat{\tilde{f}}$ are faster than $n^{-1/2}$, which is the best-case rate of estimation of the main effects by ordinary least squares. Without the regularization terms $\lambda_1\mathrm{TV}_1(f)$ and $\lambda_2\mathrm{TV}_2(f)$, the speed of estimation of f^0 would be limited by the estimation of the main effects. We therefore propose a regularized method, the so-called fused Lasso (Tibshirani et al., 2005), to estimate both $f^0(\cdot,\circ)$ and $f^0(\circ,\cdot)$ at a faster rate than $n^{-1/2}$.

Also the noise term ϵ can be decomposed into four orthogonal components:

$$\epsilon = \epsilon(\circ,\circ)\psi^{1,1} + \epsilon(\cdot,\circ) + \epsilon(\circ,\cdot) + \tilde{\epsilon}.$$

All the four terms of the decomposition present some correlation structure. This is however not a problem for the analysis of the respective estimators, since the four terms can be seen as the projections onto four mutually orthogonal linear subspaces of $\mathbb{R}^{n_1\times n_2}$. Indeed, in the analysis of $\hat{f}(\cdot,\circ)$, the empirical process $\mathrm{trace}(\epsilon(\cdot,\circ)'f(\cdot,\circ)/n)$ appears. By the idempotence of projection matrices, we have that $\mathrm{trace}(\epsilon(\cdot,\circ)'f(\cdot,\circ)/n) = \mathrm{trace}((\epsilon(\circ,\circ)\psi^{1,1} + \epsilon(\cdot,\circ))'f(\cdot,\circ)/n)$, where $\epsilon(\circ,\circ)\psi^{1,1}+\epsilon(\cdot,\circ)$ has rowwise iid entries. Similarly, in the analysis of $\hat{\tilde{f}}$ it holds that $\mathrm{trace}(\tilde{\epsilon}'\tilde{f}/n) = \mathrm{trace}(\epsilon'\tilde{f}/n)$.

A slow rate for $\hat{f}(\cdot,\circ)$ and $\hat{f}(\circ,\cdot)$ of order $n^{-2/3}$ is shown in Mammen and van de Geer (1997) using entropy calculations but with large constants and in Ortelli and van de Geer (2019b, 2020) with small constants but an additional logarithmic term.

The adaptivity of the estimators $\hat{f}(\cdot,\circ)$ and $\hat{f}(\circ,\cdot)$ has been established in Lin et al. (2017); Guntuboyina et al. (2020); Dalalyan et al. (2017); Ortelli and van de Geer (2018, 2019b, 2020). Let s denote the number of jumps in any column of $f^0(\cdot,\circ)$ or the number of jumps in any row of $f^0(\circ,\cdot)$. We give the fast rates exposed in these papers for the case that the s jumps lie on a regular grid.

Lin et al. (2017) obtain the rate $\mathcal{O}\left(\mathrm{s}\left((\log \mathrm{s} + \log\log n)\log n + \sqrt{\mathrm{s}}\right)/n\right)$, under the choice of the tuning parameter $\lambda \asymp n^{-\frac{1}{2}}\mathrm{s}^{-\frac{1}{4}}$, for n large enough.

Under the choice of the tuning parameter $\lambda \asymp \sqrt{\log n/n}$, Dalalyan et al. (2017) obtain an oracle inequality with the rate $\mathcal{O}\left(s\log n\left(\mathrm{s} + \log n\right)/n\right)$.

Under the choice of the tuning parameter $\lambda \asymp \sqrt{\log n/(\mathrm{s}n)}$, Ortelli and van de Geer (2018, 2020, 2019b) obtain the rate $\mathcal{O}\left(\mathrm{s}\log^2 n/n\right)$, which is improved by a log term by Guntuboyina et al. (2020) for a choice of the tuning parameter depending on f^0.

Since the estimation of $f^0(\circ,\circ)$, $f^0(\cdot,\circ)$, $f^0(\circ,\cdot)$ can be undertaken in a satisfactory way with estimators already widely studied in the literature, we are going to focus on establishing a slow rate and the adaptivity for the estimator of the interaction terms $\hat{\tilde{f}}$. For slow rates it will turn out that the part limiting the speed of estimation of f^0 is the estimation of the interaction terms, while for fast rates we will show that the interaction terms too can be estimated in an adaptive manner.

4. A Taste of the Main Result

We present our main result, Theorem 19 from Section 7, for the special case of a square image ($n_1 = n_2$) and an active set S defining a regular grid of cardinality $\sqrt{s}\times\sqrt{s}$. To be understood in its generality, Theorem 19 requires the background knowledge from Section 6.

Theorem 6 (Main result for fast rates: a special case) *Let $n_1 = n_2$. Let $g \in \mathbb{R}^{n_1\times n_2}$ be arbitrary. Let S be an arbitrary subset of size $s := |S|$ of $[3 : n_1 - 1]\times[3 : n_2 - 1]$ defining a regular grid of cardinality $\sqrt{s}\times\sqrt{s}$ parallel to the coordinate axes. Choose*

$$\lambda \geq 4\sigma\sqrt{\frac{\log(2n)}{n\sqrt{s}}}.$$

Then, with probability at least $1 - 1/n$, it holds that

$$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n \leq \|g - \tilde{f}^0\|_2^2/n + 4\lambda\|(\Delta g)_{-S}\|_1 + \left(\sigma\sqrt{\frac{s}{n}} + \sigma\sqrt{\frac{2\log(2n)}{n}} + \lambda\sqrt{\frac{8s^2 n\log(e^2 n)}{(\sqrt{n}-1)^2}}\right)^2.$$

If we choose $g = \tilde{f}^0$, S to be the active set of $\tilde{f}^0$ (given it is a regular grid) and $\lambda = 4\sigma\sqrt{\log(2n)}/\sqrt{n\sqrt{s}}$, then, with probability at least $1 - 1/n$, we have that $\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n = \mathcal{O}\left(s^{3/2}\log^2(n)/n\right)$. If we instead make the choice $\lambda = 4\sigma\sqrt{\log(2n)/n}$, which does not depend on S, then, with probability at least $1 - 1/n$, we have that $\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n = \mathcal{O}\left(s^2\log^2(n)/n\right)$.

In both cases, the dependence on s is worse than the linear dependence which has been proven for the one-dimensional case by Dalalyan et al. (2017); Ortelli and van de Geer (2019b); Guntuboyina et al. (2020). However, if s is constant the rate is parametric, up to the log factor. Fang et al. (2019) also prove a parametric rate up to log terms, however in a slightly different setting.

Theorem 6 gives us theoretical guarantees holding for all $g \in \mathbb{R}^{n_1\times n_2}$ and active sets S defining a regular grid, no matter the structure of f^0. Therefore g and S (under some

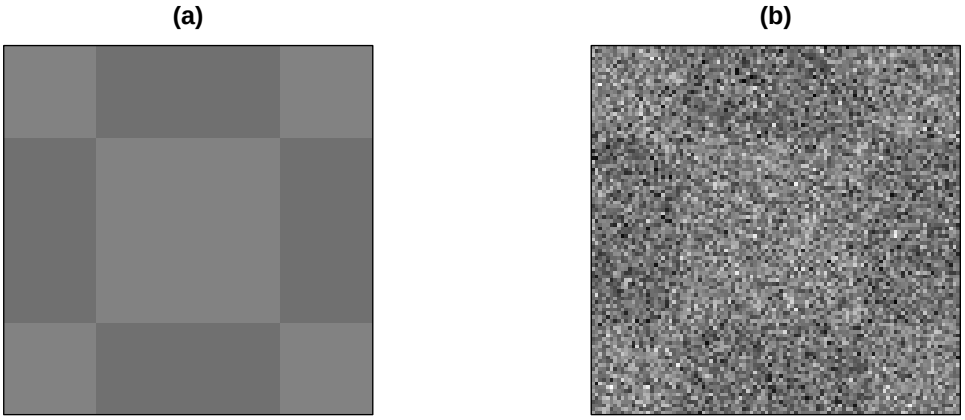

Figure 2: Plot (a) shows $\tilde{f}^0$ for f^0 as in Equation (1). Plot (b) shows $\tilde{Y} = \tilde{f}^0 + \tilde{\epsilon}$ for a realization of $\tilde{\epsilon}$ with $\sigma = 1$.

constraints) can be seen as free parameters. The upper bound can be minimized over all $g \in \mathbb{R}^{n_1 \times n_2}$ and all active sets S defining a regular grid. However, minimizing the upper bound requires the knowledge of $\tilde{f}^0$. A pair (f^*, S^*) minimizing the upper bound is called "an oracle", since $\tilde{f}^0$ is typically unknown. Theorem 6 is an oracle inequality in the sense that it guarantees that the (properly tuned) estimator behaves almost as good as if it would know the aspects of $\tilde{f}^0$ required to minimize the upper bound and optimally trade off all of its terms.

Theorem 6 is also an adaptive result: the estimator $\hat{\tilde{f}}$ is shown to adapt to the underlying image $\tilde{f}^0$, in particular to the number and location of its jumps. The adaptation to $\tilde{f}^0$ is achieved by means of the optimal tradeoff between the approximation of $\tilde{f}^0$ by the oracle f^* and the almost parametric rate of estimation of the rectangular pieces defined by the oracle active set S^*.

We will expose the more general version of this theorem holding for active sets S not necessarily defining a regular grid in Section 7.

For $n_1 = n_2$ being a multiple of 4 consider the image $f^0 \in \mathbb{R}^{n_1 \times n_2}$ defined as

$$f^0_{j,k} = 1_{\{n_1/4+1 \leq j \leq 3n_1/4\}} 1_{\{n_2/4+1 \leq k \leq 3n_2/4\}}, \ (j,k) \in [n_1] \times [n_2]. \tag{1}$$

Figure 2 shows $\tilde{f}^0$ and $\tilde{Y} = \tilde{f}^0 + \tilde{\epsilon}$ for $n_1 = 100$ and $\sigma = 1$. Figure 3 shows some simulation results for denoising the image $\tilde{Y} = \tilde{f}^0 + \tilde{\epsilon}$ with $\sigma = 1$, where, for $n_1 = n_2 \in \{4, 8, \ldots, 196, 200\}$, f^0 is taken as in Equation (1). For such images, Δf^0 has 4 nonzero components. Therefore we chose $s = 4$. The estimator was computed via a detour through its synthesis formulation (see Section 5), which allowed to use the R package `glmnet`.

The results of the simulation support our findings: if tuned with $\lambda = \sqrt{\log(2n)/(2n)}$, the estimator $\hat{\tilde{f}}$ converges at an almost parametric rate to the underlying piecewise rectangular image $\tilde{f}^0$. However, the rate of convergence n^{-1} (up to log terms) is achieved only for n large enough and the tuning parameter has to be chosen smaller by a constant factor than the smallest theoretical choice $\lambda = 4\sqrt{\log(2n)/(2n)}$ suggested by Theorem 6.

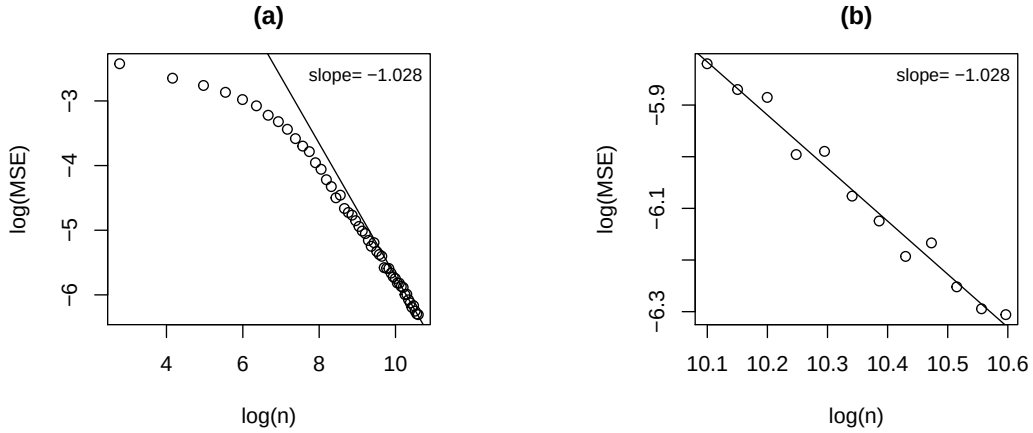

Figure 3: Plot (a) displays the logarithm of the average mean squared error of the estimator over 40 realizations of the noise term versus $\log(n)$, for $n_1 = n_2 \in \{4, 8, \ldots, 196, 200\}$. The least squares fit with slope -1.028 is based on the values for $n_1 = n_2 \in \{156, \ldots, 200\}$ shown in detail in plot (b).

5. Synthesis Form

Recall the analysis estimator

$$\hat{\hat{f}} := \underset{\mathrm{f} \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left\{ \|\tilde{Y} - \mathrm{f}\|_2^2 / n + 2\lambda \mathrm{TV}(\mathrm{f}) \right\}.$$

Analysis estimators approximate the observations under a penalty on the norm of a linear operator—a so-called analysis operator—applied to the candidate estimator f, in this case $\mathrm{TV}(\mathrm{f}) = \|\Delta \mathrm{f}\|_1$.

Analysis estimators can be rewritten as synthesis estimators (Elad et al., 2007). A well-known instance of synthesis estimator is the Lasso (Tibshirani, 1996). The synthesis approach to estimation is constructive: the signal is approximated by a linear combination of atoms under a penalty on the norm of the coefficients of this linear combination. By looking at the properties of the collection of atoms—the so-called dictionary—one can gain some insights into the structure of the estimator.

In our case, the synthesis formulation shows that the estimator $\hat{\hat{f}}$ produces piecewise rectangular estimates. Moreover, the synthesis formulation of $\hat{\hat{f}}$ will be of great help in computing "bounds on the antiprojections" (see Definition 10 and Lemmas 29 and 31), which are essential ingredients of Lemma 13 to control the increments of the empirical process. For a detailed discussion on the relation between analysis and synthesis estimators we refer to Elad et al. (2007) and to Ortelli and van de Geer (2019a), who focus on analysis and synthesis in total variation regularization.

We first express a matrix f as linear combination of dictionary matrices. We then show that $\hat{\hat{f}}$ can be written as a synthesis estimator using these dictionary matrices. Here the notation with arguments instead of subscripts comes in handy.

Consider some $f \in \mathbb{R}^{n_1 \times n_2}$. We may write for $j \in [n_1]$ and $k \in [n_2]$,

$$f(j,k) = \sum_{j'=1}^{n_1} \sum_{k'=1}^{n_2} \beta_{j',k'} \psi^{j',k'}(j,k),$$

where for $(j',k') \in [n_1] \times [n_2]$ the dictionary matrices are $\psi^{j',k'}$ with

$$\psi^{j',k'}(j,k) = 1_{\{j \geq j', k \geq k'\}}, \ (j,k) \in [n_1] \times [n_2]$$

and

$$\beta_{j',k'} := \begin{cases} f(1,1), & (j'k') = (1,1), \\ f(j',1) - f(j'-1,1), & (j',k') \in [2:n_1] \times [1], \\ f(1,k') - f(1,k'-1), & (j',k') \in [1] \times [2:n_2], \\ (\Delta f)_{j',k'}, & (j',k') \in [2:n_1] \times [2:n_2]. \end{cases}$$

We call the collection of matrices $\{\psi^{j',k'}\}_{(j',k') \in [n_1] \times [n_2]}$ the dictionary. The dictionary consists of a collection of indicator functions of half intervals. Therefore, a sparse linear combination of elements of the dictionary will be piecewise constant on rectangular sets.

Define

$$\tilde{\psi}^{j,k} := \begin{cases} \psi^{1,1}, & (j,k) = (1,1), \\ \psi^{j,1} - \psi^{j,1}(\circ,\circ) = \mathrm{A}_{\mathrm{span}(\psi^{1,1})} \psi^{j,1}, & (j,k) \in [2:n_1] \times [1], \\ \psi^{1,k} - \psi^{1,k}(\circ,\circ) = \mathrm{A}_{\mathrm{span}(\psi^{1,1})} \psi^{1,k}, & (j,k) \in [1] \times [2:n_2], \\ \psi^{j,k} - \psi^{j,k}(\cdot,\circ) - \psi^{j,k}(\circ,\cdot) - \psi^{j,k}(\circ,\circ) & \\ \quad = \mathrm{A}_{\mathrm{span}(\{\psi^{j,1}\}_{j \in [n_1]}, \{\psi^{1,k}\}_{k \in [n_2]})} \psi^{j,k}, & (j,k) \in [2:n_1] \times [2:n_2]. \end{cases}$$

The four resulting linear spaces $\mathrm{span}(\tilde{\psi}^{1,1})$, $\mathrm{span}(\{\tilde{\psi}^{j,1}\}_{j \in [2:n_1]})$, $\mathrm{span}(\{\tilde{\psi}^{1,k}\}_{k \in [2:n_2]})$ and $\mathrm{span}(\{\tilde{\psi}^{j,k}\}_{(j,k) \in [2:n_1] \times [2:n_2]})$ are mutually orthogonal. Moreover, the atoms of the dictionary $\{\tilde{\psi}^{j,k}\}_{(j,k) \in [n_1] \times [n_2]}$ are piecewise constant on rectangular sets.

Lemma 7 gives the form of the coefficients needed to express an image f as a linear combination of the matrices $\{\tilde{\psi}^{j,k}\}_{(j,k) \in [n_1] \times [n_2]}$.

Lemma 7 (Construct a piecewise rectangular image) *It holds that*

$$f = \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \tilde{\beta}_{j,k} \tilde{\psi}^{j,k},$$

where

$$\tilde{\beta}_{j,k} = \begin{cases} f(\circ,\circ), & (j,k) = (1,1), \\ f(j,\circ) - f(j-1,\circ), & (j,k) \in [2:n_1] \times [1], \\ f(\circ,k) - f(\circ,k-1), & (j,k) \in [1] \times [2:n_2], \\ (\Delta f)_{j,k}, & (j,k) \in [2:n_1] \times [2:n_2]. \end{cases}$$

Proof See Appendix A.1. ■

The next lemma, based on Lemma 7, gives a synthesis form of the estimator $\tilde{\hat{f}}$.

Lemma 8 (Synthesis formulation) *We have*

$$\hat{\tilde{f}} = \sum_{j=2}^{n_1} \sum_{k=2}^{n_2} \hat{\tilde{\beta}}_{j,k} \tilde{\psi}^{j,k},$$

where

$$\hat{\tilde{\beta}}_{j,k} = \underset{\{\beta_{j,k}\}_{(j,k)\in[2:n_1]\times[2:n_2]}}{\arg\min} \left\{ \|Y - \sum_{j=2}^{n_1} \sum_{k=2}^{n_2} \beta_{j,k} \tilde{\psi}^{j,k}\|_2^2/n + 2\lambda \sum_{j=2}^{n_1} \sum_{k=2}^{n_2} |\beta_{j,k}| \right\}.$$

Proof See Appendix A.2. ■

Note that $\sum_{j=2}^{n_1} \sum_{k=2}^{n_2} |\hat{\tilde{\beta}}_{j,k}| = \mathrm{TV}(\hat{\tilde{f}}) = \mathrm{TV}(\hat{f})$.

6. Oracle Inequalities

In this section we expose standard techniques used to derive oracle inequalities with fast and slow rates. We closely follow Hütter and Rigollet (2016); Dalalyan et al. (2017); Ortelli and van de Geer (2019b, 2020). This section can therefore be viewed as a preparatory step, which frames the work that has to be done in order to establish adaptivity as well as the rate $n^{-5/8} \log^{3/8} n$ for the estimator of the interaction terms.

Indeed, we do not yet exploit the specific properties of the two-dimensional total derivative operator Δ. These properties will be further explored in Sections 7 and 8. The current section can be seen as the background knowledge already present in the literature. It is complemented by our results in Sections 7 and 8, which are new and specific for total variation image denoising.

For simplicity, in this section we look at matrices as if they were vectors by concatenating their entries by columns. We define the dictionary

$$\tilde{\Psi} := \{\tilde{\psi}^{j,k}\}_{(j,k)\in\{2,\dots,n_1\}\times\{2,\dots,n_2\}} \in \mathbb{R}^{n_1 n_2 \times (n_1-1)(n_2-1)}$$

and the two-dimensional discrete derivative operator

$$\Delta := \big(D_1 \otimes D_2\big) \in \{-1, 0, +1\}^{(n_1-1)(n_2-1)\times n_1 n_2},$$

where $\otimes$ denotes the Kronecker product. Note that the dictionary $\tilde{\Psi}$ is the remainder of the projection of the last $(n_1-1)(n_2-1)$ columns of the dictionary

$$\Psi := \{\psi^{j,k}\}_{(j,k)\in[n_1]\times[n_2]} \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$$

onto its first $n_1 + n_2 - 1$ columns.

Recall the estimator

$$\hat{\tilde{f}} = \arg\min_{\mathrm{f}\in\mathbb{R}^n} \left\{ \|\tilde{Y} - \mathrm{f}\|_2^2/n + 2\lambda \|\Delta \mathrm{f}\|_1 \right\}, \lambda > 0.$$

To guarantee favorable error bounds the tuning parameter λ has to be chosen carefully. It has to be chosen large enough to overrule the noise, but not too large.

A choice of the tuning parameter λ that guarantees that all the noise is overruled is the "universal choice"

$$\lambda_0(t) := \sigma\sqrt{\frac{2\log(2n)+2t}{n}}, t > 0.$$

This choice results from the assumption that all noise has to be overruled by the penalty: the the structure encoded in the analysis operator Δ and in the active set S is not considered. The projection arguments by Dalalyan et al. (2017) exposed in Lemma 13 in Subsection 6.1 take into account the structure encoded in Δ and show that only the part of the noise not being correlated with the candidate structure of the estimator needs to be overruled by the tuning parameter λ. Thus, more favorable error bounds can be obtained with a choice of λ smaller than the universal choice $\lambda_0(t)$. How much $\lambda_0(t)$ has to be downscaled depends then on the correlation in the structure encoded in Δ and S.

The universal choice $\lambda_0(t)$ can therefore be seen as a worst-case choice, which always overrules the noise. It always does its job, but not always the best job.

The following inequality is the starting point for the proof of oracle inequalities with both fast and slow rates and can be found for instance in van de Geer (2016); Hütter and Rigollet (2016); Ortelli and van de Geer (2018, 2019b, 2020).

Lemma 9 (Basic inequality) *For all $g \in \mathbb{R}^n$ we have that*

$$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n + \|\hat{\tilde{f}} - g\|_2^2/n \leq \|g - \tilde{f}^0\|_2^2/n + 2\frac{\tilde{\epsilon}^T(\hat{\tilde{f}} - g)}{n} + 2\lambda(\|\Delta g\|_1 - \|\Delta\hat{\tilde{f}}\|_1).$$

Proof See Appendix B.1. ■

6.1 Bounding the Increments of the Empirical Process

Let $S \subseteq [(n_1-1)(n_2-1)]$. Let $\tilde{\Psi}_i$ denote the i^{th} column of $\tilde{\Psi}$. We write $\tilde{\Psi}_S := \{\tilde{\Psi}_i\}_{i\in S}$ and $\tilde{\Psi}_{-S} := \{\tilde{\Psi}_i\}_{i\notin S}$. Denote by $\mathrm{P}_S := \tilde{\Psi}_S(\tilde{\Psi}_S^T\tilde{\Psi}_S)^{-1}\tilde{\Psi}_S^T$ the orthogonal projection matrix onto the column span of $\tilde{\Psi}_S$ and by $\mathrm{A}_S := \mathrm{I}_n - \mathrm{P}_S$ the corresponding antiprojection matrix.

Empirical processes and their relevance for statistics are discussed for instance in van de Geer (2007, 2009). In this subsection we are going to expose a high-probability upper bound for the increments of the empirical process by projection arguments proposed by Dalalyan et al. (2017).

The increments of the empirical process we study are given by

$$\left\{\frac{\tilde{\epsilon}^T f}{n} : f \in \mathbb{R}^n\right\} = \left\{\frac{\tilde{\epsilon}^T \tilde{f}}{n} : f \in \mathbb{R}^n\right\} = \left\{\frac{\epsilon^T \tilde{f}}{n} : f \in \mathbb{R}^n\right\},$$

where the equality holds because of the idempotence of projection matrices. The basis of the technique to bound the increments of the empirical process by Dalalyan et al. (2017) is to decompose them into a part projected onto a low-rank linear space and a remainder, the so-called antiprojection:

$$\frac{\tilde{\epsilon}^T\tilde{f}}{n} = \frac{\tilde{\epsilon}^T\mathrm{P}_S\tilde{f}}{n} + \frac{\tilde{\epsilon}^T\mathrm{A}_S\tilde{f}}{n}. \tag{2}$$

We now define the bound on the antiprojections, the inverse scaling factor and the noise weights, which are needed to control the increments of the empirical process by projection arguments.

Definition 10 (Bound on the antiprojections) *A bound on the antiprojections $\tilde{v} \in \mathbb{R}^{(n_1-1)(n_2-1)}$ is a vector (or matrix), such that*

$$\tilde{v}_i \geq \|(\mathrm{I}_n - \mathrm{P}_S)\tilde{\Psi}_i\|_2/\sqrt{n}, \ \forall i \in [(n_1-1)(n_2-1)].$$

Based on the bound on the antiprojections $\tilde{v}$ we define the inverse scaling factor and the noise weights, which will be important in determining the choice of the tuning parameter λ and the bound on the effective sparsity, respectively.

Definition 11 (Inverse scaling factor) *Let $\tilde{v}$ be a bound on the antiprojections. The inverse scaling factor $\tilde{\gamma} \in \mathbb{R}$ is defined as $\tilde{\gamma} := \|\tilde{v}_{-S}\|_\infty$.*

The inverse scaling factor $\tilde{\gamma}$ depends on the analysis operator Δ via the dictionary $\tilde{\Psi}$ and on the active set S.

Definition 12 (Noise weights) *Let $\tilde{v}$ be a bound on the antiprojections and $\tilde{\gamma}$ the corresponding inverse scaling factor. A vector of noise weights is any vector $v \in \mathbb{R}^{(n_1-1)(n_2-1)}$ such that $v \geq \tilde{v}/\tilde{\gamma} \in [0,1]^{(n_1-1)(n_2-1)}$, componentwise.*

The following lemma is inspired by the proof of Theorem 1 in Dalalyan et al. (2017) and can be found in a more general form as Lemma A.2 in Ortelli and van de Geer (2019b).

Lemma 13 (Control the increments of the empirical process with projections) *For $x, t > 0$ choose*

$$\lambda \geq \tilde{\gamma}\lambda_0(t).$$

Then, $\forall f \in \mathbb{R}^{n_1 n_2}$, with probability at least $1 - e^{-x} - e^{-t}$ it holds that

$$\frac{\tilde{\epsilon}^T f}{n} \leq \frac{\|f\|_2}{\sqrt{n}}\left(\sigma\sqrt{\frac{2x}{n}} + \sigma\sqrt{\frac{s}{n}}\right) + \lambda\|v_{-S} \odot (\Delta f)_{-S}\|_1.$$

Proof See Appendix B.2. ■

Lemma 13 can be interpreted as a bound on the increments of the empirical process tailored to the structure of the estimation problem. Indeed, the linear space onto which the noise is projected is chosen depending on the analysis operator Δ and on the candidate active set S. As a consequence of the projection arguments used in its proof, one can choose the tuning parameter smaller than the universal choice $\lambda_0(t)$ by a factor $\tilde{\gamma}$, which depends on the structure encoded in Δ (or $\tilde{\Psi}$) and S. The universal choice of the tuning parameter is retrieved by choosing $S = \emptyset$, which is equivalent to neglecting all structure.

6.2 Fast Rates

Oracle inequalities with fast rates are characterized by the presence of the so-called effective sparsity in the upper bound. To define the effective sparsity we need the notion of sign configurations. Indeed, we will apply the definition of effective sparsity to an image, the signs of whose jumps we do not know. Therefore we look for a bound on the effective sparsity holding for all sign configurations.

Definition 14 (Sign configuration) *Let* $q \in [-1,1]^{(n_1-1)(n_2-1)}$ *be s.t.*

$$q_i \in \begin{cases} \{-1,+1\}, & i \in S, \\ [-1,1], & i \notin S. \end{cases}$$

We call $q_S \in \{-1,0,1\}^{(n_1-1)(n_2-1)}$ *a sign configuration.*

We now define the effective sparsity as in Ortelli and van de Geer (2019b). This definition will be reformulated in matrix form in Definition 20 in Section 7.

Definition 15 (Effective sparsity) *Let* S *be an active set,* $q_S \in \{-1,0,1\}^{(n_1-1)(n_2-1)}$ *be a sign configuration and* $v \in [0,1]^{(n_1-1)(n_2-1)}$ *be noise weights. The effective sparsity* $\Gamma(S, v_{-S}, q_S) \in \mathbb{R}$ *is defined as*

$$\Gamma(S, v_{-S}, q_S) = \max\{q_S^T (\Delta f)_S - \|(1-v)_{-S} \odot (\Delta f)_{-S}\|_1 : \|f\|_2^2/n = 1\}.$$

Moreover we write

$$\Gamma(S, v_{-S}) := \max_{q_S} \ \Gamma(S, v_{-S}, q_S).$$

The definition of effective sparsity consists of two parts: a first term representing approximately the number of jumps—the sparsity—and a second term which is a discount due to the correlation of the non-active dictionary atoms with the active ones. Hence the name "effective sparsity". The larger this correlation, the larger the discount for the effective sparsity.

An oracle inequality with fast rate is shown in the following theorem, which corresponds to Theorem 2.1 in Ortelli and van de Geer (2020) and to the adaptive bound of Theorem 2.2 in Ortelli and van de Geer (2019b).

Theorem 16 (Oracle inequality with fast rates) *Let* $g \in \mathbb{R}^n$ *and* $S \subseteq [(n_1-1)(n_2-1)]$ *be arbitrary. For* $x, t > 0$*, choose* $\lambda \geq \tilde{\gamma}\lambda_0(t)$*. Then, with probability at least* $1 - e^{-x} - e^{-t}$*, it holds that*

$$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n \leq \|g - \tilde{f}^0\|_2^2/n + 4\lambda\|(\Delta g)_{-S}\|_1 + \left(\sigma\sqrt{\frac{2x}{n}} + \sigma\sqrt{\frac{s}{n}} + \lambda\Gamma(S, v_{-S}, q_S)\right)^2,$$

where $q_S = sign((\Delta g)_S)$.

Proof See Appendix B.3. ■

The fast rate of Theorem 16 is given by $\lambda^2\Gamma^2(S, v_{-S}) \asymp \log(n)\tilde{\gamma}^2\Gamma^2(S, v_{-S})/n$. Typically, we expect $\Gamma^2(S, v_{-S})$ to scale approximately as $\Gamma^2(S, v_{-S}) \asymp s/\tilde{\gamma}^2$. We will prove in Lemma 25 in Section 7 that for image denoising with total variation regularization the effective sparsity scales as $\Gamma^2(S, v_{-S}) \asymp s^{3/2}\log(n)/\tilde{\gamma}^2$. We have an extra factor $s^{1/2}$ due to the two-dimensional nature of the problem and a log factor due to the noise.

Using Lemma 13 to bound the increments of the empirical process has two effects: on the one side we can choose a tuning parameter $\lambda = \tilde{\gamma}\lambda_0(t)$ smaller than the universal choice $\lambda_0(t)$. Thus, the rate that would be obtained with $\lambda = \lambda_0(t)$ can be obtained with $\lambda = \tilde{\gamma}\lambda_0(t)$ and a bound on the effective sparsity larger by a factor $1/\tilde{\gamma}^2$. On the other side, the effective sparsity is increased by an additive $\|v_{-S} \odot (\Delta f)_{-S}\|_1$.

To prove adaptivity, we need to find an appropriate bound on the antiprojections $\tilde{v}$, the corresponding scaling factor $\tilde{\gamma}$, the noise weights v and finally prove a bound on the effective sparsity $\Gamma(S, v_{-S}, q_S)$ holding for all sign confgurations q_S. This will be the topic of Section 7.

6.3 Slow Rates

The next theorem corresponds to Theorem 2.2 in Ortelli and van de Geer (2020) and to the non-adaptive bound of Theorem 2.2 in Ortelli and van de Geer (2019b).

Theorem 17 (Oracle inequality with slow rates) *Let $g \in \mathbb{R}^n$ and $S \subseteq [(n_1 - 1)(n_2 - 1)]$ be arbitrary. For $x, t > 0$, choose $\lambda \geq \tilde{\gamma}\lambda_0(t)$. Then, with probability at least $1-e^{-x}-e^{-t}$, it holds that*

$$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n \leq \|g - \tilde{f}^0\|_2^2/n + 4\lambda\|\Delta g\|_1 + \left(\sigma\sqrt{\frac{2x}{n}} + \sigma\sqrt{\frac{s}{n}}\right)^2.$$

Proof See Appendix B.4. ■

To obtain the rate $n^{-5/8}\log^{3/8} n$, we need to choose S in a way that optimally trades off the term s/n and the term $\tilde{\gamma}\lambda_0(t)\|\Delta g\|_1$. This will be the topic of Section 8.

7. Adaptive Rates for Total Variation Image Denoising

Our objective for this section is to establish that $\hat{f}$ can adapt to the number of jumps in the main effects and the interaction terms. The main effects can be dealt with by using the results for the one-dimensional total variation regularized estimator, see Dalalyan et al. (2017); Guntuboyina et al. (2020); Ortelli and van de Geer (2019b). Thus, our main result will be to show that the estimator $\hat{\tilde{f}}$ of the interaction terms is adaptive in that it can adapt to the underlying true interaction terms $\tilde{f}^0$. We will prove an upper bound on the mean squared error of $\hat{\tilde{f}}$ which can be different for different values of $\tilde{f}^0$. In practice $\tilde{f}^0$ is unknown. However adaptivity guarantees that the estimator can "sense" different structures in the underlying $\tilde{f}^0$ and adapt to them.

To prove adaptivity, we establish a bound for the effective sparsity (Definition 20) using interpolating matrices (see Lemma 22). This way of bounding the effective sparsity is an extension to the two-dimensional case of the bound on the effective sparsity for one-dimensional total variation regularized estimators based on interpolating vectors exposed in Ortelli and van de Geer (2019b). The combination of the new bound on the effective sparsity (see Lemma 25) with the standard Theorem 16 will lead to our main result.

The roadmap for this section is the following: in Subsection 7.1 we will state our main result. The result follows by combining the general oracle inequality for analysis estimators given in Theorem 16 with the results of Subsections 7.2–7.4 and will be proved in the conclusive Subsection 7.5. In Subsection 7.2 we define interpolating matrices and show how to carry out discrete partial integration in two dimensions. In Subsection 7.3 we prove a bound on the effective sparsity and in Subsection 7.4 we show how to find suitable noise weights.

7.1 Main Result

We present the main result: an oracle inequality for the estimator $\hat{\tilde{f}}$ of the interaction terms $\tilde{f}^0$.

We fix an active set $S \subseteq [3 : n_1 - 1] \times [3 : n_2 - 1]$. The discussion that follows, and in particular also Theorem 19, depends on the choice of S, which can therefore be considered as a "free parameter".

Given an active set $S \subseteq [3 : n_1 - 1] \times [3 : n_2 - 1]$, we can partition $[2 : n_1] \times [2 : n_2]$ into s subsets, consisting of the points closest to t_m, $m = 1, \ldots, s$ with respect to the city block metric. This corresponds to a Voronoi tessellation. However, a Voronoi tessellation typically has subsets of relatively irregular shape. We will require that the partition consists of rectangles to ease the construction of an interpolating matrix. The concept of interpolating matrix is presented in Section 7.2 and will be applied in the bound for the effective sparsity in Section 7.3.

Definition 18 (Rectangular tessellation) *We call $\{R_m\}_{m=1}^s$ a rectangular tessellation of $[2 : n_1] \times [2 : n_2]$ if it satisfies the following conditions:*
- *each $R_m \subseteq [2 : n_1] \times [2 : n_2]$ is a rectangle ($m = 1, \ldots, s$);*
- *$\cup_{m=1}^s R_m = [2 : n_1] \times [2 : n_2]$;*
- *for all m and $m' \neq m$, the rectangles R_m and $R_{m'}$ ($m \neq m'$) possibly share boundary points, but not interior points;*
- *for all m, the jump location t_m is an interior point of R_m.*

For a rectangular tessellation $\{R_m\}_{m=1}^s$ we let d_m^{--} be the area of the rectangle R_m North-West of t_m, d_m^{+-} the area to the South-West, d_m^{++} the area to the South-East and d_m^{-+} the area to the North-East. In other words, if $(t_{1,m}^-, t_{2,m}^-)$, $(t_{1,m}^-, t_{2,m}^+)$, $(t_{1,m}^+, t_{2,m}^+)$, $(t_{1,m}^+, t_{2,m}^-)$ are the four corners of the rectangle R_m, starting with the top-left corner and going clockwise along the boundary, then

$$d_m^{--} = d_{1,m}^- d_{2,m}^-, \qquad d_m^{-+} = d_{1,m}^- d_{2,m}^+,$$
$$d_m^{+-} = d_{1,m}^+ d_{2,m}^-, \qquad d_m^{++} = d_{1,m}^+ d_{2,m}^+,$$

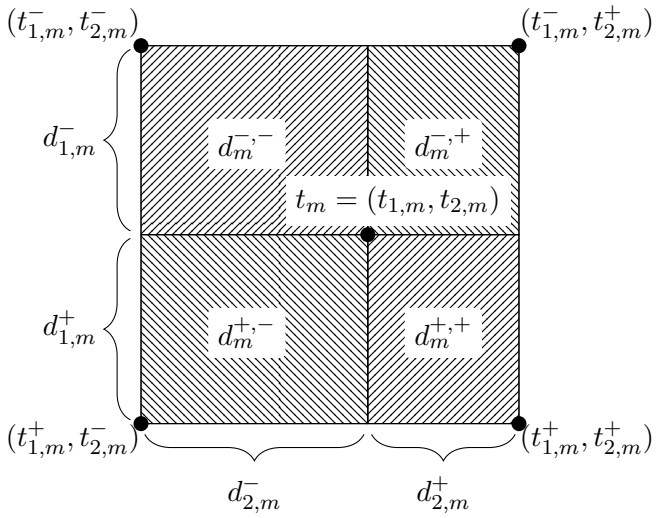

Figure 4: Illustration of a rectangle R_m of the rectangular tessellation $\{R_m\}_{m=1}^s$, defined in Definition 18.

where

$$d^-_{1,m} = (t_{1,m} - t^-_{1,m}), \qquad d^-_{2,m} = (t_{2,m} - t^-_{2,m}),$$
$$d^+_{1,m} = (t^+_{1,m} - t_{1,m}), \qquad d^+_{2,m} = (t^+_{2,m} - t_{2,m}).$$

The rectangle R_m is illustrated in Figure 4.

Fix a set $S \subseteq [3 : n_1 - 1] \times [3 : n_2 - 1]$ and a rectangular tessellation $\{R_m\}_{m=1}^s$. Let $d_{1,\max}(S) := \max_{m\in[1:s]} \max\{d^-_{1,m}, d^+_{1,m}\}$ and $d_{2,\max}(S) := \max_{m\in[1:s]} \max\{d^-_{2,m}, d^+_{2,m}\}$. The quantity $d_{1,\max}(S)$ (respectively $d_{2,\max}(S)$) denote the maximal horizontal (respectively vertical) distance from a jump location to the boundary of the corresponding rectangular region in the rectangular tessellation $\{R_m\}_{m=1}^s$.

For simplicity, we do not elaborate on the dependence on the sign configuration in the effective sparsity and focus instead on a worst-case upper bound holding for all sign configurations. In other words, we bound the worst case $\Gamma(S, v_{-S}) := \max_{q_S} \Gamma(S, v_{-S}, q_S)$ rather than $\Gamma(S, v_{-S}, q_S)$.

Theorem 19 (Adaptivity of image denoising with total variation) *Let $g \in \mathbb{R}^{n_1\times n_2}$ be arbitrary. Let $x,\ t > 0$. Choose*

$$\lambda \geq 2\sqrt{\frac{d_{1,\max}(S)}{n_1} + \frac{d_{2,\max}(S)}{n_2}}\ \lambda_0(t).$$

Then, with probability at least $1 - e^{-x} - e^{-t}$, it holds that

$$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n \leq \|g - \tilde{f}^0\|_2^2/n + 4\lambda\|(\Delta g)_{-S}\|_1 + \left(\sigma\sqrt{\frac{s}{n}} + \sigma\sqrt{\frac{2x}{n}} + \lambda\Gamma(S, v_{-S})\right)^2,$$

where

$$\Gamma^2(S, v_{-S}) \leq \frac{1}{2}\Big(\log(en_1) + \log(en_2)\Big) \sum_{m=1}^{s} \Big(\frac{n}{d_m^{--}} + \frac{n}{d_m^{-+}} + \frac{n}{d_m^{++}} + \frac{n}{d_m^{+-}} \Big). \tag{3}$$

We note that the upper bound on the mean squared error in the above theorem depends on $\tilde{f}^0$, g and S. The true underlying interaction terms $\tilde{f}^0$ are typically fixed, while the choices of g and—under some constraints—of S are arbitrary. As such, the upper bound can be optimized over g and S. A pair $(g = f^*, S = S^*)$ optimizing the upper bound is called "oracle" and depends on $\tilde{f}^0$. The optimized upper bound depends therefore on $\tilde{f}^0$ and consists of the optimal tradeoff between the approximation error of $\tilde{f}^0$ by the oracle signal f^* and the estimation error of the piecewise rectangular structure encoded in the oracle active set S^*. This piecewise rectangular structure is estimated almost at a parametric rate, this means, almost as if the number and the locations of the elements of S^* were known.

In this sense, Theorem 19 is an adaptive result: the bound on the mean squared error of the estimator $\hat{\tilde{f}}$ of the interaction terms varies depending on the underlying interaction terms $\tilde{f}^0$ to estimate. The estimator $\hat{\tilde{f}}$ can therefore sense the structure in $\tilde{f}^0$—as for instance the number and the location of its jumps—and adapt to it.

Theorem 19 gives also a theoretical justification for choosing the tuning parameter smaller than the universal choice $\lambda = \lambda_0(t)$. However the active set of the true image f^0 or of its oracle approximation might not be known in practice, so that one might have to choose $\lambda = \lambda_0(t)$. The choice $\lambda = \lambda_0(t)$ in the setting of Theorem 6 with $s_1 = s_2$ results then in an oracle bound of order s^2/n, up to log-terms.

Theorem 19 can be obtained from Theorem 16 by finding a bound on the (worst-case) effective sparsity $\Gamma(S, v_{-S})$, which is the main contribution of Section 7.

In Subsection 7.2, we expose some tools needed to bound the effective sparsity. Of particular interest is the concept of interpolating matrix, which is an adaptation of the interpolating vector by Ortelli and van de Geer (2019b) to two dimensions.

In Subsection 7.3 we will take the interpolating matrix as given and bound the effective sparsity based on it and on the tools exposed in Subsection 7.2.

The results of Subsection 7.4 will show that the interpolating matrix given in Subsection 7.3 is indeed a valid interpolating matrix.

Subsection 7.5 combines the results of Subsections 7.2–7.4 to prove Theorem 19.

7.2 Interpolating Matrix and Partial Integration

We now rewrite the definition of effective sparsity (Definition 15) in matrix instead of vector form. Let $q_S = \{(q_S)_{j,k}\}_{(j,k)\in[2:n_1]\times[2:n_2]}$ and $v = \{v_{j,k}\}_{(j,k)\in[2:n_1]\times[2:n_2]}$ be a sign configuration and noisy weights written in matrix form.

Definition 20 (Effective sparsity in matrix form) *The effective sparsity is defined as*

$$\Gamma(S, v_{-S}, q_S) = \max\{\text{trace}(q_S^T (D_1 f D_2^T)_S) - \|(1-v)_{-S} \odot (D_1 f D_2^T)_{-S}\|_1 : \|f\|_2^2/n = 1\}.$$

Moreover we write

$$\Gamma(S, v_{-S}) := \max_{q_S} \ \Gamma(S, v_{-S}, q_S).$$

We define an interpolating matrix. The interpolating matrix will be a tool for finding a bound on the effective sparsity. The concept of interpolating vector (and matrix) is inspired by the dual certificate by Candès and Fernandez-Granda (2014). Related concepts appear also earlier in the literature in Fuchs (2004); Candès and Recht (2013). Ortelli and van de Geer (2019b) make a connection between interpolating matrix and effective sparsity. This connection is here extended to the two-dimensional case.

Definition 21 (Interpolating matrix) *Let $q_S \in \{-1,0,1\}^{(n_1-1)\times(n_2-1)}$ be a sign configuration and $v \in [0,1]^{(n_1-1)\times(n_2-1)}$ be a matrix of weights. We call an interpolating matrix for the sign configuration q_S and the weights v a matrix $w(q_S) = \{w_{j,k}(q_S)\}_{(j,k)\in[2:n_1]\times[2:n_2]}$ having the following properties:*
- $w_{t_m}(q_S) = q_{t_m}$, $\forall m \in [1:s]$,
- $|w_{j,k}(q_S)| \leq 1 - v_{j,k}$, $\forall (j,k) \notin S$.

The interpolating matrix $w(q_S)$ can be interpreted to belong to the subdifferential of $\|\Delta h\|_1$ for some matrix $h \in \mathbb{R}^{n_1\times n_2}$ with the same sign configuration. The choices of the active set S, the sign configuration q_S and the matrix h are not tied to the estimator $\hat{\hat{f}}$.

For completeness, we give the matrix version of Lemma 4.2 by Ortelli and van de Geer (2019b).

Lemma 22 (How to bound the effective sparsity) *We have*

$$\Gamma^2(S, v_{-S}, q_S) \leq n \min_{w(q_S)} \|D_1^T w(q_S) D_2\|_2^2$$

where the minimum is over all interpolating matrices $w(q_S)$ for the sign configuration q_S.

Proof See Appendix C.2. ■

The proof of Lemma 22 uses the equation

$$\text{trace}(w^T D_1 f D_2^T) = \text{trace}(D_2^T w^T D_1 f).$$

When $D_1 f D_2^T = \Delta f$ this equality is called partial integration. We study it further in the next lemma.

Lemma 23 (Partial integration in two dimensions with zero boundaries) *Choose arbitrarily a matrix $w = \{w_{j,k}\}_{(j,k)\in[2:n_1]\times[2:n_2]}$ with its boundary entries equal to zero, i.e.,*

$$w_{j,2} = w_{j,n_2} = 0, \ \forall\ j \in [2:n_1] \text{ and } w_{2,k} = w_{n_1,k} = 0, \ \forall\ k \in [2:n_2].$$

Then

$$\text{trace}(w^T \Delta f) = \sum_{k=2}^{n_2}\sum_{j=2}^{n_1} w_{j,k}(\Delta f)_{j,k} = \sum_{j=2}^{n_1-1}\sum_{k=2}^{n_2-1} (\Delta w)_{j+1,k+1} f_{j,k}.$$

Proof See Appendix C.1. ■

To obtain a bound on the effective sparsity, we now have to find a suitable interpolating matrix. We then compute the Frobenius norm of its total derivative with the help of partial integration.

7.3 A Bound for the Effective Sparsity

Given a set $S \subseteq [3, n_1 - 1] \times [3, n_2 - 1]$, let $\{R_m\}_{m=1}^s$ be a rectangular tessellation. Each jump location t_m is an interior point of R_m. The rectangle R_m consists of a North-West rectangle R_m^{--} a North-East rectangle R_m^{-+} a South-East rectangle R^{++} and a South-West rectangle R^{+-}. Thus

$$\begin{aligned} R_m^{--} &:= \{(j,k) : \ t_{1,m}^- \le j \le t_{1,m}, \ t_{2,m}^- \le k \le t_{2,m}\}, \\ R_m^{-+} &:= \{(j,k) : \ t_{1,m}^- \le j \le t_{1,m}, \ t_{2,m} \le k \le t_{2,m}^+\}, \\ R_m^{++} &:= \{(j,k) : \ t_{1,m} \le j \le t_{1,m}^+, \ t_{2,m} \le k \le t_{2,m}^+\}, \\ R_m^{+-} &:= \{(j,k) : \ t_{1,m} \le j \le t_{1,m}^+, \ t_{2,m}^- \le k \le t_{2,m}\}. \end{aligned}$$

Let $z_1, \ z_2 \in \{+,-\}$. For $m = 1, \ldots, s$ and $(j,k) \in R_m^{z_1 z_2}$ the weights will be

$$v_{j,k} = 1 - \frac{1}{2}\left(1 - \sqrt{\frac{|j - t_{1,m}|}{d_{1,m}^{z_1}}}\right)\left(1 - \frac{|k - t_{2,m}|}{d_{2,m}^{z_2}}\right) - \frac{1}{2}\left(1 - \frac{|j - t_{1,m}|}{d_{1,m}^{z_1}}\right)\left(1 - \sqrt{\frac{|k - t_{2,m}|}{d_{2,m}^{z_2}}}\right), \tag{4}$$

where $(d_{1,m}^-, d_{1,m}^+, d_{2,m}^-, d_{2,m}^+)$ are given in Subsection 7.1.

For $q_S \in \{-1, 0, 1\}^{(n_1-1)\times(n_2-1)}$, take

$$w_{j,k}(q_S) = \begin{cases} +1 - v_{j,k}, & q_{t_m} = +1 \\ -1 + v_{j,k}, & q_{t_m} = -1 \end{cases}, \ (j,k) \in R_m, \ m \in [s]. \tag{5}$$

Then $w(q_S)$ is an interpolating matrix for q_S. Moreover, it has the property that $w_{j,k}(z) = 0$ as soon as (j,k) is at the boundary of R_m for some $m \in [s]$.

Remark 24 (Dependence on the rectangular tessellation) *Given S, the weights in* (4) *and the interpolating matrix in* (5) *both depend on the rectangular tessellation $\{R_m\}_{m\in[s]}$ chosen. Thus, also the bound on the effective sparsity derived in the next lemma depends on $\{R_m\}_{m\in[s]}$ and can be interpreted to hold, given a set S, for an arbitrary rectangular tessellation $\{R_m\}_{m\in[s]}$.*

Lemma 25 (Bound on the (worst-case) effective sparsity) *With the weights v given in (4) we have*

$$\Gamma^2(S, v_{-S}) \le \frac{1}{2}\Big(\log(en_1) + \log(en_2)\Big) \sum_{m=1}^{s} \left(\frac{n}{d_m^{--}} + \frac{n}{d_m^{-+}} + \frac{n}{d_m^{++}} + \frac{n}{d_m^{+-}}\right).$$

Proof We say that the interpolating matrix w has product structure if it is of the form $w(j,k) = w_1(j)w_2(k)$ for all $(j,k) \in [2 : n_1] \times [2 : n_2]$. Clearly, if it has this structure, then

$$(\Delta w)_{j,k} = (D_1 w_1)_j (D_2 w_2)_k.$$

We examine now a prototype rectangle $[-d_1^- : d_1^+] \times [-d_2^- : d_2^+]$. Consider the four rectangles

$$\begin{aligned} R^{--} &:= [-d_1^- : 0] \times [-d_2^- : 0], & R^{-+} &:= [-d_1^- : 0] \times [0 : d_2^+], \\ R^{+-} &:= [0 : d_1^+] \times [-d_2^- : 0], & R^{++} &:= [0 : d_1^+] \times [0 : d_2^+], \end{aligned}$$

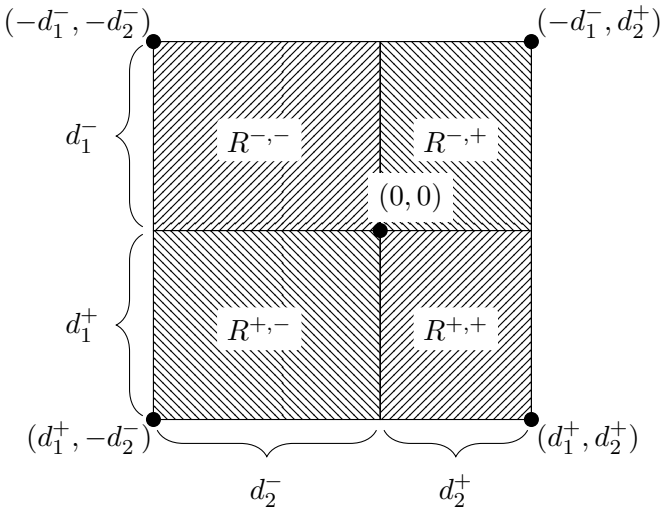

Figure 5: Illustration of the prototype rectangle R used in the Proof of Lemma 25.

and let $R := R^{--} \cup R^{-+} \cup R^{+-} \cup R^{++} = [-d_1^- : d_1^+] \times [-d_2^- : d_2^+]$. Thus R is a rectangle surrounding the origin $(0,0)$. The prototype rectangle R is illustrated in Figure 5.

For $z_1,\ z_2 \in \{+,-\}$, and $(j,k) \in R^{z_1 z_2}$, take

$$w_{j,k} := \frac{1}{2}\left(1 - \sqrt{\frac{|j|}{d_1^{z_1}}}\right)\left(1 - \frac{|k|}{d_2^{z_2}}\right) + \frac{1}{2}\left(1 - \frac{|j|}{d_1^{z_1}}\right)\left(1 - \sqrt{\frac{|k|}{d_2^{z_2}}}\right).$$

Then $w_{0,0} = 1$ and $w_{j,k} = 0$ for all (j,k) at the border of R.

Because R^{--}, R^{-+}, R^{++} and R^{+-} are rectangles aligned with the coordinate axes, w is the sum of two terms with product structure. We see that

$$|\Delta w_{j,k}| \leq \frac{1}{2}\frac{1}{\sqrt{d_1^{z_1}|j|}}\frac{1}{d_2^{z_2}} + \frac{1}{2}\frac{1}{d_1^{z_1}}\frac{1}{\sqrt{d_2^{z_2}|k|}}, (j,k) \in R^{z_1 z_2}.$$

Invoking the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for real numbers a and b, we conclude that

$$\begin{aligned}
\sum_{(j,k)\in R}\left(\Delta w_{j,k}\right)^2 &\leq \sum_{(z_1,z_2)\in\{+,-\}^2}\left(\frac{1}{2}\frac{1}{d_1^{z_1}(d_2^{z_2})^2}\sum_{j=1}^{d_1^{z_1}}\sum_{k=1}^{d_2^{z_2}}\frac{1}{j} + \frac{1}{2}\frac{1}{(d_1^{z_1})^2 d_2^{z_2}}\sum_{j=1}^{d_1^{z_1}}\sum_{k=1}^{d_2^{z_2}}\frac{1}{k}\right) \\
&\leq \frac{1}{2}\frac{1}{d_1^- d_2^-}\left(\log(ed_1^-) + \log(ed_2^+)\right) + \frac{1}{2}\frac{1}{d_1^- d_2^+}\left(\log(ed_1^-) + \log(ed_2^-)\right) \\
&\quad + \frac{1}{2}\frac{1}{d_1^+ d_2^+}\left(\log(ed_1^+) + \log(ed_2^+)\right) + \frac{1}{2}\frac{1}{d_1^+ d_2^-}\left(\log(ed_1^+) + \log(ed_2^-)\right).
\end{aligned}$$

The interpolating matrices $w(q_S)$ given by (5) are of the above form on each R_m. Moreover, they are equal to zero on their borders. The final result follows from glueing the $\{R_m\}_{m=1}^s$ together. ■

7.4 Dealing with the Noise

We start with an auxiliary lemma, which we will use to find a convenient formula for the interpolating matrix w based on the noise weights v. Both w and v are informally added to the statement of the lemma under the terms to which they correspond in the application of the lemma.

Lemma 26 (Auxiliary lemma) *For all* $(x, y) \in [0, 1]^2$

$$\underbrace{\frac{(1-\sqrt{x})(1-y)+(1-x)(1-\sqrt{y})}{2}}_{\text{“}w\text{”}} \leq 1 - \underbrace{\frac{\sqrt{x}+\sqrt{y}}{2}}_{\text{“}v\text{”}}$$

Proof See Appendix C.3. ■

The next lemma will be used to obtain the inverse scaling factor $\tilde{\gamma}$ and the noise weights v from the bound on the antiprojections $\tilde{v}$. Here too $\tilde{v}$, v and $\tilde{\gamma}$ are added below the terms to which they correspond in the application of the lemma.

Lemma 27 (Finding noise weights) *For any* $((t_1, t_2), (d_1, d_2)) \in \mathbb{N}^4$, *for* $j \in [t_1 : t_1+d_1]$ *and* $k \in [t_2 : t_2 + d_2]$,

$$\underbrace{\sqrt{\frac{j-t_1}{n_1}+\frac{k-t_2}{n_2}}}_{\text{“}\tilde{v}\text{”}} \leq \underbrace{\left(\sqrt{\frac{j-t_1}{d_1}}+\sqrt{\frac{k-t_2}{d_2}}\right)}_{\text{“}2v\text{”}} \underbrace{\sqrt{\frac{d_1}{n_1}+\frac{d_2}{n_2}}}_{\text{“}\tilde{\gamma}/2\text{”}}.$$

Proof See Appendix C.4. ■

By Lemma 13, we can bound the antiprojections using the distance of the inactive variables $\{\tilde{\psi}^{j,k}\}_{(j,k)\notin S}$ from the linear space spanned by the active ones $(\{\tilde{\psi}^{j,k}\}_{(j,k)\in S})$. As a consequence of the next lemma, we may also look at the original variables ${\psi^{j,k}}_{(j,k)\in[n_1]\times[n_2]}$ instead.

Lemma 28 (Projections) *Consider the linear spaces* $\mathcal{U} = \operatorname{span}(\{u_j\})$ *and* $\mathcal{W}$. *Define the linear space* $\tilde{\mathcal{U}} := \operatorname{span}(\{\tilde{u}_j\})$, *where* $\tilde{u}_j = \mathrm{P}_{\mathcal{W}} u_j$ *for all* j. *For any* z *define* $\tilde{z} = \mathrm{P}_{\mathcal{W}} z$. *Then* $\|\tilde{z} - \mathrm{P}_{\tilde{\mathcal{U}}}\tilde{z}\|_2 \leq \|z - \mathrm{P}_{\mathcal{U}} z\|_2$.

Proof See Appendix C.5 ■

In the next lemma we bound the antiprojections using the distance of the original inactive variables $\{\psi^{j,k}\}_{(j,k)\notin S}$ from the linear space spanned by the active ones $(\{\psi^{j,k}\}_{(j,k)\in S})$.

Let $\mathcal{U} := \operatorname{span}\left(\{\psi^{t_m}\}_{m=1}^{s}\right)$.

Lemma 29 (Finding a bound on the antiprojections) *For* $m \in [s]$ *and all* $(j, k) \in R_m$

$$\|\mathrm{A}_{\mathcal{U}}\psi^{j,k}\|_2^2/n \leq \frac{|j-t_{1,m}|}{n_1} + \frac{|k-t_{2,m}|}{n_2}.$$

Proof See Appendix C.6. ■

7.5 Proof of Theorem 19

Theorem 19 follows from the results of Subsections 7.2–7.4 combined with Theorem 16.

We thus need to find suitable $\tilde{v},\ \tilde{\gamma},\ v,\ w$ and an upper bound on the effective sparsity. Let $S \subseteq [3 : n_1 - 1] \times [3 : n_2 - 1]$ be arbitrary. Let

$$\mathcal{U} := \text{span}\left(\{\psi^{t_m}\}_{m\in[s]}\right),\ \tilde{\mathcal{U}} := \text{span}\left(\{\tilde{\psi}^{t_m}\}_{m\in[s]}\right)$$

and

$$\mathcal{W} := \text{span}\left(\{\psi^{j,k}\}_{(j,k)\in\{1\}\times[n_2]\cup[n_1]\times\{1\}}\right).$$

Note that $\|\mathrm{A}_{\tilde{\mathcal{U}}}\tilde{\psi}^{j,k}\|_2/\sqrt{n}$ is the same quantity as $\|(\mathrm{I}_n - \mathrm{P}_S)\tilde{\Psi}_i\|_2^2/\sqrt{n}$ found in Definition 10.

- By Lemma 28 we have that $\|\mathrm{A}_{\tilde{\mathcal{U}}}\tilde{\psi}^{j,k}\|_2/\sqrt{n} = \|\mathrm{A}_{\tilde{\mathcal{U}}}\tilde{\psi}^{j,k}\|_2/\sqrt{n} \leq \|\mathrm{A}_{\mathcal{U}}\psi^{j,k}\|_2/\sqrt{n}$, since $\tilde{\psi}^{j,k} = \mathrm{P}_{\mathcal{W}}\psi^{j,k}$, $(j,k) \in [2 : n_1] \times [2 : n_2]$.
- Upper bounds on the values of $\|\mathrm{A}_{\mathcal{U}}\psi^{j,k}\|_2/\sqrt{n}$ are given by Lemma 29. These upper bounds yield a bound on the antiprojections $\tilde{v}$ and the corresponding inverse scaling factor $\tilde{\gamma}$.
- By applying Lemma 27 and Lemma 26 to the results of Lemma 29 we see that for v as in (4) we have $v_{j,k} \geq \tilde{v}_{j,k}/\tilde{\gamma},\ \forall (j,k) \in [2 : n_1] \times [2 : n_2]$, where

$$\tilde{\gamma} = 2\sqrt{\frac{d_{1,\max}(S)}{n_1} + \frac{d_{2,\max}(S)}{n_2}}$$

  and v reaches its maximal values of 1 on the boundaries of the rectangles $\{R_m\}_{m=1}^s$.
- Therefore $w(q_S)$ given in (5) is an interpolating matrix for the sign configuration q_S.

Lemma 25 bounds the effective sparsity by using the interpolating matrix $w(q_S)$ given in (5). ■

8. Slow Rates for Total Variation Image Denoising

The goal of this section is to prove a slow rate for $\hat{\tilde{f}}$. It will turn out that this rate is $n^{-5/8}$, up to log terms, and is faster than the rate $n^{-3/5}$ in Mammen and van de Geer (1997).

We combine the standard Theorem 17 with the insight that, if the arbitrary active set S is chosen in a careful way, we can obtain a value of $\tilde{\gamma}$ which is smaller than the one obtained in Section 7. The key idea is that an active set defining a "mesh grid" (see Definition 30 and Figure 6) results in a more favorable $\tilde{\gamma}$ than an active set S defining a regular grid.

We consider the two-dimensional grid $[2 : n_1] \times [2 : n_2]$. Let $t_1, t_2 \in \mathbb{N}$. Assume that $n_1/(t_1^2 + 1)$ and $n_2/(t_2^2 + 1)$ are integers.

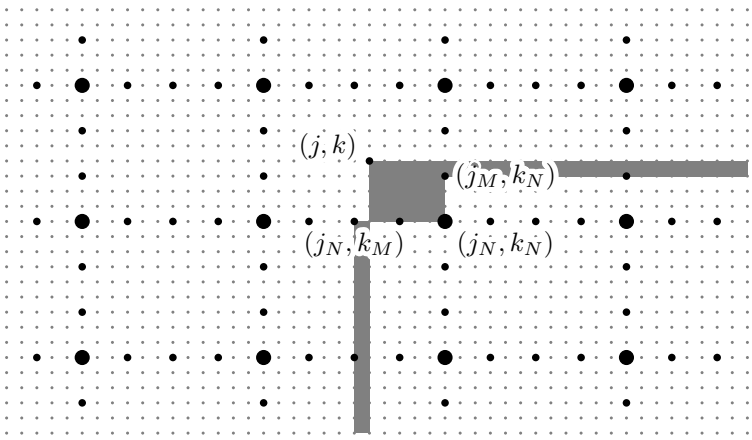

Figure 6: Illustration of the two-dimensional grid $[2:30]\times[2:51]$. The black points represent a mesh grid S_M with $t_1=3$ and $t_2=4$. The thicker points belong to the set of nodes S_N. The grey area corresponds to the error made when approximating $\psi^{j,k}$ by a linear combination of $\psi^{j_N,k_N},\psi^{j_N,k_M},\psi^{j_M,k_N}\in\mathcal{M}$ in the Proof of Lemma 31.

We define

$$\begin{aligned}
M_1 &:= \left\{1+\frac{n_1}{t_1^2+1},1+\frac{2n_1}{t_1^2+1},\dots,1+\frac{t_1^2n_1}{t_1^2+1}\right\}\\
N_1 &:= \left\{1+\frac{\lceil t_1/2\rceil n_1}{t_1^2+1},1+\frac{(\lceil t_1/2\rceil+t_1)n_1}{t_1^2+1},\dots,1+\frac{(\lceil t_1/2\rceil+t_1(t_1-1))n_1}{t_1^2+1}\right\}\\
M_2 &:= \left\{1+\frac{n_2}{t_2^2+1},1+\frac{2n_2}{t_2^2+1},\dots,1+\frac{t_2^2n_2}{t_2^2+1}\right\}\\
N_2 &:= \left\{1+\frac{\lceil t_2/2\rceil n_2}{t_2^2+1},1+\frac{(\lceil t_2/2\rceil+t_2)n_2}{t_2^2+1},\dots,1+\frac{(\lceil t_2/2\rceil+t_2(t_2-1))n_2}{t_2^2+1}\right\}
\end{aligned}$$

Definition 30 (Mesh grid) *An active set $S_M\subseteq[2:n_1]\times[2:n_2]$ is said to define a mesh grid if $S_M:=M_1\times N_2\cup N_1\times M_2$.*

We define the set of the nodes S_N of the mesh grid S_M as $S_N:=N_1\times N_2$.

An example of a mesh grid is illustrated in Figure 6. If S_M is a mesh grid, then $s_M=t_1^2t_2+t_1t_2^2-t_1t_2=t_1t_2(t_1+t_2-1)$. Note that we can add t_1t_2 points to the mesh grid S_M to obtain an active set of cardinality $t_1t_2(t_1+t_2)$. The antiprojections do not become larger. With some abuse of notation we write from now on $s_M=t_1t_2(t_1+t_2)$.

The next lemma is an analogon of Lemma 29, but the active set is constrained to take the form of a mesh grid. This allows us to find a more favorable uniform bound for the antiprojections, which is smaller than the one we would have if the active set were a regular grid like S_N instead of a mesh grid.

Let S_M be a mesh grid and let $\mathcal{M}:=\operatorname{span}\left(\{\psi^{j,k}\}_{(j,k)\in S_M}\right)$.

Lemma 31 (Finding a bound on the antiprojections when S is a mesh grid) *For all $(j,k) \in [2:n_1] \times [2:n_2]$ it holds that*

$$\|\mathrm{A}_{\mathcal{M}} \psi^{j,k}\|_2^2/n \leq \frac{1}{t_1^2} + \frac{1}{t_2^2} + \frac{\lceil t_1/2 \rceil \lceil t_2/2 \rceil}{t_1^2 t_2^2}.$$

Proof Let $(j,k) \in [2:n_1] \times [2:n_2]$ be arbitrary. We define

$$j_M := \arg\min_{i \in M_1} |i-j| \qquad j_N := \arg\min_{i \in N_1} |i-j|$$
$$k_M := \arg\min_{i \in M_2} |i-k| \qquad k_N := \arg\min_{i \in N_2} |i-k|.$$

We now approximate $\psi^{j,k}$ by a linear combination of $\psi^{j_N,k_N}, \psi^{j_N,k_M}, \psi^{j_M,k_N} \in \mathcal{M}$ (cf. Figure 6) as

$$\begin{aligned} \|\psi^{j,k} - \psi^{j_M,k_N} - \psi^{j_N,k_M} + \psi^{j_N,k_N}\|_2^2 &\leq j|k-k_M| + k|j-j_M| + |j_N - j_M||k_N - k_M| \\ &\leq \frac{n_1 n_2}{t_2^2+1} + \frac{n_1 n_2}{t_1^2+1} + \frac{n_1 n_2 \lceil t_1/2 \rceil \lceil t_2/2 \rceil}{(t_1^2+1)(t_2^2+1)}. \end{aligned}$$

■

Let us now define the class

$$\mathcal{F}(C) = \left\{ \tilde{f} \in \mathbb{R}^{n_1 \times n_2} : \mathrm{TV}(\tilde{f}) \leq C \right\}, \ C > 0.$$

For the ease of exposition, we consider square images ($n_1 = n_2$) and we choose $t_1 = t_2 =: \tau$ to be even. The following theorem holds.

Theorem 32 (Slow rates for total variation image denoising) *The estimator $\hat{\tilde{f}}$ has the following properties.*

- **Dependence of S_M and λ on f^0 allowed.**
  Choose S_M such that

  $$s_M = \left\lceil \frac{2^{5/4} 3^{3/4} \log^{3/8}(2n) \mathrm{TV}^{3/4}(f^0)}{\sigma^{3/4}} \right\rceil \text{ and } \lambda = \frac{3^{3/4} \sigma^{5/4} \log^{3/8}(2n)}{2^{1/12} n^{5/8} \mathrm{TV}^{1/4}(f^0)} \geq \tilde{\gamma} \lambda_0(\log(2n)).$$

  Then, with probability at least $1 - 1/n$ it holds that

  $$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n \leq \frac{24 \sigma^{5/4} C^{3/4} \log^{3/8}(2n)}{n^{5/8}} + \frac{4\sigma^2 \log(2n)}{n} + \frac{2\sigma^2}{n}.$$

- **Dependence of S_M and λ on f^0 not allowed.**
  Choose S_M such that

  $$s_M = \left\lceil 2^{5/4} 3^{3/4} \log^{3/8}(2n) \right\rceil \text{ and } \lambda = \frac{3^{3/4} \log^{3/8}(2n)}{2^{1/12} n^{5/8}} \geq \tilde{\gamma} \lambda_0(\log(2n)).$$

  Then, with probability at least $1 - 1/n$ it holds that

  $$\|\hat{\tilde{f}} - \tilde{f}^0\|_2^2/n \leq \frac{12 \sigma (C+\sigma) \log^{3/8}(2n)}{n^{5/8}} + \frac{4\sigma^2 \log(2n)}{n} + \frac{2\sigma^2}{n}.$$

Proof Our starting point is Theorem 17. We choose $x = t = \log(2n)$, $g = \tilde{f}^0$ and $S = S_M$ and we apply the inequality $(\sqrt{2x} + \sqrt{s_M})^2 \leq 4x + 2s_M$. Note that $s_M = 2\tau^3$. By Lemma 31, $\tilde{\gamma} = 3/(2\tau) = 2^{-2/3} 3 s_M^{-1/3}$.

The two results of the above theorem follow by optimally choosing s_M to trade off $2s_M/n$ and $4\lambda \mathrm{TV}(f^0)$ or $4\lambda/\sigma$, respectively. ■

Remark 33 (Comments on Theorem 32)

- *Theorem 32 improves on the rate $n^{-3/5}$ found in Mammen and van de Geer (1997). If in the proof of Theorem 32 one chooses an active set defining a regular grid and bounds the antiprojections with Lemma 29, then the rate $n^{-3/5}$ by Mammen and van de Geer (1997) is retrieved, up to a $\log^{2/5} n$ term.*

- *Since we assume t to be an even integer, s_M should be the smallest number, which is twice an even cube and greater than the s_M we propose in the statement of Theorem 32.*

- *In the second part of Theorem 32 the choice of λ is completely data-driven, does not depend on $\tilde{f}^0$ and is smaller than the universal choice $\lambda_0(\log(2n))$.*

- *Under the assumption that $C > 0$ is a constant that does not depend on n, the two rates in Theorem 32 are equal.*

- *The results of Theorem 32 are "constant-friendly" in the sense that we can trace the constants for the choice of the tuning parameter and for the upper bound on the mean squared error. Moreover, these constants are small. This situation has to be contrasted with results relying on entropy calculations, which possibly produce very large constants both in the tuning parameter and in the upper bound, for instance Mammen and van de Geer (1997).*

Remark 34 (Comparision with Fang et al. (2019)) *Fang et al. (2019) study the estimator*

$$\hat{f}_C := \arg \min_{f \in K(C)} \|Y - f\|_2^2 / n,$$

where, for $C > 0$, $K(C) := \{f \in \mathbb{R}^{n_1 \times n_2} : \sum_{j=2}^{n_1} |f_{j,1} - f_{j-1,1}| + \sum_{k=2}^{n_2} |f_{1,k} - f_{1,k-1}| + \mathrm{TV}(f) \leq C\}$. Fang et al. (2019) show that the minimax rate of estimation for $f^0 \in K(C)$ is of order $n^{-2/3} C^{2/3}$ up to log terms and $\hat{f}_C$ attains it.

With the entropy bound by Blei et al. (2007) used by Fang et al. (2019), one could prove a similar result for the penalized version of the estimator $\hat{f}_C$, which differs from our estimator $\hat{f}$, since the one-dimensional total variation is defined in a different way and the tuning parameters are choosen to be $\lambda_1 = \lambda_2 = \lambda$.

The minimax rate in the class $\bar{K}(C) := \{f \in \mathbb{R}^{n_1 \times n_2} : \mathrm{TV}_1(f) \leq C, \mathrm{TV}_2(f) \leq C, \mathrm{TV}(f) \leq C\}$ follows, by the ANOVA decomposition, from the minimax rate for the class $\mathcal{F}(C) = \{\tilde{f} \in \mathbb{R}^{n_1 \times n_2} : \mathrm{TV}(\tilde{f}) \leq C\}$. The class $\bar{K}(C)$ is larger than the class $K(C)$, as for $f \in K(C)$ it holds that $\mathrm{TV}_1(f) \leq C$ and $\mathrm{TV}_2(f) \leq C$.

9. Conclusion

We showed that the estimator for the interaction terms satisfies an oracle inequality with fast rates and can adapt to the number and the locations of the unknown "jumps" to optimally trade off approximation and estimation error as if it would know the true image f^0.

As in our previous work on one-dimensional (higher-order) total variation regularization, the projection arguments by Dalalyan et al. (2017) are central to make sure that the effective sparsity is "small enough" and thus being able to prove adaptivity. These arguments exploit the strong correlation between the atoms constituting the dictionary. It also turns out that the technique to bound the effective sparsity proposed by Ortelli and van de Geer (2019b) and inspired by Candès and Fernandez-Granda (2014) can be generalized to the two-dimensional case.

The slow rate $n^{-5/8} \log^{3/8} n$ improves on the rate $n^{-3/5}$ by Mammen and van de Geer (1997).

Both for fast and slow rates, the estimator enjoys the most favorable theoretical properties when the tuning parameter is chosen to be smaller than the universal choice of order $\sqrt{\log n/n}$. The most favorable choice of the tuning parameter might depend on some aspects of the true image f^0. However we show that there are choices of λ which are completely data driven and still confer to the estimator adaptivity (fast rates) or the rate $n^{-5/8} \log^{3/8} n$ (slow rates).

Acknowledgments

We would like to acknowledge support for this project from the the Swiss National Science Foundation (SNF grant 200020_169011). We moreover thank the action editor and the referees for the careful reading of the manuscript and for their valuable comments.

Appendix A. Proofs of Section 5

A.1 Proof of Lemma 7

We note that, for $(j,k) \in [n_1] \times [n_2]$, the global mean of a dictionary atom is given by $\psi^{j,k}(\circ,\circ) = \psi^{1,k}(\circ,\circ)\psi^{j,1}(\circ,\circ)$ and its main effects are given by $\psi^{j,k}(\cdot,\circ) = \psi^{1,k}(\circ,\circ)\psi^{j,1} - \psi^{j,k}(\circ,\circ)\psi^{1,1}$ and $\psi^{j,k}(\circ,\cdot) = \psi^{j,1}(\circ,\circ)\psi^{1,k} - \psi^{j,k}(\circ,\circ)\psi^{1,1}$. Thus

$$\psi^{j,k} = \tilde{\psi}^{j,k} + \psi^{1,k}(\circ,\circ)\tilde{\psi}^{j,1} + \psi^{j,1}(\circ,\circ)\tilde{\psi}^{1,k} + \psi^{j,k}(\circ,\circ)\psi^{1,1}.$$

From the definition of $\tilde{\psi}^{j,k},\ (j,k) \in [n_1] \times [n_2]$ it follows that

$$f = \tilde{\beta}_{1,1}\tilde{\psi}^{1,1} + \sum_{j=2}^{n_1} \tilde{\beta}_{j,1}\tilde{\psi}^{j,1} + \sum_{k=2}^{n_2} \tilde{\beta}_{1,k}\tilde{\psi}^{1,k} + \sum_{j=2}^{n_1}\sum_{k=2}^{n_2} \tilde{\beta}_{j,k}\tilde{\psi}^{j,k},$$

where

$$\tilde{\beta}_{j,k} = \begin{cases} \beta_{1,1} + \sum_{j=2}^{n_1} \beta_{j,1}\psi^{j,1}(\circ,\circ) + \sum_{k=2}^{n_2} \beta_{1,k}\psi^{1,k}(\circ,\circ) + \sum_{j=2}^{n_1}\sum_{k=2}^{n_2} \beta_{j,k}\psi^{j,k}(\circ,\circ),\ (j,k) = (1,1) \\ \beta_{j,1} + \sum_{k=2}^{n_2} \beta_{j,k}\psi^{1,k}(\circ,\circ),\ (j,k) \in [2:n_1] \times [1], \\ \beta_{1,k} + \sum_{j=2}^{n_1} \beta_{j,k}\psi^{j,1}(\circ,\circ),\ (j,k) \in [1] \times [2:n_2], \\ \beta_{j,k},\ (j,k) \in [2:n_1] \times [2:n_2]. \end{cases}$$

Note that $\psi^{1,k}(\circ,\circ) = 1 - (k-1)/n_2$, $\psi^{j,1}(\circ,\circ) = 1 - (j-1)/n_1$ and

$$\sum_{j=2}^{n_1} \beta_{j,1}\psi^{j,1}(\circ,\circ) = -f(1,1) + \frac{1}{n_1}\sum_{j=1}^{n_1} f(j,1).$$

Analogously, it holds that

$$\sum_{k=2}^{n_2} \beta_{j,k}\psi^{1,k}(\circ,\circ) = -f(1,1) + \frac{1}{n_2}\sum_{k=1}^{n_2} f(1,k).$$

By plugging in the expressions for $\beta_{j,k},\ (j,k) \in [n_1] \times [n_2]$ and by using the above equations the result follows. ■

A.2 Proof of Lemma 8

Since $f - \tilde{f}$ and $\tilde{f}$ are orthogonal, $\mathrm{trace}(\tilde{f}^T(f - \tilde{f})) = 0$. We have

$$\begin{aligned} \|Y - f\|_2^2 - \|Y\|_2^2 &= -2\mathrm{trace}(Y^T f) + \|f\|_2^2 \\ &= -2\mathrm{trace}(Y^T(f - \tilde{f})) - 2\mathrm{trace}(Y^T\tilde{f}) + \|f - \tilde{f}\|_2^2 + \|\tilde{f}\|_2^2 \\ &= \|Y - (f - \tilde{f})\|_2^2 + \|Y - \tilde{f}\|_2^2 - 2\|Y\|_2^2. \end{aligned}$$

The result now follows from Lemma 7. ■

Appendix B. Proofs of Section 6

For the proof of Lemma 9 we use the dual norm inequality and the polarization identity.

Let $y, z \in \mathbb{R}^n$. The dual norm of $\|z\|_1$ is $\sup_{\|y\|_1 \leq 1} |y^T z| = \max_{i \in [n]} |z_i| = \|z\|_\infty$. In particular, the dual norm inequality holds

$$|y^T z| \leq \|y\|_1 \|z\|_\infty, \ \forall y, z \in \mathbb{R}^n.$$

For $y, z \in \mathbb{R}^n$ the polarization identity holds

$$2y^T z = \|y\|_2^2 + \|z\|_2^2 - \|y - z\|_2^2.$$

B.1 Proof of Lemma 9

The estimator $\hat{\tilde{f}}$ satisfies the optimality conditions (KKT conditions)

$$\frac{\tilde{Y} - \hat{\tilde{f}}}{n} = \lambda \Delta^T \partial \|\Delta \hat{\tilde{f}}\|_1,$$

where $\Delta' \partial \|\Delta \hat{\tilde{f}}\|_1$ is a subgradient of $\|\Delta \hat{\tilde{f}}\|_1$ by the chain rule of the subgradient (Theorem 23.9 in Rockafellar (1970)). In other words $\Delta' \partial \|\Delta \hat{\tilde{f}}\|_1$ is any vector s.t. $\hat{\tilde{f}}' \Delta' \partial \|\Delta \hat{\tilde{f}}\|_1 = \|\Delta \hat{\tilde{f}}\|_1$. Since Δ is fixed, $\partial \|\Delta \hat{\tilde{f}}\|_1$ can be any vector in $\mathbb{R}^{(n_1-1)(n_2-1)}$ s.t.

$$(\partial \|\Delta \hat{\tilde{f}}\|_1)_i \in \begin{cases} \text{sign}((\Delta \hat{\tilde{f}})_i), & (\Delta \hat{\tilde{f}})_i \neq 0, \\ [-1, 1], & (\Delta \hat{\tilde{f}})_i = 0. \end{cases}$$

The set of subgradients of $\|\Delta \hat{\tilde{f}}\|_1$ is called subdifferential of $\|\Delta \hat{\tilde{f}}\|_1$. By multiplying the KKT conditions by $\hat{\tilde{f}}$ and an arbitrary $g \in \mathbb{R}^n$, we therefore obtain

$$\frac{\hat{\tilde{f}}^T (\tilde{Y} - \hat{\tilde{f}})}{n} = \lambda \hat{\tilde{f}}^T \Delta^T \partial \|\Delta \hat{\tilde{f}}\|_1 = \lambda \|\Delta \hat{\tilde{f}}\|_1 \tag{6}$$

and

$$\frac{g^T (\tilde{Y} - \hat{\tilde{f}})}{n} = \lambda g^T \Delta^T \partial \|\Delta \hat{\tilde{f}}\|_1 \leq \lambda \|\Delta g\|_1. \tag{7}$$

The last inequality follows by the dual norm inequality and the fact that $\|\partial \|\Delta \hat{\tilde{f}}\|_1\|_\infty \leq 1$. Subtracting Equation (6) from Equation (7), together with the identity $\tilde{Y} = \tilde{f}^0 + \tilde{\epsilon}$ yields

$$\frac{(g - \hat{\tilde{f}})^T (\tilde{Y} - \hat{\tilde{f}})}{n} = \frac{(g - \hat{\tilde{f}})^T (\tilde{f}^0 - \hat{\tilde{f}})}{n} - \frac{\tilde{\epsilon}^T (\hat{\tilde{f}} - g)}{n} \leq \lambda (\|\Delta g\|_1 - \|\Delta \hat{\tilde{f}}\|_1).$$

The application of the polarization identity gives

$$2 \frac{(g - \hat{\tilde{f}})^T (\tilde{f}^0 - \hat{\tilde{f}})}{n} = \|\hat{\tilde{f}} - \tilde{f}^0\|_2^2 / n + \|\hat{\tilde{f}} - g\|_2^2 / n - \|g - \tilde{f}^0\|_2^2 / n$$

and Lemma 9 follows.

B.2 Proof of Lemma 13

We start by decomposing the empirical process as in Equation (2). For the first part, we have that

$$\tilde{\epsilon}^T \mathrm{P}_S \tilde{f}/n \leq \|\mathrm{P}_S \tilde{\epsilon}\|_2 \|\tilde{f}\|_2/n \leq \|\mathrm{P}_S \tilde{\epsilon}\|_2 \|f\|_2/n$$

and by applying Lemma 1 by Laurent and Massart (2000) or Lemma 8.6 by van de Geer (2016) for $x > 0$ it holds that with probability at least $1 - e^{-x}$

$$\tilde{\epsilon}^T \mathrm{P}_S \tilde{f}/n \leq \|f\|_2 \sigma(\sqrt{2x} + \sqrt{s})/n.$$

For the second part, note that by Lemma 7 we can write

$$\tilde{\epsilon}^T \mathrm{A}_S \tilde{f}/n = \tilde{\epsilon}^T \mathrm{A}_S \tilde{\Psi} \tilde{\beta}/n,$$

where $\tilde{\beta} = \Delta f$.

Then by Lemma 17.5 in van de Geer (2016), for $t > 0$ we have that with probability at least $1 - e^{-t}$

$$\frac{\tilde{\epsilon}^T \mathrm{A}_S \tilde{f}}{n} \leq \lambda \|\tilde{v}_{-S} \odot (\Delta \tilde{f})_{-S}\|_1 / \tilde{\gamma} = \lambda \|v_{-S} \odot (\Delta f)_{-S}\|_1,$$

if we choose $\lambda \geq \tilde{\gamma} \lambda_0(t)$. Thus the claim follows. ■

B.3 Proof of Theorem 16

For $S \subseteq [(n_1 - 1)(n_2 - 1)]$ and $q_S = \mathrm{sign}((\Delta g)_S)$, we have that, by the triangle inequality,

$$\|\Delta g\|_1 - \|\Delta \hat{\tilde{f}}\|_1 \leq q_S^T (\Delta(g - \hat{\tilde{f}}))_S - \|(\Delta(g - \hat{\tilde{f}}))_{-S}\|_1 + 2\|(\Delta g)_{-S}\|_1.$$

By combining the above inequality with Lemma 9 and Lemma 13 applied to $g - \hat{\tilde{f}}$, using the definition of effective sparsity, and applying the convex conjugate inequality to the term involving $\|g - \hat{\tilde{f}}\|_2/\sqrt{n}$ we obtain the claim. ■

B.4 Proof of Theorem 17

Note that Lemma 13 implies that, for $x, t > 0$ and $\lambda \geq \tilde{\gamma} \lambda_0(t)$, with probability at least $1 - e^{-x} - e^{-t}$ it holds that

$$2\frac{\tilde{\epsilon}^T (g - \hat{\tilde{f}})}{n} \leq \frac{\|g - \hat{\tilde{f}}\|_2^2}{n} + \left(\sqrt{\frac{2x}{n}} + \sqrt{\frac{s}{n}}\right)^2 + 2\lambda(\|\Delta g\|_1 + \|\Delta \hat{\tilde{f}}\|_1).$$

Combining the above claim with the basic inequality (Lemma 9) proves the theorem. ■

Appendix C. Proofs of Section 7

C.1 Proof of Lemma 23

Partial integration in one dimension gives, for $N \in \mathbb{N}$ and sequences $\{a_j\}_{j=2}^N : a_2 = a_N = 0$ and $\{b_j\}_{j=1}^N$ of real numbers,

$$\sum_{j=2}^{N} a_j (b_j - b_{j-1}) = a_N b_N - a_2 b_1 - \sum_{j=2}^{N-1} (a_{j+1} - a_j) b_j = -\sum_{j=2}^{N-1} (a_{j+1} - a_j) b_j.$$

Apply this to the inner and the outer sum. ■

C.2 Proof of Lemma 22

Let $f \in \mathbb{R}^{n_1 \times n_2}$ be arbitrary and let q_S be a sign configuration. Then

$$\begin{aligned} & \operatorname{trace}(q_S^T \odot (D_1 f D_2^T)_S) - \|(1-v)_{-S} \odot (D_1 f D_2^T)_{-S}\|_1 \\ \leq\ & \operatorname{trace}(q_S^T \odot (D_1 f D_2^T)_S) - \|w_{-S}(q_S) \odot (D_1 f D_2^T)_{-S}\|_1 \\ =\ & \operatorname{trace}(w(q_S)^T D_1 f D_2^T) = \operatorname{trace}(D_2^T w(q_S)^T D_1 f) \\ \leq\ & \sqrt{n}\|D_1^T w(q_S) D_2\|_2 \|f\|_2/\sqrt{n}. \end{aligned}$$

■

C.3 Proof of Lemma 26

By direct calculation

$$\begin{aligned} 1 - \frac{1}{2}(1-\sqrt{x})(1-y) - \frac{1}{2}(1-x)(1-\sqrt{y}) &= \frac{1}{2}\sqrt{x} + \frac{1}{2}(1-\sqrt{x})y + \frac{1}{2}\sqrt{y} + \frac{1}{2}x(1-\sqrt{y}) \\ &\geq (\sqrt{x}+\sqrt{y})/2 \end{aligned}$$

where we used that $(x, y) \in [0,1]^2$ so that $(1-\sqrt{x})y$ and $x(1-\sqrt{y})$ are non-negative. ■

C.4 Proof of Lemma 27

For $j \in \{t_1, \ldots, t_1 + d_1\}$ and $k \in \{t_2, \ldots, t_2 + d_2\}$

$$\begin{aligned} \sqrt{\frac{j-t_1}{n_1} + \frac{k-t_2}{n_2}} \leq \sqrt{\frac{j-t_1}{n_1}} + \sqrt{\frac{k-t_2}{n_2}} &= \sqrt{\frac{j-t_1}{d_1}}\sqrt{\frac{d_1}{n_1}} + \sqrt{\frac{k-t_2}{d_2}}\sqrt{\frac{d_2}{n_2}} \\ &\leq \left(\sqrt{\frac{j-t_1}{d_1}} + \sqrt{\frac{k-t_2}{d_2}}\right)\sqrt{\frac{d_1}{n_1} + \frac{d_2}{n_2}}. \end{aligned}$$

■

C.5 Proof of Lemma 28

Clearly $\|\mathrm{P}_{\mathcal{W}}(z - \mathrm{P}_{\mathcal{U}} z)\|_2 \leq \|z - \mathrm{P}_{\mathcal{U}} z\|_2$. We moreover have for some vector γ

$$\mathrm{P}_{\mathcal{U}} z = \sum_j \gamma_j u_j \text{ s.t. } \mathrm{P}_{\mathcal{W}}(z - \mathrm{P}_{\mathcal{U}} z) = \tilde{z} - \sum_j \gamma_j \tilde{u}_j.$$

Thus

$$\|\tilde{z} - \mathrm{P}_{\tilde{U}} \tilde{z}\|_2 = \min_c \|\tilde{z} - \sum_j c_j \tilde{u}_j\|_2 \leq \|\tilde{z} - \sum_j \gamma_j \tilde{u}_j\|_2 = \|\mathrm{P}_{\mathcal{W}}(z - \mathrm{P}_{\mathcal{U}} z)\|_2 \leq \|z - \mathrm{P}_{\mathcal{U}} z\|_2.$$

■

C.6 Proof of Lemma 29

For $(j,k) \in R_m$,

$$\|\mathrm{A}_{\mathcal{U}}\psi^{j,k}\|_2^2 \leq \|\psi^{j,k} - \psi^{t_m}\|_2^2 \leq |j - t_{1,m}|n_2 + |k - t_{2,m}|n_1.$$

■

References

R. Abergel and L. Moisan. The Shannon total variation. *Journal Math Imaging Vis*, 59: 341–370, 2017.

E. Arias-Castro, J. Salmon, and R. Willett. Oracle inequalities and minimax rates for nonlocal means and related adaptive kernel-based methods. *SIAM Journal on Imaging Sciences*, 5(3):944–992, 2012.

F. Bach. Shaping level sets with submodular functions. *Neural Information Processing Systems (NIPS)*, pages 10–18, 2011.

P. Bellec. Sharp oracle inequalities for least-squares estimators in shape restricted regression. *Annals of Statistics*, 46(2):745–780, 2018.

P. Bellec, J. Salmon, and S. Vaiter. A sharp oracle inequality for graph-slope. *Electronic Journal of Statistics*, 11(2):4851–4870, 2017.

P. Bellec, G. Lecué, and A. Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. *Annals of Statistics*, 46(6B):3603–3642, 2018.

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

R. Blei, F. Gao, and W. Li. Metric entropy of high dimensional distributions. *Proceedings of the American Mathematical Society*, 135(12):4009–4018, 2007.

E. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

E. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Math. Program. Ser. A*, 141:577–589, 2013.

V. Caselles, A. Chambolle, and M. Novaga. Total variation in imaging. *Handbook of Mathematical Methods in Imaging: Volume 1, Second Edition*, 1(1):1455–1499, 2015.

A. Chambolle and P. L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

A. Chambolle, V. Caselles, D. Cremers, and M. Novaga. An introduction to total variation for image analysis. In *Radon Series Comp. Appl. Math. 9*. De Gruyter, 2010.

A. Chambolle, S. E. Levine, and B. J. Lucier. An upwind finite-difference method for total variation-based image smoothing. *Siam J. Imaging Sciences*, 4(1):277–299, 2011.

A. Chambolle, V. Duval, G. Peyré, and C. Poon. Geometric properties of solutions to the total variation denoising problem. *Inverse Problems*, 33(1), 2017.

S. Chatterjee and S. Goswami. New risk bounds for 2d total variation denoising. *arXiv:1902.01215v2*, 2019.

L. Condat. Discrete total variation: New definition and minimization. *SIAM Journal on Imaging Sciences*, 10(3):1258–1290, 2017.

K. Dabov, A. Foi, and K. Egiazarian. Video denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on Control of Network Systems*, 16(8):2080–2095, 2007.

A. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Annals of Statistics*, 40(4):2327–2355, 2012.

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

M. Elad. *Sparse and Redundant Representations*. Springer, 2010.

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(947), 2007.

A. Elsener and S. van de Geer. Sharp oracle inequalities for stationary points of nonconvex penalized M-estimators. *IEEE Transactions on Information Theory*, 65(3):1452–1472, 2019.

J. Fadili and G. Peyre. Total variation projection with first order schemes. *IEEE Transactions on Image Processing*, 20(3):657–669, 2011.

B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. *arXiv:1903.01395v1*, 2019.

J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma. Image denoising review: from classical to state-of-the-art approaches. *Information Fusion*, 55:220–244, 2020.

A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *Annals of Statistics*, 48(1):205–229, 2020.

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

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

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

K. Lounici, M. Pontil, S. van de Geer, and A. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(4):2164–2204, 2011.

E. Mammen and A. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *Annals of Statistics*, 23(2):502–524, 1995.

E. Mammen and S. van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.

F. Ortelli and S. van de Geer. On the total variation regularized estimator over a class of tree graphs. *Electronic Journal of Statistics*, 12:4517–4570, 2018.

F. Ortelli and S. van de Geer. Synthesis and analysis in total variation regularization. *ArXiv ID 1901.06418v1*, 2019a.

F. Ortelli and S. van de Geer. Prediction bounds for (higher order) total variation regularized least squares. *ArXiv ID 1904.10871*, 2019b.

F. Ortelli and S. van de Geer. Oracle inequalities for square root analysis estimators with application to total variation penalties. *Information and Inference: A Journal of the IMA*, (iaaa002), 2020.

O. H. M. Padilla, J. Scott, J. Sharpnack, and R. Tibshirani. The DFS fused Lasso: linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:1–36, 2018.

J. Polzehl and V. Spokoiny. Image denoising: pointwise adaptive approach. *Annals of Statistics*, 31(1):30–57, 2003.

R. T. Rockafellar. *Convex analysis*, volume 196. Princeton university press, 1970.

L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

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

V. Sadhanala, Y. X. Wang, J. Sharpnack, and R. Tibshirani. Higher-order total variation classes on grids: minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, pages 5801–5811, 2017.

J. Sharpnack, A. Rinaldo, and A. Singh. Sparsistency of the edge Lasso over graphs. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 22:1028–1036, 2012.

B. Stucky and S. van de Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18:1–29, 2017.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58 (1):267–288, 1996.

R. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

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

S. van de Geer. Oracle Inequalities and Regularization. In *Lectures on Empirical Processes*, pages 191–252. European Mathematical Society, 2007.

S. van de Geer. *Empirical Processes in M-estimation*, volume 6 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge, 2009.

S. van de Geer. *Estimation and Testing under Sparsity*, volume 2159. Springer, 2016.

S. van de Geer. On tight bounds for the Lasso. *Jorunal of Machine Learning Research*, 19: 1–48, 2018.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

V. E. Waghmare. Leaf Shapes Database. URL `http://imageprocessingplace.net/downloads_V3/root_downloads/image_databases/leafshapedatabase/leaf_shapes_downloads.htm`.

Y.-X. Wang, J. Sharpnack, A. Smola, and R. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17:15–147, 2016.

K. Zhang, W. Zuo, and L. Zhang. FFDNet: toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.