Multiscale Total Variation Estimators for Regression and Inverse Problems

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
“Doctor rerum naturalium”
der Georg-August-Universität zu Göttingen

im Promotionsprogramm
“PhD School of Mathematical Sciences (SMS)”
der Georg-August University School of Science (GAUSS)

vorgelegt von
Miguel del Álamo
aus Soria, Spanien

Göttingen, 2019

Betreuungsausschuss:

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Thorsten Hohage
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Mitglieder der Prüfungskommission:

Referent:
Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Korreferent:
Prof. Dr. Thorsten Hohage
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Russell Luke
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Prof. Dr. Anja Sturm
Institut für Mathematische Stochastik, Universität Göttingen

Dr. Frank Werner
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Ingo Witt
Mathematisches Institut, Universität Göttingen

Tag der mündlichen Prüfung: 24.5.2019

Summary

In the context of nonparametric regression and inverse problems, variational multiscale methods combine multiscale dictionaries (such as wavelets or overcomplete curvelet frames) with regularization functionals in a variational framework. In recent years, these methods have gained popularity in nonparametric statistics due to their good reconstruction properties. Nevertheless, their theoretical performance is, with few exceptions, poorly understood. Further, the computation of these estimators is challenging, as it involves non-smooth large scale optimization problems.

In this thesis we apply variational multiscale methods to the estimation of functions of bounded variation (BV). BV functions are relevant in many applications, since they involve minimal smoothness assumptions and give simplified, interpretable cartoonized reconstructions. These functions are however remarkably difficult to analyze, and there is to date no statistical theory for the estimation of BV functions in dimension $d \geq 2$.

The main theoretical contribution of this thesis is the proof that a class of multiscale estimators with a BV penalty is minimax optimal up to logarithms for the estimation of BV functions in regression and inverse problems in any dimension. Conceptually, our proof exploits a connection between multiscale dictionaries and Besov spaces. We thus leverage tools from harmonic analysis, such as interpolation inequalities, for our theoretical analysis.

Regarding the efficient computation of variational multiscale estimators, we present two approaches: a primal-dual method, and the semismooth Newton method applied to a regularized problem and combined with the path-following technique. We discuss the implementation of these methods and use them to illustrate the performance of multiscale BV estimators in simulations.

The theoretical analysis presented in Chapters 2 and 3 has been partially submitted for publication, and is available under del Álamo et al. (2018) and del Álamo and Munk (2019).

Acknowledgments

I want to thank my supervisor Prof. Axel Munk for his constant support during my PhD. He has given me a lot of freedom to learn and pursue ideas, and at the same time he was always ready to discuss results and share his views and enthusiasm. I also thank him for directing a working group in which an atmosphere of friendship and mutual help reigns. It has been a joy to work here. I also want to thank my second supervisor Prof. Thorsten Hohage for his guidance in inverse problems and for letting me join his oberseminar.

I began to work 6 years ago with Axel and PD Timo Aspelmeier on a project on fluorescence microscopy. I thank very much Timo for his scientific guidance, but also for his sense of humor and his genuinely curious approach to science. Special thanks go to Dr. Housen Li for his support throughout my PhD. His readiness to discuss any topic and his wide understanding of statistics have been of substantial help for my research. Further, I thank both Housen and Dr. Frank Werner for sharing their knowledge on optimization and their code: the computational analysis in this thesis owes very much to them. I am also grateful to Prof. Russell Luke and Dr. Yurii Malitskyi for discussions on how to accelerate the primal-dual algorithm and for sharing their code.

It has been a pleasure to work at the IMS: I therefore thank my colleagues, including secretaries and IT personnel. I wish to thank specially my officemates Dr. Florian Pein, Gytis Kulaitis and Marco Seiler, and also the members of our reading group.

On the financial side, I am grateful for the support of the RTG 2088 “Discovering structure in complex data: Statistics meets Optimization and Inverse Problems”.

Finally, I want to thank my friends: to the ones that are mathematicians, for pouring more fuel to the fire, and to the non-mathematicians for keeping the fire under control and forcing me broaden my view. I cannot thank enough my family for their support and their example: specially their serene approach to problems has always been an inspiration to me. And many thanks go to my girlfriend Marieke, who throughout the years has brought balance and joy to my life.

Contents

CHAPTER 1

Introduction

We consider the problem of estimating a real-valued function f given observations of Tf in the commonly used white noise regression model (see e.g. Brown and Low (1996), Reiß (2008) and Tsybakov (2009))

$$dY(x) = Tf(x)\,dx + \frac{\sigma}{\sqrt{n}}\,dW(x), \quad x \in \mathbb{M}. \tag{1.1}$$

Here $\mathbb{M}$ denotes a Borel-measurable open subset of $\mathbb{R}^d$, $T : L^2(\mathbb{R}^d) \to L^2(\mathbb{M})$ is a linear, bounded operator, and dW denotes a Gaussian white noise process on $L^2(\mathbb{M})$ (defined in Section 2.1). The domain $\mathbb{M}$ in which the data dY is defined is given by the inverse problem under consideration. It is e.g. $\mathbb{M} = \mathbb{R}^d$ if T is a convolution operator or the identity, or $\mathbb{M} = \mathbb{R} \times S^{d-1}$ if T is the Radon transform (Natterer, 1986), where S^{d-1} denotes the d-dimensional unit sphere. See Figure 1.1 for an illustration. The parameter $\sigma n^{-1/2} > 0$ serves as a noise level, and we may assume it to be known, since otherwise it can be estimated efficiently (see e.g. Spokoiny (2002) or Munk et al. (2005)). The parametrization $\sigma n^{-1/2}$ is motivated by the fact that the white noise model (1.1) is an idealization of a nonparametric regression model with n design points and independent normal noise with variance σ^2 (see Section 1.10 in Tsybakov (2009)). Consequently, we see n informally as the sample size, and have the following intuition: the larger n, the lower the noise level in (1.1) and the easier it is to reconstruct f.

In this setting, our goal is to reconstruct the function f from observations dY in (1.1), and to quantify the reconstruction error as the sample size n grows.

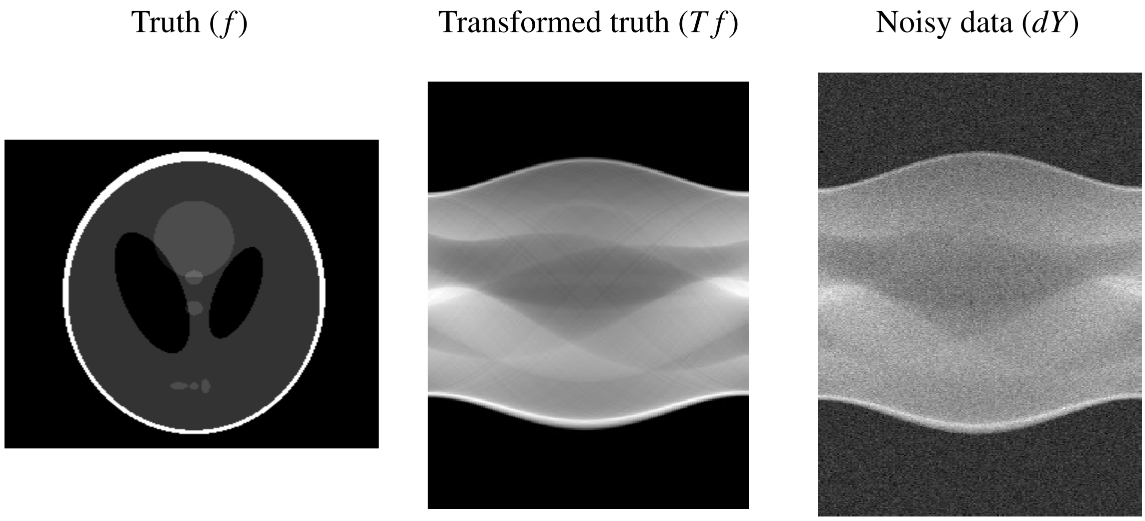
Two clarifications are due: first, observing dY in the model (1.1) means that we have access to a *finite* number of projections

$$\langle \phi, dY \rangle := \langle \phi, Tf \rangle_{L^2} + \frac{\sigma}{\sqrt{n}} \int_{\mathbb{M}} \phi(x)\,dW(x) \tag{1.2}$$

for "test functions" $\phi \in L^2(\mathbb{M})$. The integral against white noise dW is a random variable, as defined in Section 2.1. We stress the word finite, since we want our reconstruction procedure to be computable in finite time. And second, the meaning of "reconstruct f" or "estimate f" here is to come up with a procedure that, based on observations (1.1), produces a function that *resembles* f in some sense. We will measure "resemblance" in an L^q sense, and our benchmark for good performance will be the minimax risk, defined in (1.12).

Without further assumptions, our task seems hopeless: if f can be just any function, then knowing a finite amount of information is not enough for estimating it in a meaningful sense. A way of solving this problem is to impose restrictions on f: these could either concern some qualitative property (e.g. monotony or a general shape constraint (Dümbgen (2003), Guntuboyina and Sen (2018))), or measure smoothness in a quantitative way (e.g. Hölder or Sobolev smoothness (Tsybakov, 2009)). The challenge here is to find conditions that make estimation possible, while still being realistic in applications.

In this thesis we work with the assumption that f is a function of bounded variation (BV), written $f \in BV$, meaning it is in L^1 and its weak partial derivatives of first order are finite Radon measures on $\mathbb{R}^d$. This restriction is not too burdensome: plenty of applications can be modeled with functions of bounded variation. Crucially, the main finding of this thesis is that this restriction is sufficient to enable the reconstruction of f in a statistical setting.



Truth (f)

Transformed truth (Tf)

Noisy data (dY)

Figure 1.1: Shepp-Logan phantom f, its Radon transform Tf and data dY generated by adding Gaussian white noise. Tf is defined on $L^2(\mathbb{R} \times [0, 2\pi))$.

1.1 Methodology

Statistical models of inverse problems like (1.1) are relevant in plenty of applications, such as medical imaging and tomography (Natterer and Wübbeling, 2001), astronomy and microscopy (Bertero et al., 2009), oceanography and weather modeling (Wunsch, 1996), and geology and mining (Tahmasebi et al., 2016), to mention just a few. Most inverse problems of interest are ill-posed, meaning that the operator T does not have a bounded inverse. Consequently, a naive application of T^{-1} to the data dY will amplify the error. This motivates the use of a form of regularization. To that end, several alternative approaches have been proposed, of which we mention a few representative ones: the spectral method based on the singular value decomposition (SVD, see e.g. Bissantz et al. (2007)); dictionary methods, where the observations are projected onto a suitable frame in which denoising and inversion are performed (Cohen et al. (2004), Hoffmann and Reiss (2008)); variational regularization methods, such as Tikhonov (-Phillips) regularization (Phillips (1962), Morozov (1966), Scherzer et al. (2009)); iterative methods with a form of regularization either in the iteration schema or as an early stopping rule (see e.g. Bauer et al. (2009), Blanchard and Mathé (2012)); and Bayesian methods, in which a prior distribution on the function space modeling f has a regularizing effect (see e.g. Stuart (2010), Knapik et al. (2011)). Most related to this work are dictionary-based methods and variational methods, which we briefly discuss from the perspective of this thesis.

a) Dictionary methods. The essential idea of dictionary methods is that, even though T does not have a bounded inverse, it may have *locally* a bounded inverse. We distinguish two variants of this approach, depending on the nature of the localization:

(i) Singular value decomposition (SVD). Let $\{\phi_j\}$ denote an orthonormal basis of L^2 that consists of singular vectors of the adjoint operator T^*, i.e. they satisfy

$$T^*\phi_j = \kappa_j \phi_j$$

for singular values $\kappa_j \to 0$ as $j \to \infty$. Such a basis exists if we assume T to be a compact operator (see the spectral theorem for compact self-adjoint operators, e.g. Theorem VII.3 in Reed and Simon (1972)). The SVD works as follows: if we project the data dY onto the basis ϕ_j, we get

$$\begin{aligned}\langle \phi_j, dY\rangle &= \langle \phi_j, Tf\rangle + \sigma n^{-1/2} \langle \phi_j, dW\rangle \\ &= \kappa_j \langle \phi_j, f\rangle + \sigma n^{-1/2} \epsilon_j.\end{aligned}$$

Roughly, the projections $\langle \phi_j, dY\rangle$ rescaled by the singular value κ_j equal the coefficients of f with respect to the basis ϕ_j plus noise. At this stage, truncation or thresholding of these noisy coefficients yields an estimator for f. Even though SVD-based methods are widely used and enjoy theoretical guarantees for estimating Sobolev and Hölder functions (Cavalier, 2011), they have a crucial weakness: the user has no freedom in choosing the basis $\{\phi_j\}$, which is given solely by the operator T. If it happens that the unknown function f is not sparse in this basis (or if its coefficients do not decay fast enough), then SVD is bound to perform poorly for reconstructing f. This brings us to the second kind of dictionary method.

(ii) Wavelet-vaguelette decomposition (WVD). Donoho (1995) introduced the WVD in order to mitigate the deficiency of the SVD presented above. Given a linear operator T and a wavelet basis $\{\psi_j\}$, his idea was to construct *vaguelette systems* $\{u_j\}$ and $\{v_j\}$ satisfying

$$\begin{aligned} T\psi_j &= \kappa_j v_j \\ T^* u_j &= \kappa_j \psi_j, \end{aligned}$$

along with some additional regularity conditions. Once we have such systems, we project the observations dY onto u_j, which gives us the wavelet coefficients of f rescaled by the singular value κ_j. Performing thresholding in the wavelet domain and transforming back to the image domain, which is known to perform optimally for nonparametric regression (Donoho and Johnstone, 1998), yields a minimax optimal reconstruction of f (Donoho, 1995). The success of this approach and its superiority with respect to the SVD stems from the localizing nature of wavelet bases. A disadvantage of this approach: not all operators have a WVD. However, extensions of the WVD to deal with this problem have been proposed (see e.g. Picard and Kerkyacharian (2006) and references therein).

However, it is known that for denoising multiscale dictionary methods combined with thresholding or truncation may generate artifacts (Gibbs phenomenon). The reason for that is of computational nature: dictionary methods (especially wavelets) are designed for compression, in which a function is represented with as few dictionary elements as possible within a given error, typically measured by an L^q-loss. But having few dictionary elements, which are often oscillatory functions, induces oscillation artifacts in the reconstruction. One way to circumvent this issue is to use overcomplete dictionaries or frames: in doing so, we give up compression properties but gain reconstruction accuracy (Grasmair et al., 2018). Another way to solve this issue is given by variational regularization methods.

b) Variational regularization. This technique uses the assumption (or prior knowledge) that the function f we wish to reconstruct is not arbitrary, but satisfies some regularity property, such as being in a certain function space. Assuming that the regularity of f is measured well by a functional $\mathcal{R}(\cdot)$, we may pose the estimation problem as an optimization problem: Find a function $\hat{f}$ with a *small* $\mathcal{R}(\cdot)$ value and such that $T\hat{f}$ is *close* to the observed data dY, i.e.,

$$\hat{f} \in \operatorname*{argmin}_{g} \mathcal{R}(g) + \mathcal{S}(Tg, dY), \tag{1.3}$$

where $\mathcal{S}(Tg, dY)$ measures the similarity between Tg and dY. A usual choice of $\mathcal{S}(\cdot,\cdot)$ is a Hilbert space distance, although alternatives exist (see e.g. Nemirovski (1985) and Candès and Tao (2007)). On the other hand, a common choice of the regularization functional $\mathcal{R}(\cdot)$ are Sobolev norms, but more subtle alternatives such a Besov (Hohage and Miller, 2019) or BV seminorms (Rudin et al., 1992) have been considered. We remark that the estimator (1.3) has the advantage of automatically producing a function of the right regularity (as measured by $\mathcal{R}$), which limits the effect of artifacts. On the other hand, variational estimators typically lack the spatial adaptation properties characteristic of wavelet methods. The reason is that, for analytical and numerical simplicity, researchers have mostly concentrated on regularization functionals $\mathcal{R}$ that are too smoothing (e.g. a Hilbert space norm). This has the effect of producing oversmoothed reconstructions.

This dichotomy is the starting point of this work: Multiscale dictionary methods are locally adaptive but prone to artifacts, and variational methods avoid artifacts at the price of losing spatial adaptation. In this thesis we propose an estimation framework that combines the local adaptation of multiscale dictionaries with the smoothness guaranties of variational regularization with the BV seminorm (see Section 2.1). Since the BV seminorm is mildly smoothing, it preserves the local reconstruction properties of dictionary methods. We prove that the proposed estimators are minimax optimal up to logarithmic factors for estimating BV functions in any dimension for a variety of inverse problems, including denoising ($T = id$), Radon inversion and deconvolution.

Functions of bounded variation

Functions of bounded variation are L^1 functions whose weak gradients are finite Radon measures. They satisfy very weak regularity properties, and are suitable to model objects with discontinuities. This is a desirable property for instance in medical imaging applications, where sharp transitions between tissues occur, and smoother functions would represent them inadequately. Consequently, BV functions have been studied extensively in the applied and computational analysis literature, see e.g. Chambolle and Lions (1997), Meyer (2001), Rudin et al. (1992), Scherzer et al. (2009) and references therein.

Remarkably, the very reason for the success of functions of bounded variation in applications, namely their low smoothness, has hindered the development of a rigorous theory for the corresponding estimators in a statistical setting. In dimension $d = 1$, Mammen and van de Geer (1997) showed that the least squares estimator with a total variation (TV) penalty attains the minimax optimal convergence rates when T is the identity operator. Further, Donoho and Johnstone (1998) proved the optimality of wavelet thresholding over BV in $d = 1$ and $T = id$, while Donoho (1995) extended these results to operators T admitting a WVD. In contrast, there are to the best of our knowledge no statistical guarantees for estimating BV functions in dimension $d \geq 2$. Roughly speaking, the main challenges in higher dimensions are twofold: first, the embedding $BV \hookrightarrow L^\infty$ *fails* if $d \geq 2$; and second, the space BV does not admit a characterization in terms of the size of wavelet coefficients. This makes wavelet thresholding unsuitable for estimating BV functions. More generally, the space BV does not admit an unconditional basis (see Sections 17 and 18 in Meyer (2001)). In statistical terms this means that purely dictionary-based methods are doomed to perform poorly for estimating BV functions.

On the other hand, the failure of the embedding into L^∞ for $d \geq 2$ is related to the fact that BV behaves roughly like Sobolev spaces $W^{s,p}$ with $s < d/p$. These spaces contain discontinuous functions, and statistical estimation there is challenging and has received little attention. One contribution of this thesis is to characterize the minimax estimation rates in these spaces.

An alternative route to estimating BV functions in higher dimension is to discretize the observational model. This approach has seen recent successes (see e.g. Hütter and Rigollet (2016), Dalalyan et al. (2017)), which we discuss in more detail in Section 1.4 below.

1.2 Multiscale total variation estimation

As stressed above, we want to construct a variational estimator of the form (1.3) which enjoys the benefits of multiscale dictionaries. A way to achieve that is to include a multiscale dictionary in the data-fidelity $\mathcal{S}(\cdot,\cdot)$. While there are several ways of doing so, we propose to use

$$\mathcal{S}(Tg, dY) := \max_{\omega\in\Omega_n} \left|\langle u_\omega, Tg\rangle - \langle u_\omega, dY\rangle\right|, \tag{1.4}$$

where $\{u_\omega\}$ is a vaguelette system associated with the operator T, and Ω_n is a finite set of indices, typically corresponding to different locations and scales. In this thesis we consider the variational estimator (1.3) with data-fidelity (1.4) in *constrained* form, i.e.,

$$\hat{f}_n \in \underset{g\in\mathcal{F}_n}{\operatorname{argmin}} |g|_{BV} \text{ subject to } \max_{\omega\in\Omega_n} \left|\langle u_\omega, Tg\rangle - \langle u_\omega, dY\rangle\right| \leq \gamma_n, \tag{1.5}$$

where γ_n is a threshold to be chosen, and we minimize over a set of functions $\mathcal{F}_n$ to be specified later. Notice that the operator T is inverted indirectly by the dictionary elements u_ω. Indeed, by the definition of the vaguelettes, the data-fidelity (1.4) is actually a constraint on the wavelet coefficients of g: they are forced to be close to the wavelet coefficients of the unknown function

f, up to noise terms. Consequently, $\hat{f}_n$ will enjoy the spatial adaptation properties of wavelet methods, while the regularization term $|g|_{BV}$ in (1.5) ensures that $\hat{f}_n$ is well-behaved in the BV norm.

Example 1. In order to illustrate the estimator $\hat{f}_n$, consider the situation where $d = 2$, $T = id$, and the multiscale dictionary consists of normalized indicator functions of dyadic squares (Nemirovski, 2000),

$$\Phi = \left\{ \frac{1}{\sqrt{|B|}} 1_B(x) \,\middle|\, B \text{ dyadic square } \subseteq [0,1]^2 \right\},$$

where $|B|$ denotes the Lebesgue measure of the set B. Consider a particular estimator $\hat{f}_n$ of the form (1.5) as

$$\hat{f}_n \in \underset{g \in \mathcal{F}_n}{\operatorname{argmin}} |g|_{BV} \text{ s.t. } \max_{\text{dyadic } |B| \geq \frac{1}{n}} \frac{1}{\sqrt{|B|}} \left| \int_B g(x) - f(x)\, dx - \frac{\sigma}{\sqrt{n}} \int_B dW(x) \right| \leq \gamma_n, \tag{1.6}$$

that is, Ω_n consists of all squares $B \subseteq [0,1]^2$ of size $|B| \geq 1/n$ with vertices at dyadic positions. The main peculiarity of $\hat{f}_n$ is the data-fidelity term, which encourages proximity of $\hat{f}_n$ to the truth f *simultaneously* at all large enough dyadic squares B. This results in an estimator that preserves features of the truth in both the large and the small scales, thus giving a *spatially adaptive* estimator. This is illustrated in Figure 1.2 (see Chapter 4 for an algorithmic implementation): the multiscale TV-estimator $\hat{f}_n$ is represented in the lower left corner, and it succeeds to reconstruct the image well at both the large (sky and building) and small scales (stairway). We show for comparison the classical L^2-TV-regularization estimator, also known as Rudin-Osher-Fatemi (ROF) estimator (Rudin et al., 1992)

$$\hat{f}_\lambda \in \underset{g}{\operatorname{argmin}} \|g - Y\|_2^2 + \lambda |g|_{BV}, \tag{1.7}$$

which employs a global L^2 data-fidelity term. The parameter λ is chosen here in an oracle way so as to minimize the distance to the truth, where we measure the "distance" by the symmetrized Bregman divergence of the BV seminorm (see Chapter 5). As seen in Figure 1.2, the L^2-TV estimator successfully denoises the image in the large scales at the cost of details in the small scales. The reason is simple: the use of the L^2 norm as a data-fidelity, which measures the proximity to the data *globally*. This means that the optimal parameter λ is forced to achieve the best trade-off between regularization and data fidelity *in the whole image*: in particular, in rich enough images there will be regions where one either over-regularizes or under-regularizes, e.g. in the stairway in Figure 1.2. Finally, we also show the curvelet thresholding estimator in Figure 1.2. As expected, curvelet thresholding performs excellently on elongated structures (stairway), but it introduces artifacts in locally constant regions (sky, building). In Chapter 5 we present a broader quantitative comparison study of different methods.

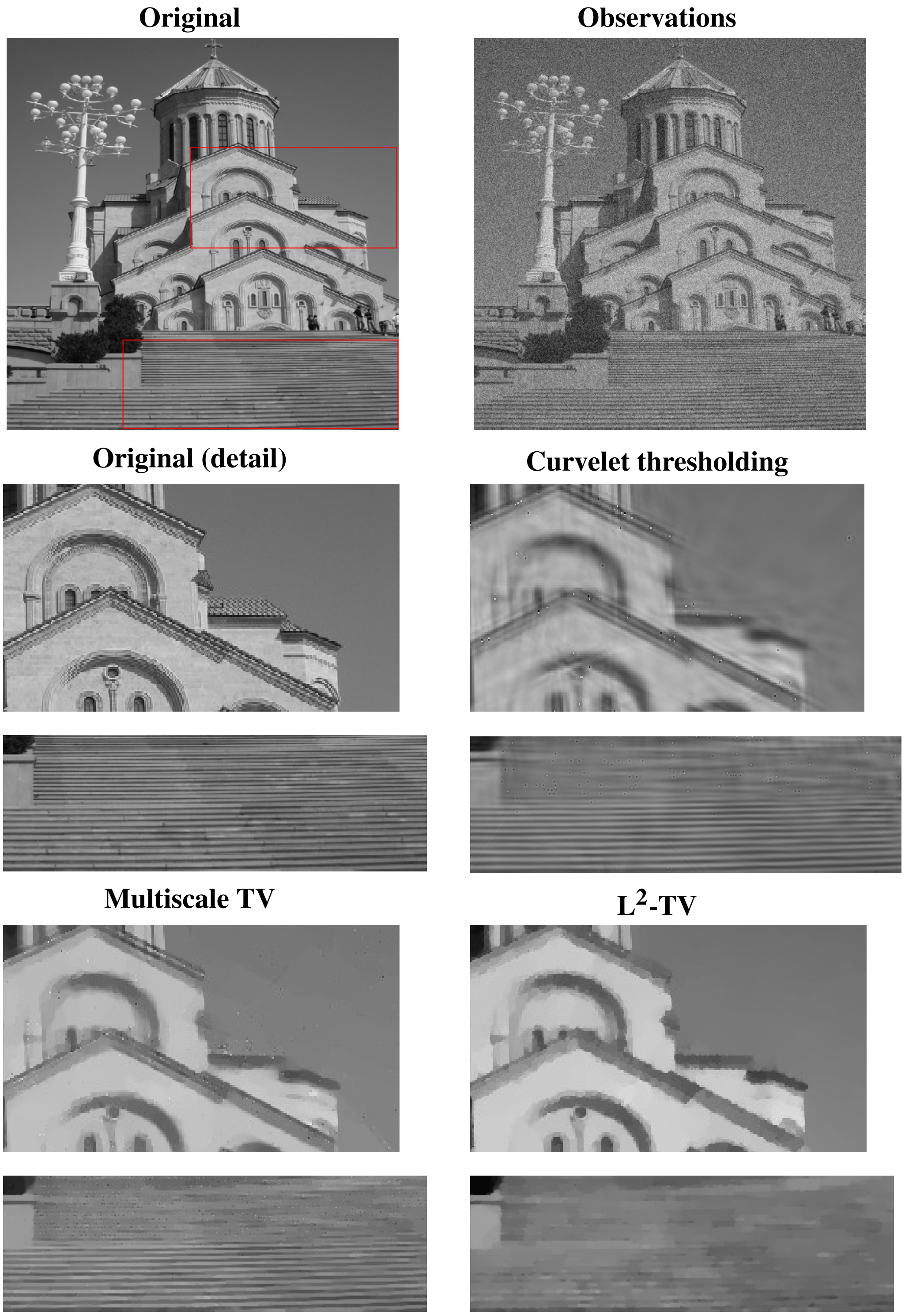


Figure 1.2: Row-wise, from top to bottom: original image and noisy version with signal-to-noise ratio $\sigma^{-1}\,\|f\|_{L^\infty} = 5$; zoom in of the original image and of the curvelet thresholding estimator; zoom in of the multiscale TV-estimator (1.5) and of the estimator $\hat{f}_\lambda$ from (1.7) with oracle $\lambda^* = \operatorname{argmin} \mathbb{E}[D_{BV}(\hat{f}_\lambda, f)]$, where $D_{BV}(\cdot,\cdot)$ denotes the symmetrized Bregman divergence of the BV seminorm. See Chapter 5 for the details of the simulation.

Choice of the threshold γ_n

Both the constrained minimization (1.5) and the penalized minimization problem (1.3) involve tuning parameters γ_n and λ that have to be chosen. Crucially, there is an optimal choice for γ_n and λ, in the sense that choosing a smaller parameter leads to overfitting the data, and choosing a larger parameter induces oversmoothing.

In penalized estimation, the optimal parameter λ typically depends on the unknown function f, and there are data-driven approaches to estimate it, such as e.g. cross validation (Wahba, 1977), or a version of Lepskii's balancing principle (Lepskii, 1991) for inverse problems (Mathé and Pereverzev, 2003).

We prefer constrained over penalized minimization because the optimal γ_n depends on the noise model but not on f, and it can be computed using known or simulated quantities only. To see that the optimal γ_n is independent of f, consider the following trade-off: the smaller γ_n, the fewer functions satisfy the constraint in (1.5). Since the best reconstruction we can hope for is the true regression function f, the optimal γ_n is the one that is large enough to let f be a feasible function, but no larger. In this sense, note that f satisfies the constraint in (1.5) precisely when

$$\max_{\omega\in\Omega_n} \left|\langle u_\omega, Tf\rangle - \langle u_\omega, dY\rangle\right| = \max_{\omega\in\Omega_n} \frac{\sigma}{\sqrt{n}}\left|\langle u_\omega, dW\rangle\right| \leq \gamma_n. \tag{1.8}$$

Assume for a moment that $u_\omega \in L^2$ with $\|u_\omega\|_{L^2} = 1$ for all ω. Then the left-hand side is the maximum of the absolute value of $\#\Omega_n$ standard normal random variables times $\sigma n^{-1/2}$. Consequently, a simple computation (see the claim in equation (2.12)) implies that (1.8) holds asymptotically almost surely for the *universal threshold*

$$\gamma_n = \kappa\sigma n^{-1/2}\sqrt{2\log\#\Omega_n}, \tag{1.9}$$

with κ depending on the dictionary Φ in an explicit way (see Theorem 4). This argument can be adapted to the case that the u_ω do not have norm one, as long as they remain bounded above and below by positive constants. We remark that this universal choice of the parameter γ_n appears to us as a great conceptual and practical advantage of the estimator (1.5), in contrast to penalized estimators such as (1.7) requiring more complex parameter-choice methods (e.g. Lepskii (1991) or Wahba (1977)).

Multiscale data-fidelity

There are several reasons why the multiscale data-fidelity (1.4) is preferable over more classical choices, such as the L^2-norm. For the sake of simplicity, we illustrate them here in the case where T is the identity and $\{u_\omega\}$ is an orthonormal wavelet basis. In that case, the multiscale constraint in (1.5) requires the wavelet coefficients of $\hat{f}_n$ to be close to the coefficients of f, up to noise terms:

$$\left|\langle u_\omega, \hat{f}_n\rangle - \langle u_\omega, f\rangle - \sigma n^{-1/2}\langle u_\omega, dW\rangle\right| \leq \gamma_n \ \ \forall \omega \in \Omega_n.$$

In particular, similarity between $\hat{f}_n$ and f is required at all positions in all scales. On the other hand, using the L^2 data-fidelity and writing it in terms of the wavelet basis (which is possible by orthonormality) imposes a constraint of the form

$$\sum_{\omega\in\Omega_n} \left|\langle u_\omega, \hat{f}_n\rangle - \langle u_\omega, f\rangle - \sigma n^{-1/2}\langle u_\omega, dW\rangle\right|^2 \leq L_n^2. \tag{1.10}$$

This is a constraint on the average error, and it enforces similarity between $\hat{f}_n$ and f on average, and not pointwise. We have seen above that the optimal choice of γ_n is given by (1.9), which implies that (1.8) holds asymptotically almost surely. For the L^2 data-fidelity we choose the threshold L_n analogously, i.e., such that the true function $\hat{f}_n = f$ satisfies (1.10) with high probability. In that case, the summands in (1.10) would be squares of independent normal random variables (by orthogonality of u_ω), so L_n^2 should be a quantile of a χ^2 random variable with $\#\Omega_n$ degrees of freedom. This gives roughly $L_n \sim \sigma n^{-1/2}\sqrt{\#\Omega_n}$. The difference between the multiscale and L^2 constraints is now apparent:

$$\text{multiscale constraint: } \ell^\infty \text{ ball of radius } \sigma n^{-1/2}\sqrt{2\log\#\Omega_n},$$
$$L^2 \text{ constraint: } \ell^2 \text{ ball of radius } \sigma n^{-1/2}\sqrt{\#\Omega_n},$$

where both constraints are on the wavelet domain. Due to the norm equivalence $\|x\|_{\ell^\infty} \leq \|x\|_{\ell^2} \leq \sqrt{\#\Omega_n}\,\|x\|_{\ell^\infty}$, $\forall x \in \ell^\infty(\Omega_n)$, the difference between the constraints may not seem excessive. However, the difference is considerable. Indeed, in this thesis we choose the number of constraints $\#\Omega_n$ to behave polynomially in n (see Assumption 4). Consequently, the radius in the multiscale constraint tends to zero as $n \to \infty$, while the radius in the L^2 constraint tends to a constant or diverges if $n = O(\#\Omega_n)$. Hence, the multiscale constraint set is much smaller for n large, and we expect the multiscale data-fidelity to produce more faithful reconstructions.

The constraint in (1.5) can also be interpreted from a hypothesis testing perspective (Lehmann and Romano, 2006). Given a candidate function g, we can ask how likely it is that the observed data dY arose from g. The question can be made precise by testing, for each $\omega \in \Omega_n$, the hypothesis

$$H_\omega : \langle u_\omega, g\rangle = \langle u_\omega, f\rangle \quad \text{against} \quad K_\omega : \langle u_\omega, g\rangle \neq \langle u_\omega, f\rangle.$$

The log-likelihood ratio test for testing this hypothesis under model (1.1) is given by $|\langle u_\omega, g\rangle - \langle u_\omega, dY\rangle|$, so the multiscale data-fidelity (1.4) is a test statistic for testing the hypotheses H_ω simultaneously for all $\omega \in \Omega_n$. Choosing γ_n appropriately, the constraint in (1.5) includes exactly the functions that pass all these tests.

Finally, there is a seemingly unrelated yet crucial reason for using (1.4) as a data-fidelity term. For $T = id$ and $\{u_\omega\}$ a smooth enough wavelet basis, the multiscale data-fidelity (1.4) is a truncation of the Besov $B^{-d/2}_{\infty,\infty}$ norm of $g - dY$, seen as a random temperate distribution. More precisely, we have

$$||g||_{B^{-d/2}_{\infty,\infty}} \leq C \max_{\omega \in \Omega_n} \left|\langle u_\omega, g \rangle\right| + C\frac{||g||_{L^\infty}}{\sqrt{n}} \tag{1.11}$$

for any function $g \in L^\infty$ and a suitable set Ω_n. This is a Jackson-type inequality (Cohen, 2003), representing how well a function can be approximated in the Besov $B^{-d/2}_{\infty,\infty}$ norm by its coefficients with respect to $\{u_\omega\}$. It is well-known that smooth enough wavelet bases satisfy this condition (Cohen, 2003). In Section 2.4 we will show (1.11) for more general multiscale systems, e.g. systems of indicator functions of dyadic cubes, and mixed frames of wavelets and curvelets and of wavelets and shearlets. Remarkably, inequality (1.11) allows us to relate the statistical multiscale constraint in (1.4) to an analytic object: the Besov norm. This connection allows us to leverage tools from harmonic analysis to analyze the performance of the estimator (1.5).

Besides the mathematical reasons just given, there is also a practical motivation for using multiscale data-fidelites. In fact, multiscale dictionaries are widely used and known to perform well since the introduction of wavelets (see e.g. Daubechies (1992) and Donoho (1993)). Moreover, overcomplete frames such as curvelets (Candès and Donoho, 2000), shearlets (Labate et al. (2005), Guo et al. (2006)) and other multiresolution systems (see Haltmeier and Munk (2014) for a survey) have been shown to perform well in theory and numerical applications, specially in imaging. Several works have proposed variants of the multiscale data-fidelity (1.4) in a variational estimation setting (Meyer (2001), Starck et al. (2001) Durand and Froment (2001), Malgouyres (2001), Candès and Guo (2002), Malgouyres (2002), Osher et al. (2003), Haddad and Meyer (2007) Garnett et al. (2007)). Closer to our work, multiscale methods using overcomplete frames in combination with a BV penalty have been empirically shown to yield promising results for function estimation (Malgouyres (2002), Candès and Guo (2002), Dong et al. (2011), Frick et al. (2012), Frick et al. (2013)). The theory in those cases is still lacking, which motivates the present work.

Challenges

Until now we have motivated the estimator (1.5) as a synthesis of very successful techniques for solving inverse problems, and we have illustrated and explained the multiscale constraint. Before we turn to the discussion of the optimal convergence properties of $\hat{f}_n$, let us admit two limitations of the multiscale TV-estimator. First, not every operator T has an associated vaguelette system $\{u_\omega\}$, as we use in (1.5). In fact, only reasonably homogeneous operators have such a system (see Donoho (1995)). On the other hand, for our theory we do not need the whole generality of the WVD (see Assumption 4 in Chapter 3), and many practically relevant operators such as the Radon transform, convolution or integration satisfy our assumptions (see Examples 2 in Chapter 3).

The second limitation concerns the solution of the optimization problem in (1.5), which is a non-smooth, high dimensional optimization problem (since n and $\#\Omega_n$ might be large). Due to the non-smoothness, standard interior point methods (Nesterov and Nemirovsky, 1994) are not applicable here, and the large number of variables makes it a challenging optimization problem. However, the computation of (1.1) is now feasible due to recent progress in convex optimization, e.g. in primal-dual methods (Chambolle and Pock, 2011) and acceleration thereof (Malitsky and Pock, 2018), and in semismooth Newton methods with the path-following technique (Clason et al., 2010). In Chapter 4 we present different approaches to compute the minimum in (1.5), and discuss their advantages and disadvantages in terms of runtime and precision.

1.3 Main results

The main result of this thesis states that the estimator (1.5) is minimax optimal (up to logarithmic factors) for estimating BV functions in any dimension for a family of inverse problems. The concept of minimax optimality is based on the notion of minimax risk over a set of functions X, which is a measure of the difficulty of a statistical problem and a benchmark for the performance of estimators. It is defined as the error of the best estimator in the most difficult instance in the set X, i.e.,

$$\mathcal{R}(L^q, X) := \inf \left\{ \sup_{f \in X} \mathbb{E}_f \|\hat{f} - f\|_{L^q} \,\middle|\, \hat{f} \text{ is an estimator using (1.1)} \right\}, \tag{1.12}$$

where the infimum runs over *all* estimators, i.e., over all measurable functions $\hat{f} : \mathcal{Y}_n \to L^2(\mathbb{R}^d)$, where $\mathcal{Y}_n$ is the sample space where the process in (1.1) takes values (see Section 1.2.2 in Giné and Nickl (2015) for more details). Here, the expectation is taken with respect to the measure that generates the observations, which depends on f. The error is measured here in an L^q-sense. The minimax rate over X with respect to the L^q-risk is defined as the rate at which $\mathcal{R}(L^q, X)$ tends to zero as the noise level in (1.1) tends to zero, i.e., as $n \to \infty$.

In order to formulate our results, define for $L > 0$ the parameter set

$$BV_L := \{g \in BV \cap \mathcal{D}(T) \,\big|\, |g|_{BV} \leq L, \;\; \|g\|_{L^\infty} \leq L, \;\; \operatorname{supp} g \subseteq [0,1]^d\}, \tag{1.13}$$

where $\mathcal{D}(T) \subset L^2(\mathbb{R}^d)$ denotes the domain of the operator T. In Theorem 5 below we show that the minimax rate over the set BV_L satisfies

$$\liminf_{n\to\infty} n^{\min\{\frac{1}{d+2\beta+2}, \frac{1}{(d+2\beta)q}\}} \mathcal{R}(L^q, BV_L) > 0,$$

where $\beta \geq 0$ is the degree of ill-posedness of the operator T. This means that no estimator can have an L^q-error tending to zero strictly faster than $n^{-\min\{\frac{1}{d+2\beta+2}, \frac{1}{(d+2\beta)q}\}}$ uniformly over BV_L. For given $d, \beta \geq 0$ and $q \in [1, \infty]$, define the number

$$\vartheta_{q,\beta} := \begin{cases} \frac{1}{d+2\beta+2} & \text{for } q \leq 1 + 2/(d+2\beta) \\ \frac{1}{q\,(d+2\beta)} & \text{for } q > 1 + 2/(d+2\beta). \end{cases} \tag{1.14}$$

Our main theorem can be stated informally as follows.

Theorem 4 (Informal)**.** Let the dimension $d \geq 2$, and for $\beta \geq 0$ let T have a WVD with singular values behaving as $\kappa_j = 2^{-j\beta}$ (see Assumption 4 in Chapter 3). Let the threshold γ_n be as in (1.9) for $\kappa > \kappa^*$ depending on T and d only. Then the estimator $\hat{f}_n$ attains the *minimax optimal* rate of convergence over BV_L up to a logarithmic factor,

$$\sup_{f \in BV_L} \mathbb{E}[\|\hat{f}_n - f\|_{L^q}] \leq C_L\, n^{-\vartheta_{q,\beta}} \log n \tag{1.15}$$

for n large enough, for any $q \in [1, \infty)$, any $L > 0$ and a constant $C_L > 0$ independent of n, but dependent on L, σ, d and T. For $d = 1$, (1.15) holds with an additional $\log n$ factor.

The estimator $\hat{f}_n$ is nearly optimal in the sense that there exists no estimator such that the left-hand side of (1.15) is $o(n^{-\vartheta_{q,\beta}})$.

The theorem refers to inverse problems for which T has a WVD. As we show in Chapter 3, this includes the cases of regression $T = id$, Radon inversion, and deconvolution.

The theorem proves convergence when the function f is supported on the unit cube, as stated in (1.13). The reason for this constraint is that, since we only have a finite amount of information, we cannot hope to recover a function with infinite support. The restriction to the unit cube is in a sense arbitrary: any regular enough compact set would do. While the restriction to compactly supported functions is a common practice in nonparametric statistics, there is an alternative: to assume that the regression function f is periodic, i.e. defined on the torus $\mathbb{T}^d$. See for instance Grasmair et al. (2018) for an example of function estimation under a periodicity assumption.

The proof of Theorem 4 relies on the compatibility between the multiscale constraint and the $B^{-d/2-\beta}_{\infty,\infty}$ norm, as expressed in (1.11) for $\beta = 0$. This allows us to use techniques from harmonic analysis to analyze $\hat{f}_n$, such as the interpolation inequality between the spaces $B^{-d/2-\beta}_{\infty,\infty}$ and BV,

$$\|g\|_{L^q} \leq C \|g\|_{B^{-d/2-\beta}_{\infty,\infty}}^{\frac{2}{d+2\beta+2}} \|g\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}} \quad \forall g \in B^{-d/2-\beta}_{\infty,\infty} \cap BV \tag{1.16}$$

for any $q \in [1, \frac{d+2\beta+2}{d+2\beta}]$, $d \geq 2$. A variant of this inequality was proven in Cohen et al. (2003)

by a delicate analysis of the wavelet coefficients of functions of bounded variation (see Ledoux (2003) for an alternative approach). The inequality (1.16) is the first step towards bounding the L^q-risk of $\hat{f}_n$: inserting $g = \hat{f}_n - f$ we can bound it in terms of the $B^{-d/2-\beta}_{\infty,\infty}$ and the BV-risks. The BV-risk is bounded by a constant with high probability, while the $B^{-d/2-\beta}_{\infty,\infty}$-risk can be related to the multiscale data-fidelity in (1.5). In fact, under suitable assumptions we have

$$\begin{aligned} \|\hat{f}_n - f\|_{B^{-d/2-\beta}_{\infty,\infty}} &\leq C \max_{\omega\in\Omega_n} \left|\langle u_\omega, T\hat{f}_n\rangle - \langle u_\omega, Tf\rangle\right| + C\,\|\hat{f}_n - f\|_{L^\infty}\, n^{-1/2} \\ &\leq C \max_{\omega\in\Omega_n} \left|\langle u_\omega, T\hat{f}_n\rangle - \langle u_\omega, dY\rangle\right| + C\frac{\sigma}{\sqrt{n}} \max_{\omega\in\Omega_n} \left|\int_{\mathbb{M}} u_\omega(x)\, dW(x)\right| \\ &\quad + C\,\|\hat{f}_n - f\|_{L^\infty}\, n^{-1/2}. \end{aligned}$$

The first term is bounded by $\gamma_n = O(n^{-1/2}\sqrt{\log \#\Omega_n})$ by construction, and it represents the error that we allow the minimization procedure to make. The second term behaves as $O(n^{-1/2}\sqrt{\log \#\Omega_n})$ asymptotically almost surely, and it represents the stochastic error arising from the randomness of the observations. The third term is a truncation error, stemming from the use of only a finite amount of information. Inserting the result in (1.16) yields the conclusion that $\|\hat{f}_n - f\|_{L^q} \leq C\, n^{-\frac{1}{d+2\beta+2}} \log n$ with high probability for $q \leq 1 + 2/(d+2\beta)$. The bound for $q > 1 + 2/(d+2\beta)$ follows from Hölder's inequality applied between $L^{1+2/(d+2\beta)}$ and L^∞. For $d = 1$ we proceed analogously with some modifications. In Section 2.3 we give a more detailed sketch of the proof.

Minimax risk over Besov spaces

As stated in Theorem 4, the minimax rate over BV_L presents a sharp transition depending on the L^q-risk: it is $n^{-\frac{1}{d+2\beta+2}}$ for $q \leq 1 + 2/(d+2\beta)$, and it deteriorates to $n^{-\frac{1}{q(d+2\beta)}}$ otherwise. A remarkable consequence is that the L^∞ minimax risk does not tend to zero, i.e., there is no estimator that is L^∞-consistent uniformly over BV functions.

More generally, this behavior is characteristic of Besov spaces $B^s_{p,t}$ for $s \leq d/p$. This was observed for the first time by Goldenshluger and Lepskii (2014) and Lepskii (2015) in the context of density and function estimation, respectively. They considered anisotropic Nikolskii spaces, which in the isotropic case coincide with the Besov spaces $B^s_{p,\infty}$, and in general allow for different smoothness and integrability indices for different spatial directions. In Theorem 6 we generalize their results in the isotropic case and establish the minimax rates for regression and mildly ill-posed inverse problems over all spaces

$$(B^s_{p,t} \cap L^\infty)_L := \{g \in B^s_{p,t} \cap L^\infty \mid \|g\|_{B^s_{p,t}} \leq L,\ \ \|g\|_{L^\infty} \leq L,\ \ \operatorname{supp} g \subseteq [0,1]^d\} \tag{1.17}$$

for $s \leq d/p$, $s > 0$, $p, t \in [1,\infty]$ and $L > 0$.

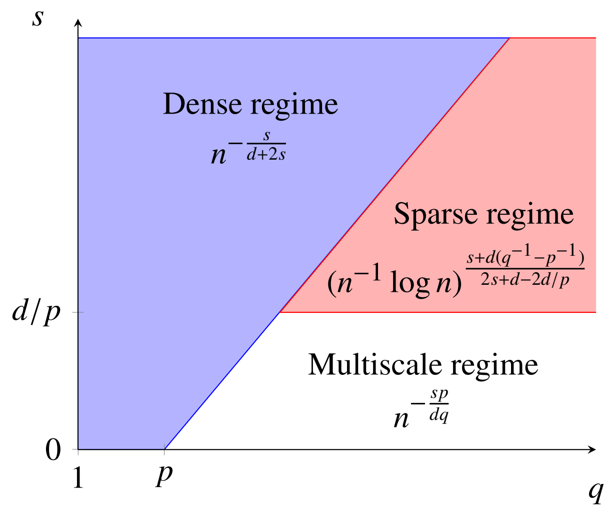


Figure 1.3: Regimes for the minimax rates for regression ($\beta = 0$) over Besov $B^s_{p,t}$ spaces, together with the associated rates. The sloped line is given by $q = p(1 + 2s/d)$.

Our result completes the picture of minimax rates over Besov spaces. Beyond the well-known dense and sparse regimes, which correspond to $q/p < 1+2s/(d+2\beta)$ and $q/p \geq 1+2s/(d+2\beta)$, $s > d/p$, respectively, our results concern the regime $q/p \geq 1 + 2s/(d + 2\beta)$ and $s \leq d/p$. The three regimes are depicted in Figure 1.3 for $\beta = 0$. The new regime, in which the minimax rate behaves differently than in the others, is in a sense a middle ground between the dense and the sparse regime. Indeed, the minimax risk in the dense regime is driven by functions with mass everywhere, meaning that those functions are the most challenging to estimate. On the other hand, the minimax risk in the sparse regime is driven by localized spikes. In the new regime, the risk is driven by blocks of spikes at different locations and scales, and the precise amount of spikes depends on the quantity $d - sp \geq 0$. For that reason, we refer to it as *multiscale regime*.

1.4 Related work and contributions

In spite of the success of BV functions in imaging applications (see e.g. Scherzer et al. (2009) and references therein), there are surprisingly few works that analyze the estimation of BV functions in a statistical setting. Starting with the seminal paper of Rudin et al. (1992) that proposed the TV-regularized least squares (ROF) estimator for image denoising, the subsequent development of TV-based estimators depends greatly on the spatial dimension.

In dimension $d = 1$, Mammen and van de Geer (1997) showed that the ROF-estimator attains the optimal rates of convergence in the discretized nonparametric regression model, and Donoho and Johnstone (1998) proved the optimality of wavelet thresholding for estimation over BV. We also refer to Davies and Kovac (2001) and Dümbgen and Kovac (2009) for a combination of TV-regularization with related multiscale data-fidelity terms in $d = 1$, and to Li et al. (2017) for the combination of a multiscale constraint with a jump penalty for segmentation of one-dimensional functions. In statistical inverse problems, the only work proving minimax optimal convergence rates for the estimation of BV is, to the best of our knowledge, Donoho (1995). He shows that thresholding of the WVD is minimax optimal over a range of Besov spaces $B^s_{p,t}$ and for a class of β-smoothing inverse problems, meaning that the singular values of the operator T behave as $\kappa_j = 2^{-j\beta}$. In the case relevant for BV ($s = p = 1$), minimax optimality holds for the range $\beta < 1 - d/2$, i.e. for β-smoothing operators in dimension $d = 1$ and $\beta \in [0, 1/2)$. The present work is hence an improvement, since we do not impose any limitation on β nor on the dimension d. On the other hand, our estimator is suboptimal by the $\log n$ factor in (1.15), while Donoho's estimator achieves the exact minimax rate.

In higher dimensions, the situation becomes more involved due to the low regularity of functions of bounded variation. There are roughly two approaches to deal with this: either employ a finer data-fidelity term, or discretize the problem. Concerning the first approach, we distinguish three different variants of the ROF-model that are related to our setting. First, Meyer (2001) proposed the replacement of the L^2-norm in the ROF functional by a weaker norm designed to match the smoothness of Gaussian noise. Several algorithms and theoretical frameworks using the Besov norm $B^{-1}_{\infty,\infty}$ (Garnett et al., 2007), the G-norm (Haddad and Meyer, 2007) and the Sobolev norm H^{-1} in $d = 2$ (Osher et al., 2003) were proposed, but the statistical performance of these estimators was not analyzed. A different approach started with Durand and Froment (2001), Malgouyres (2001) and Malgouyres (2002), who proposed estimators of the form (1.5) with a wavelet basis. Following this approach and the development of curvelets (see e.g. Candès and Donoho (2000) for an early reference), Candès and Guo (2002) and Starck et al. (2001) proposed the estimator (1.5) with a curvelet frame and a mixed curvelet and wavelet family, respectively, which showed good numerical behavior. A third line of development that leads to the estimator (1.5) began with Nemirovski (1985) (see also Nemirovski (2000)). He proposed a variational estimator for nonparametric regression over Hölder and Sobolev spaces that used a data-fidelity term based on the combination of local likelihood ratio tests: the *multiresolution norm*. That type of data-fidelities were also proposed by Frick et al. (2012) and Frick et al. (2013) in combination with a BV penalty. In statistical inverse problems, Dong et al. (2011) proposed an estimator using TV-regularization constrained by the *sum* of local averages of residuals, instead of the maximum we employ in (1.5). In a nutshell, the situation (both in regression and in inverse problems) for the estimation of BV functions in dimension $d \geq 2$ is the following: a plethora of estimation procedures has been proposed, many of which employ data-fidelity terms weaker than the L^2-norm. Nevertheless, no convergence guaranty has been proven for any of these methods. In that sense, this thesis presents the first statistical analysis of a method for estimating BV functions in regression and inverse problems in higher dimensions. Moreover, we prove that such method is optimal in a minimax sense up to logarithms.

The other approach to TV-regularization in higher dimensions is to discretize the observational model (1.1), thereby reducing the problem of estimating a function $f \in BV$ to that of estimating a vector of function values $(f(x_1), \dots, f(x_n)) \in \mathbb{R}^n$. In particular, the risk is measured by the *Euclidean norm* of $\mathbb{R}^n$, and not by the continuous L^2-norm. TV-regularized least squares in this discrete setting is nowadays fairly well understood. The recent works by Hütter and Rigollet (2016) and Dalalyan et al. (2017) proved convergence of the TV least squares estimator in any dimension in a variety of discretized models, including functions defined on certain graphs. These rates were shown to be minimax optimal (Sadhanala et al., 2016). Also, the generalization from BV to trend-filtering is a current research topic (Guntuboyina et al. (2017), Wang et al. (2016)). However, this discretized model is radically different from the continuous model we consider. To see that, notice that BV functions are indistinguishable from Sobolev $W^{1,1}$ functions in the discretized model. Conversely, BV functions can have jump singularities, which makes their estimation significantly more challenging than estimating a Sobolev function. Therefore, the analysis of discrete TV-regularization is inspiring, but it regrettably does not solve the problem in the continuous setting: different and genuinely continuous tools are needed, such as the interpolation inequality (1.16). Another drawback of this approach is that the BV seminorm is quite sensitive to discretization. In fact, it has been shown that the minimizers of the discretized TV-regularized least squares estimator do not necessarily converge to their continuous counterparts in a reasonable sense as the discretization tends to zero (see Lassas and Siltanen (2004) and Section 4.2 below for more details). Besides, a limitation of discretized models is that they typically discretize the functions and the BV seminorm with respect to the *same* grid. The discretization of the signals is usually determined by the application, but different discretizations of the BV seminorm can have different effects, so it might be desirable to choose how to discretize it (see e.g. Condat (2017)). It is hence useful to study the estimation of BV functions in the continuous setting, since it gives insight on how the estimation problem is, independently of the discretization of signals or functionals.

An interesting connection of our result with discrete models is that the minimax rate of estimation of BV functions with respect to the discrete L^2-risk was shown by Sadhanala et al. (2016) to be $n^{-\min\{\frac{1}{d+2}, \frac{1}{2d}\}}$ up to logarithms. This coincides with the rate in Theorem 4 for $q = 2$, so our results explain the phase transition in this rate as arising from the use of the L^2 risk. Furthermore, the same rate was shown by Han et al. (2017) to be minimax for estimating bounded, componentwise isotone function in the discrete model, again with respect to the discrete L^2-risk. This means that the statistical complexity of estimating BV functions equals the complexity of estimating isotone functions: this result is well-known in dimension $d = 1$, but we are not aware of any such result in $d \geq 2$.

At a technical level, our work is inspired by several sources. We have already mentioned Donoho (1995), who introduced the WVD as a means for using wavelet methods in inverse problems (see also Abramovich and Silverman (1998) for a variant of the WVD, and Candès and Donoho (2002) for a refined approach for Radon inversion). Besides these works, there have been several approaches that implicitly use the WVD idea. We refer to Schmidt-Hieber et al. (2013) and Proksch et al. (2018) for hypothesis testing in inverse problems, where multiscale dictionaries adapted to the operator T are employed. Another source of inspiration for our work are nonparametric methods that combine variational regularization techniques with multiscale dictionaries. Here we refer to Candès and Guo (2002), Dong et al. (2011) and Frick et al. (2012) for an empirical analysis of such methods in simulations, and to Nemirovski (1985) and Grasmair et al. (2018) for a theoretical analysis. Moreover, the proof of our main result is based on the above mentioned interpolation technique: an interpolation inequality of the form (1.16) is used to relate the risk functional, the regularization functional and the data-fidelity. This technique was used by Nemirovski (1985) and Grasmair et al. (2018) for estimating Sobolev functions, using an extension of the Gagliardo-Nirenberg interpolation inequalities (Nirenberg, 1959), and we use it here for the estimation of BV functions employing generalizations thereof (Meyer (2001), Cohen et al. (2003)).

The second main contribution of this thesis is the study of the minimax rates over Besov spaces $B^s_{p,t}$ with $s \leq d/p$, which determine the minimax rates over BV. This parameter regime has remained mainly ignored in the statistics literature, presumably due to the technical difficulties it presents. Only Goldenshluger and Lepskii (2014) and Lepskii (2015) have considered estimation in an anisotropic generalization of these spaces. Our results complement theirs and show that the minimax rates for regression and inverse problems behave differently than in the other better-known regimes.

Finally, in this thesis we also consider the efficient numerical computation of the estimator (1.5). The challenge of solving the minimization problem in (1.5) lies on the high dimensionality of the constraint set ($\#\Omega_n$ is typically larger than n), and on the non-smoothness of the objective function. An approach for solving this kind of optimization problems was proposed by Frick et al. (2012) and Li (2016). It uses an Alternating Direction Method of Multipliers (ADMM) approach that alternatively minimizes the objective and projects to the constraint set. The drawback of this approach is the projection step, which is typically extremely time consuming. Instead, in this thesis we propose two alternative approaches that circumvent the projection step and can be efficiently implemented: a primal-dual method based on the Chambolle-Pock algorithm (Chambolle and Pock, 2011), and a semismooth Newton method combined with the path-following technique (see e.g. Hintermüller (2010)). We discuss the implementation of these methods and illustrate their performance in simulations.

Organization of the thesis

In Chapter 2 we consider the regression problem ($T = id$): we introduce the main assumptions on the multiscale dictionaries, and state our main theorem. We also sketch the proof of the theorem, give concrete examples of dictionaries $\{\psi_\omega\}$, and discuss how to adapt our results to the nonparametric regression model. In Chapter 3 we consider linear inverse problems: we state our assumptions and main theorem, and illustrate the examples of deconvolution and Radon inversion explicitly. We also present a result concerning the minimax rates for regression and inverse problems over Besov spaces. In Chapter 4 we present different methods for solving the optimization problem (1.5) and discuss their implementation. In Chapter 5 we illustrate the performance of the multiscale TV-estimator in simulations in $d = 1$ and $d = 2$ for regression and deconvolution. We also compare the multiscale TV-estimator quantitatively with other estimation methods. In Chapter 6 we discuss our results and present open questions and extensions. The main proofs are given in Chapter 7, while some independent results from harmonic analysis are reproduced in Appendix A.

CHAPTER 2

Regression in the white noise model

In this chapter we consider nonparametric regression in a white noise model, i.e., the problem of estimating a function f from observations (1.1) with $T = id$. We present the main concepts needed to construct the multiscale TV-estimator (1.5), and the assumptions that guarantee that it is nearly minimax optimal over the set BV_L. We also give concrete examples of multiscale TV-estimators using particular dictionaries.

2.1 Basic definitions and notation

In this section we set some notation and give the definitions of mathematical objects that will appear throughout the thesis.

Basic notation

We denote the Euclidean norm of a vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ by $|v| := (v_1^2 + \cdots + v_d^2)^{1/2}$. The logarithm to the base $b > 1$ of a number $x > 0$ is written as $\log_b x$, while $\log x$ denotes the natural logarithm of x. For a real number x, define $\lfloor x \rfloor := \max\{m \in \mathbb{Z} \,\big|\, m \leq x\}$ and $\lceil x \rceil := \min\{m \in \mathbb{Z} \,\big|\, m > x\}$. The cardinality of a finite set X is denoted by $\#X$.
We say that two norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ in a normed space V are equivalent, and write $\|v\|_\alpha \asymp \|v\|_\beta$, if there are constants $c_1, c_2 > 0$ such that $c_1 \leq \|v\|_\beta / \|v\|_\alpha \leq c_2$ for all $v \in V$. The same notation is used to denote that two sequences a_n and b_n, $n \in \mathbb{N}$, grow at the same rate: we write $a_n \asymp b_n$ if there are constants $c_1, c_2 > 0$ such that $c_1 \leq \liminf a_n/b_n \leq \limsup a_n/b_n \leq c_2$. Moreover, we denote by C a generic positive constant that may change from line to line.
For a Borel-measurable set $\mathbb{M} \subseteq \mathbb{R}^d$, the space $L^2(\mathbb{M})$ consists of all equivalence classes of real-valued square integrable functions over $\mathbb{M}$ with respect to the Lebesgue measure on $\mathbb{R}^d$. It is a Hilbert space with the inner product

$$\langle g, h \rangle := \langle g, h \rangle_{L^2} := \int_{\mathbb{M}} g(x) h(x)\, dx, \quad g, h \in L^2(\mathbb{M}),$$

and its Hilbert space norm arises from this inner product. Whenever it is clear from the context, we will drop the symbols $\mathbb{M}$ or $\mathbb{R}^d$ from the notation of the function spaces, writing e.g. L^2 instead of $L^2(\mathbb{R}^d)$, etc.

Finally, $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with expectation $\mu \in \mathbb{R}$ and variance σ^2, for $\sigma > 0$.

Gaussian white noise process

In (1.1) we consider the Gaussian white noise process dW as a stochastic process over the Hilbert space $L^2(\mathbb{M})$. It is defined by its action on elements of L^2, given by

$$\begin{aligned} \langle g, dW \rangle &:= \int_{\mathbb{M}} g(x)\, dW(x) \sim \mathcal{N}(0, \|g\|_{L^2}^2), \\ \mathbb{E}[\langle g, dW\rangle \langle h, dW \rangle] &:= \langle g, h \rangle_{L^2}, \end{aligned}$$

for any $g, h \in L^2(\mathbb{M})$. We refer to Section 2.1 of Giné and Nickl (2015) for more details.

Functions of bounded variation over $\mathbb{R}^d$

For $k \in \mathbb{N}$, let $C^k(\mathbb{R}^d)$ denote the space of k-times continuously differentiable functions on $\mathbb{R}^d$. The space of functions of bounded variation BV consists of functions $g \in L^1$ whose weak distributional gradient $\nabla g = (\partial_{x_1} g, \cdots, \partial_{x_d} g)$ is a $\mathbb{R}^d$-valued finite Radon measure on $\mathbb{R}^d$. The finiteness implies that the bounded variation seminorm of g, defined as

$$|g|_{BV} := \sup\left\{ \int_{\mathbb{R}^d} g(x)\, \nabla \cdot h(x)\, dx \,\middle|\, h \in C^1(\mathbb{R}^d; \mathbb{R}^d),\ \|h\|_{L^\infty} \leq 1 \right\},$$

is finite. Here, $\nabla \cdot h := \sum_{i=1}^d \partial_{x_i} h_i$ denotes the divergence of the vector field $h = (h_1, \ldots, h_d)$. BV is a Banach space with the norm $\|g\|_{BV} = \|g\|_{L^1} + |g|_{BV}$ (see Evans and Gariepy (2015)). Here $C^1(\mathbb{R}^d; \mathbb{R}^d)$ denotes the set of continuously differentiable functions on $\mathbb{R}^d$ taking values on $\mathbb{R}^d$. By Lebesgue's decomposition theorem (see Section 1.6.2 in Evans and Gariepy (2015)), the weak gradient of a function of bounded variation can be decomposed as a Lebesgue-absolutely continuous measure, plus a Lebesgue singular measure. The singular measure is concentrated on sets of codimension one, and it represents jump discontinuities of the function.

Wavelet bases

For $S \in \mathbb{N}$, let $\{\psi_{j,k,e} \,|\, (j, k, e) \in \Lambda\}$ be an S-regular (see below) wavelet basis for $L^2(\mathbb{R}^d)$ whose elements are S times continuously differentiable with absolutely integrable S-th derivative. The wavelets are indexed by the set

$$\Lambda := \{(j,k,e) \,\big|\, j \geq 0,\ k \in \mathbb{Z}^d, e \in E_j\}, \tag{2.1}$$

$$E_j := \begin{cases} \{0,1\}^d & \text{if } j = 0, \\ \{0,1\}^d \backslash (0,\ldots,0) & \text{else.} \end{cases}$$

In particular, we consider wavelets of the form

$$\psi_{j,k,e}(x) = 2^{jd/2}\psi_e(2^j x - k),$$

where $\psi_e(z_1, \cdots, z_d) = \prod_{i=1}^d \psi_{e_i}(z_i)$ is the tensor product of one-dimensional wavelets, and

$$\psi_{e_i}(\cdot) = \begin{cases} \psi(\cdot) & \text{if } e_i = 1, \\ \varphi(\cdot) & \text{else,} \end{cases}$$

denotes either the mother wavelet ψ or the father wavelet φ of a wavelet basis of $L^2(\mathbb{R})$. The index $(0, \cdots, 0) \in E_0$ refers here to (shifts of) the father wavelet $\psi_{0,k,0} = \varphi(\cdot - k)$. See e.g. Section 4.2 in Giné and Nickl (2015) for the construction of such a basis.

S-regularity. The assumption of S-regularity ensures that the wavelets form a basis not only of L^2, but also of a range of Besov spaces. Even though we shall not need its precise form in this thesis, the definition of S-regularity is given for completeness in Appendix A.1.

Daubechies wavelets. Quite often in this thesis we will need S-regular wavelet bases whose elements have compact support. An example of such a basis are Daubechies wavelets, introduced by Daubechies (1992). We recall that one-dimensional Daubechies wavelets with D vanishing moments have support of size $2D-1$ (with respect to the Lebesgue measure) and are $\lfloor 0.18\cdot(D-1)\rfloor$ times continuously differentiable (see Theorem 4.2.10 in Giné and Nickl (2015)). An S-regular wavelet basis formed by tensorization of one-dimensional Daubechies wavelets needs to satisfy $D = 1 + 6S$ in order to have S continuous derivatives. Consequently, the mother and father wavelets have support of size $(12S + 1)^d$.

A subset of wavelets. In this thesis we will mainly deal with functions g supported inside the unit cube, supp $g \subseteq [0,1]^d$. We will use their wavelet expansion intensively, so for a basis of compactly supported wavelets, let us introduce the set of wavelets with nonzero overlap with the unit cube

$$\Omega = \{(j,k,e) \in \Lambda \,|\, \operatorname{supp} \psi_{j,k,e} \cap (0,1)^d \neq \emptyset\}. \tag{2.2}$$

In the following we will mostly work with the wavelets indexed by the set Ω. For each $n \in \mathbb{N}$,

$n \geq 2^d$, define the subset

$$\Omega_n := \{(j,k,e) \in \Omega \,\big|\, j = 0, \dots, J-1\}, \tag{2.3}$$

as the set of indices of wavelets at scales rougher than $J = \lfloor \frac{1}{d} \log_2 n \rfloor$. If we work with compactly supported Daubechies wavelets, which at scale $j = 0$ have support of size $(12S+1)^d$, we conclude that, for any $n \geq 2^d$,

$$2^{-d} n \leq \frac{\#\Omega_n}{(12S+1)^d} = 2^{Jd} \leq n.$$

Besov spaces

Let $\{\psi_{j,k,e} \,\big|\, (j,k,e) \in \Lambda\}$ denote an S-regular wavelet basis as defined above. For $p, q \in [1, \infty]$ and $s \in \mathbb{R}$ with $S > |s|$, the Besov norm of a (generalized) function is defined as

$$\|g\|_{B^s_{p,q}} := \Big(\sum_{j \in \mathbb{N}_0} 2^{jq\left(s + d(\frac{1}{2} - \frac{1}{p})\right)} \Big(\sum_{k \in \mathbb{Z}^d} \sum_{e \in E_j} |\langle \psi_{j,k,e}, g \rangle|^p \Big)^{q/p} \Big)^{1/q}, \tag{2.4}$$

with the usual modifications if $p = \infty$ or $q = \infty$.

If $s > 0$ and $p \in [1, \infty)$, the Besov space $B^s_{p,q}(\mathbb{R}^d)$ consists of L^p functions with finite Besov norm, while if $s > 0$ and $p = \infty$, then $B^s_{p,q}(\mathbb{R}^d)$ consists of continuous functions with finite Besov norm. In these cases, $\langle \psi_{j,k,e}, g \rangle$ denotes the coefficients of g with respect to the wavelet basis.

If $s \leq 0$, $B^s_{p,q}(\mathbb{R}^d)$ consists of temperate distributions $\mathcal{S}^*(\mathbb{R}^d)$ with finite Besov norm. Here, $\mathcal{S}^*(\mathbb{R}^d)$ denotes the space of temperate distributions, defined as the topological dual of the space $\mathcal{S}(\mathbb{R}^d)$ of Schwartz functions: infinitely differentiable functions $C^\infty(\mathbb{R}^d)$ whose derivatives decay at infinity faster than any polynomial (see Section A.2 in the Appendix). In that case, $\langle \psi_{j,k,e}, g \rangle$ is interpreted as the action of $g \in \mathcal{S}^*(\mathbb{R}^d)$ on the regular function $\psi_{j,k,e}$.

Fourier transform

The Fourier transform of a function $g \in L^1(\mathbb{R}^d)$ is defined as

$$\mathcal{F}[g](\xi) := \int_{\mathbb{R}^d} g(x)\, e^{-i\xi \cdot x}\, dx, \quad \xi \in \mathbb{R}^d,$$

and the inverse Fourier transform of $h \in L^1(\mathbb{R}^d)$ as

$$\mathcal{F}^{-1}[h](x) = \int_{\mathbb{R}^d} h(\xi)\, e^{i\xi \cdot x}\, d\xi/(2\pi)^d.$$

The Fourier transform can be extended as a bounded operator to L^2. Moreover, it maps Schwartz functions to Schwartz functions, and it can be extended by duality to temperate distributions $\mathcal{S}^*(\mathbb{R}^d)$ (see e.g. Section 4.1.1 in Giné and Nickl (2015)).

Dictionaries

In this thesis we will extensively use *dictionaries*: sets of functions that act as probe functionals. Unless otherwise stated, they will be denoted by

$$\Phi = \{\phi_\omega \,|\, \omega \in \Omega\},$$

where $\phi_\omega \in L^2(\mathbb{R}^d)$ are the elements of the dictionary, indexed by $\omega \in \Omega$, where Ω is a potentially countably infinite set. Examples of dictionaries include wavelet bases, but also frames (Christensen, 2003) and vaguelette systems (see Chapter 3). We will sometimes denote the dictionary elements by ψ_ω. In particular, the symbol ψ does *not* necessarily denote a wavelet.

2.2 Main results

The main ingredient of the multiscale TV-estimator (1.5) is the multiscale dictionary, on which we impose the following assumptions.

Assumption 1. Consider a dictionary $\Phi = \{\phi_\omega \,\big|\, \omega \in \Omega\} \subset L^2$ for a countable set Ω and functions satisfying $\|\phi_\omega\|_{L^2} = 1$ for all $\omega \in \Omega$. For each $n \in \mathbb{N}$, consider a subset $\Omega_n \subset \Omega$ of polynomial growth, meaning that

$$c\, n^\Gamma \leq \#\Omega_n \leq Q(n) \quad \text{for all } n \in \mathbb{N}$$

for a polynomial Q and constants $c, \Gamma > 0$. The sets Ω_n are assumed to satisfy the inequality

$$\|g\|_{B^{-d/2}_{\infty,\infty}} \leq C \max_{\omega \in \Omega_n} \big|\langle \phi_\omega, g\rangle\big| + C\, \|g\|_{L^\infty}\, n^{-1/2}$$

for any $g \in L^\infty$ and a constant $C > 0$ independent of n and g.

Examples 1.

a) The simplest example of a system Φ satisfying Assumption 1 is a sufficiently smooth wavelet basis. Indeed, the assumption follows from the characterization of Besov spaces in terms of S-regular wavelets with $S > \lceil d/2 \rceil$ (see Proposition 2 below).

b) Another example of a family Φ satisfying Assumption 1 is given by translations and rescalings of (the smooth approximation to) the indicator function of a cube. In Section 2.4

we verify the assumption for such a system, that has been used previously as a dictionary for function estimation (Grasmair et al., 2018).

c) In Section 2.4 we show that frames containing a smooth wavelet basis and a curvelet or a shearlet frame (which play a prominent role in imaging) satisfy Assumption 1.

Definition 1. Assume the model (1.1), and let Φ be a dictionary satisfying Assumption 1. We denote

$$\hat{f}_\Phi \in \operatorname*{argmin}_{g \in \mathcal{F}_n} |g|_{BV} \quad \text{subject to} \quad \max_{\omega \in \Omega_n} \left| \langle \phi_\omega, g \rangle - \langle \phi_\omega, dY \rangle \right| \leq \gamma_n, \tag{2.5}$$

as *multiscale TV-estimator* with respect to the dictionary Φ, where we minimize over the set

$$\mathcal{F}_n := \{ g \in BV \cap L^\infty \,\big|\, \|g\|_{L^\infty} \leq \log n, \ \operatorname{supp} g \subseteq [0,1]^d \}. \tag{2.6}$$

In (2.5) we use the convention that, whenever the "argmin" is taken over the empty set, $\hat{f}_\Phi$ is defined to be the constant zero function. ♣

The reason for requiring the support to be inside the closed unit cube in (2.6) is to make the set $\mathcal{F}_n$ closed. This is important for ensuring existence of a minimizer in (3.4) as the limit of a minimizing sequence (see Proposition 1 below).

In the following we assume that $n \geq 2$, so that we do not have to worry about the case $\log 1 = 0$. The reason for minimizing over the set $\mathcal{F}_n$ is that, in the analysis of the estimator $\hat{f}_\Phi$, we will need upper bounds on its supremum norm. As it turns out, the upper bound can be chosen to grow logarithmically in n without affecting the polynomial rate of convergence of the estimator (but yielding additional logarithmic factors in the risk). Alternatively, if we knew an upper bound L for the supremum norm of f, we could choose $\mathcal{F}_n = \{g \in BV \cap L^\infty \mid \|g\|_{L^\infty} \leq L, \ \operatorname{supp} g \subseteq [0,1]^d\}$. In that case, the risk bounds in Theorem 1 below would improve by some logarithmic factors (see Remark 1).

Proposition 1. In the setting of Definition 1, for each $n \in \mathbb{N}$ there exists almost always a minimizer $\hat{f}_\Phi \in BV \cap L^\infty$ of (2.5).

Proposition 1 guarantees that the multiscale TV-estimator as defined in (2.5) indeed exists. We give its proof in Section 7.3.1. We are now ready to state the convergence properties that the multiscale TV-estimator enjoys.

Theorem 1. Let $d \in \mathbb{N}$, and assume the model (1.1) with $f \in BV_L$ for some $L > 0$ and the set BV_L defined in (1.13). Let further $q \in [1, \infty)$.

a) Let γ_n be as in (1.9) with $\kappa > 1$, and let Φ be a family of functions satisfying Assumption 1. Then for any $n \in \mathbb{N}$ with $n \geq e^L$, the estimator $\hat{f}_\Phi$ in (2.5) with parameter γ_n satisfies

$$\sup_{f \in BV_L} \|\hat{f}_\Phi - f\|_{L^q} \leq C\, n^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}} (\log n)^{3-\min\{d,2\}} \tag{2.7}$$

with probability at least $1 - (\#\Omega_n)^{1-\kappa^2}$, for a constant $C > 0$ independent of n.

b) Under the assumptions of part a), if $\kappa^2 > 1 + \frac{1}{(d+2)\Gamma}$, then

$$\sup_{f \in BV_L} \mathbb{E}[\|\hat{f}_\Phi - f\|_{L^q}] \leq C\, n^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}} (\log n)^{3-\min\{d,2\}} \tag{2.8}$$

holds for n large enough and a constant $C > 0$ independent of n. The number $\Gamma > 0$ in the condition on κ^2 is the same as in Assumption 1.

Notice that part a) of the theorem implies that (2.7) holds asymptotically almost surely if $\kappa^2 > 2$.

Remark 1. The logarithmic factors in (2.7) and (2.8) are equal to $(\log n)^2$ for $d = 1$ and to $\log n$ for $d \geq 2$. They arise in part from the bound $\|\hat{f}_\Phi\|_{L^\infty} \leq \log n$ (that we get from minimizing over $\mathcal{F}_n$ in (2.6)), while some of them arise from the estimation procedure itself. Indeed, if we additionally constrain the estimator to $\|\hat{f}_\Phi\|_{L^\infty} \leq C$, the factors can be improved to $(\log n)^{1+\min\{\frac{1}{d+2}, \frac{1}{dq}\}}$ and $(\log n)^{\min\{\frac{1}{d+2}, \frac{1}{dq}\}}$ for $d = 1$ and $d \geq 2$, respectively. See Remark 3 below for an explanation of the different factors.

Remark 2. Recall that our parameter set BV_L in (1.13) involves a bound on the supremum norm. This bound can be relaxed to a bound on the Besov $B^0_{\infty,\infty}$ norm without changing the convergence rate $n^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}}$ for $\hat{f}_\Phi$. Indeed, assume for simplicity that Φ is an orthonormal wavelet basis of L^2, and for $n \in \mathbb{N}$ let $\Omega_n \subset \Omega$ index the wavelet coefficients with nonzero overlap with the unit cube up to level $J = \lfloor \frac{1}{d} \log_2 n \rfloor$, as in (2.3). As we will see below, the proof of Theorem 1 relies on an inequality of the form

$$\max_{(j,k,e) \in \Omega} |\langle \psi_{j,k,e}, g\rangle| \leq \max_{(j,k,e) \in \Omega_n} |\langle \psi_{j,k,e}, g\rangle| + C 2^{-Jd/2} \quad \forall J \in \mathbb{N} \tag{2.9}$$

for sufficiently smooth g with $\operatorname{supp} g \subseteq [0,1]^d$. But this inequality for all $J \in \mathbb{N}$ is equivalent to $\|g\|_{B^0_{\infty,\infty}} \leq C$ (see Jackson-type inequalities for Besov spaces, e.g. in Section 3.4 in Cohen (2003)). Consequently, Theorem 1 can be extended to show that the estimator $\hat{f}_\Phi$ with an orthonormal wavelet basis Φ attains the optimal polynomial rates of convergence uniformly over the enlarged parameter space

$$\widetilde{BV}_L := \{g \in BV \,\big|\, |g|_{BV} \leq L, \ \ \|g\|_{B^0_{\infty,\infty}} \leq L, \ \ \operatorname{supp} g \subseteq [0,1]^d\}.$$

One could ask whether the requirement $\|g\|_{B^0_{\infty,\infty}} \leq L$ can be relaxed further. This is not the case if $d \geq 2$. Indeed, since the embedding $B^1_{1,\infty} \subset B^0_{\infty,\infty}$ holds for $d = 1$ only (for functions supported on the unit cube; see the definition (2.4)), and since we have $BV \subset B^1_{1,\infty}$, we see that a typical function of bounded variation does not belong to $B^0_{\infty,\infty}$ if $d \geq 2$. Hence, the Jackson-type inequality in (2.9) cannot hold for general functions of bounded variation in $d \geq 2$. This explains why our parameter space is the intersection of a BV-ball with an L^∞-ball (or a $B^0_{\infty,\infty}$-ball). Finally, we remark that most works in function estimation deal with Hölder or Sobolev $W^{s,p}$ functions with $s > d/p$, so the assumption $f \in L^\infty$ is implicit. Alternatively, we refer to Section 3 in Lepskii et al. (1997) and to Delyon and Juditsky (1996) for examples of estimation over Besov bodies $B^s_{p,q}$ where uniform boundedness has to be assumed explicitly if $s < d/p$.

We can now state one of the main results of this thesis, which is a direct consequence of Theorem 1.

Theorem 2. Under the assumptions of Theorem 1, the estimator $\hat{f}_\Phi$ is asymptotically minimax optimal up to logarithmic factors over the parameter set BV_L defined in (1.13) with respect to the L^q-risk for $q \in [1, \infty)$ in any dimension $d \in \mathbb{N}$.

Proof. The claim follows from the fact that the minimax rate for estimation over the smaller class $(B^1_{1,1} \cap L^\infty)_L \subset BV_L$, defined in (1.17), satisfies

$$\mathcal{R}(L^q, (B^1_{1,1} \cap L^\infty)_L) \geq C_{L,\sigma}\, n^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}}$$

for $n \in \mathbb{N}$. This follows from Theorem 6 in Section 3.2 with $\beta = 0$, which states lower bounds for the minimax risk over Besov spaces. This rate matches the one in Theorem 1 up to the logarithmic factor, which implies that the multiscale TV-estimator is minimax optimal, up to the logarithm. □

We remark that the rate in Theorem 2 matches the result in Han et al. (2017) for estimation of bounded, component-wise isotone functions in the nonparametric regression model. Indeed, they show that the minimax rate with respect to the empirical L^2-risk scales as $n^{-\min\{\frac{1}{d+2}, \frac{1}{2d}\}}$, which equals the risk bound in Theorem 2 for $q = 2$. This is not entirely surprising, since bounded, component-wise isotone functions on a compact set have bounded variation. However, this correspondence is surprising in that it suggests that the class of bounded BV functions is statistically as complex as the class of bounded isotone functions. While this result is well-known in dimension $d = 1$, where any BV function there can be expressed as the difference of two monotone functions, no such result in $d \geq 2$ was known. Interestingly, Han et al. (2017) prove optimality of the slower rate $n^{-1/2d}$ by constructing a lower bound based on antichains in the set of isotone functions, while our proof of the lower bound is based on a construction of approximately dense linear combinations of wavelets.

2.3 Sketch of the proof of Theorem 1

We prove Theorem 1 in Section 7.1 as a corollary of Theorem 4, which proves convergence rates of a multiscale TV-estimator for inverse problems. In this section we give a sketch of the proof of Theorem 1. It relies on the following interpolation inequality proved by Cohen et al. (2003).

Theorem 3 (Theorem 1.5 in Cohen et al. (2003))**.** Let $s \in \mathbb{R}$ and $1 < p \leq \infty$, and assume that $\gamma := 1 + (s-1)p'/d$ satisfies either $\gamma > 1$ or $\gamma < 1 - 1/d$, where p' denotes the Hölder conjugate of p. Then for any $0 < \theta < 1$ such that

$$\frac{1}{q} = \frac{1-\theta}{p} + \theta, \quad t = (1-\theta)s + \theta$$

we have the inequality

$$\|g\|_{B^t_{q,q}} \leq C\, \|g\|_{B^s_{p,p}}^{1-\theta} \|g\|_{BV}^{\theta} \tag{2.10}$$

for any function $g \in BV \cap B^s_{p,p}(\mathbb{R}^d)$ and a universal constant $C > 0$.

The proof of part a) of Theorem 1 proceeds as follows.

1. For $n \in \mathbb{N}$, define the event

$$\mathcal{A}_n := \Big\{ \max_{\omega\in\Omega_n} \Big| \int_{\mathbb{R}^d} \phi_\omega(x)\, dW(x) \Big| \leq \frac{\sqrt{n}}{\sigma}\, \gamma_n \Big\}. \tag{2.11}$$

   This event represent the situation when the noise dW is "well-behaved". Indeed, $\mathcal{A}_n$ requires that the largest noise fluctuation in the projected data is smaller than $\sqrt{n}\,\sigma^{-1}\,\gamma_n$. Since $\int_{\mathbb{R}^d} \phi_\omega(x)\, dW(x) \sim \mathcal{N}(0,1)$ for $\|\phi_\omega\|_{L^2} = 1$, we have the bound

$$\mathbb{P}\Big(\max_{\omega\in\Omega_n} \Big| \int_{\mathbb{R}^d} \phi_\omega(x)\, dW(x) \Big| \geq t \Big) \leq \#\Omega_n\, e^{-t^2/2}, \tag{2.12}$$

   for any $n \in \mathbb{N}$ and $t \geq 0$. This bound follows from the union bound and elementary computations (see Proposition 11 in Section 7.3.2). By the choice of γ_n in (1.9), we conclude that

$$\mathbb{P}(\mathcal{A}_n) \geq 1 - (\#\Omega_n)^{1-\kappa^2},$$

which tends to one as $n \to \infty$ for $\kappa > 1$. To prove Theorem 1 we will show that (2.7) holds conditionally on the event $\mathcal{A}_n$.

2. Using a particular case of (2.10) with $s = -d/2$, $p = \infty$ and $t = 0$, and embeddings between L^q and Besov $B^0_{q,q}$ spaces (see Proposition 8 in Section 7.1), we bound the L^q-risk of $\hat{f}_\Phi$ as

$$\|\hat{f}_\Phi - f\|_{L^q} \leq C\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}}^{\frac{2}{d+2}}\|\hat{f}_\Phi - f\|_{BV}^{\frac{d}{d+2}} \tag{2.13}$$

for $d \geq 2$ and $q \leq 1 + 2/d$. A different strategy is needed for $d = 1$, see Remark 3. The rest of the proof consists in showing that the right-hand side behaves as $n^{-\frac{1}{d+2}} \log n$ conditionally on $\mathcal{A}_n$.

3. By Assumption 1, the Besov norm in (2.13) can be bounded as

$$\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}} \leq C \max_{\omega\in\Omega_n} \left|\langle\phi_\omega, \hat{f}_\Phi - f\rangle\right| + C\frac{\|\hat{f}_\Phi - f\|_{L^\infty}}{\sqrt{n}}. \tag{2.14}$$

The first term satisfies

$$\begin{aligned}\max_{\omega\in\Omega_n} \left|\langle\phi_\omega, \hat{f}_\Phi - f\rangle\right| &\leq \max_{\omega\in\Omega_n} \left|\langle\phi_\omega, \hat{f}_\Phi\rangle - \langle\phi_\omega, dY\rangle\right| + \max_{\omega\in\Omega_n} \left|\langle\phi_\omega, f\rangle - \langle\phi_\omega, dY\rangle\right| \\ &\leq \gamma_n + \frac{\sigma}{\sqrt{n}} \max_{\omega\in\Omega_n} \left|\int_{\mathbb{R}^d} \phi_\omega(x)\, dW(x)\right| \leq 2\gamma_n,\end{aligned}$$

where the second inequality follows by construction of $\hat{f}_\Phi$ and the third one holds conditionally on $\mathcal{A}_n$. The second term in (2.14) is bounded by $C\frac{\|\hat{f}_\Phi\|_{L^\infty}+\|f\|_{L^\infty}}{\sqrt{n}} \leq C\frac{L+\log n}{\sqrt{n}}$, since $\|\hat{f}_\Phi\|_{L^\infty} \leq \log n$ by construction and $\|f\|_{L^\infty} \leq L$ by $f \in BV_L$. Using the expression (1.9) for γ_n we have the bound

$$\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}} \leq C\sqrt{\frac{\log \#\Omega_n}{n}} + C\frac{L + \log n}{\sqrt{n}} \tag{2.15}$$

conditionally on $\mathcal{A}_n$.

4. The bounded variation norm in (2.13) satisfies

$$\begin{aligned}\|\hat{f}_\Phi - f\|_{BV} &= \|\hat{f}_\Phi - f\|_{L^1} + |\hat{f}_\Phi - f|_{BV} \\ &\leq \|\hat{f}_\Phi - f\|_{L^\infty} + |\hat{f}_\Phi|_{BV} + |f|_{BV},\end{aligned}$$

where we bound the L^1-norm by the L^∞-norm using the fact that $\operatorname{supp}(\hat{f}_\Phi - f) \subseteq [0,1]^d$. The supremum norm of the difference can then be bounded as in step 3. In order to bound

$|\hat{f}_\Phi|_{BV}$, notice that conditionally on $\mathcal{A}_n$ and for $n \geq e^L$ we have

$$\max_{\omega \in \Omega_n} \left| \langle \phi_\omega, f \rangle - \langle \phi_\omega, dY \rangle \right| \leq \gamma_n \quad \text{and} \quad \|f\|_{L^\infty} \leq L \leq \log n,$$

and hence the function f is feasible for the minimization problem (2.5) that defines $\hat{f}_\Phi$. Therefore we conclude that $|\hat{f}_\Phi|_{BV} \leq |f|_{BV}$ conditionally on $\mathcal{A}_n$, and we have

$$\|\hat{f}_\Phi - f\|_{BV} \leq \log n + L + 2|f|_{BV} \leq 3L + \log n,$$

where the second inequality follows from $f \in BV_L$.

5. Combining steps 3 and 4 with equation (2.13) yields the bound

$$\|\hat{f}_\Phi - f\|_{L^q} \leq C\Big(\sqrt{\frac{\log \#\Omega_n}{n}} + \frac{L + \log n}{\sqrt{n}} \Big)^{\frac{2}{d+2}} (3L + \log n)^{\frac{d}{d+2}}$$

conditionally on $\mathcal{A}_n$.

6. The previous argument gives the risk bound for $q \leq 1 + 2/d$. For $q \in [1 + 2/d, \infty)$, Hölder's inequality implies that

$$\|\hat{f}_\Phi - f\|_{L^q} \leq \|\hat{f}_\Phi - f\|_{L^{1+2/d}}^{\frac{d+2}{dq}} \|\hat{f}_\Phi - f\|_{L^\infty}^{1-\frac{d+2}{dq}} \leq C n^{-\frac{1}{dq}} \log n,$$

conditionally on $\mathcal{A}_n$.

Remark 3. While in this sketch we only considered the case $d \geq 2$, the proof for $d = 1$ is analogous, but somewhat more involved. Essentially, we use Theorem 3 to bound the $B^0_{3,3}$ risk by $O(n^{-1/(d+2)})$, and then show that for smooth enough functions $g \in L^\infty \cap BV$, the L^3-risk can be controlled by the $B^0_{3,3}$-risk at the cost of the $\log n$ factor. The difficulty here lies in the fact that the embedding $B^0_{3,3} \rightharpoonup L^3$ does not hold, so a refined analysis is needed. We then extend this risk bound to $q < 3$ using the compact support of the functions, and to $q > 3$ using Hölder's inequality.

2.4 Examples

We present now several dictionaries Φ that satisfy Assumption 1, and hence can be used to construct the multiscale TV-estimator with the minimax optimality guaranty given by Theorem 2.

Wavelet bases

For $S \in \mathbb{N}$, let $\Phi = \{\psi_{j,k,e} \,|\, (j,k,e) \in \Omega\}$ be a subset of an S-regular basis of Daubechies wavelets for $L^2(\mathbb{R}^d)$ as described in Section 2.1. Recall that the index set Ω denotes the indices such that $\operatorname{supp} \psi_{j,k,e} \cap (0,1)^d \neq \emptyset$. For $n \in \mathbb{N}$, $n \geq 2^d$, define the subset Ω_n as in (2.3), which satisfies $\#\Omega_n \asymp n$.

Proposition 2. An S-regular basis of Daubechies wavelets for L^2 as in Section 2.1 with $S > \max\{1, d/2\}$ satisfies Assumption 1 with the sets Ω_n in (2.3), a linear polynomial $Q(x) = c\,x$ and parameter $\Gamma = 1$.

The proof of Proposition 2.4 is given in Section 7.3.3. Proposition 2 implies that Φ satisfies Assumption 1, so by Theorem 2 the multiscale TV-estimator with dictionary Φ is minimax optimal up to logarithms for estimating BV functions in any dimension.

Remark 4 (Comparison with wavelet thresholding)**.** In dimension $d = 1$, Donoho and Johnstone (1998) proved that thresholding of the empirical wavelet coefficients of the observations gives an estimator that attains the minimax optimal convergence rate over BV. In contrast, our estimator combines a constraint on the wavelet coefficients with control on the BV-seminorm: this second aspect is crucial in higher dimensions. As equation (2.13) in the sketch of our proof illustrates, we bound the risk by the $B^{-d/2}_{\infty,\infty}$-norm of the residuals, which is the maximum of their wavelet coefficients, and the BV-norm of the residuals. The optimality of the estimator (2.5) depends crucially on the bound $\|\hat{f}_\Phi - f\|_{BV} \lesssim \log n$, which essentially amounts to a bound on the high frequencies of the residuals. But that is precisely the difficulty with wavelet thresholding of BV functions in higher dimensions. To the best of our knowledge, wavelet thresholding has been shown to converge over Besov spaces $B^s_{p,t}$ for $s > d(1/p - 1/q)_+$ only (see e.g. Delyon and Juditsky (1996)). This condition guaranties that the wavelet coefficients of the truth f decay fast enough, which itself allows one to control the high frequencies of the residuals. But that assumption is not satisfied for BV in dimension $d \geq 2$, since we have $B^1_{1,1} \subset BV$, which satisfies $1 > d/2$ for $d = 1$ only.

This remark matches the empirical observation that wavelet thresholding may present Gibbs-like artifacts, i.e., to present abnormally high frequencies. We verify this in simulations in Chapter 5. On the other hand, variational estimators with a suitable regularization functional automatically control the high frequencies.

General multiscale systems

In this example we present a more general multiscale system, and show that the corresponding multiscale TV-estimator is minimax optimal up to logarithms for estimating BV functions. Our motivation is to prove optimality of the estimator proposed by Frick et al. (2012), which has the form (1.5) for a multiscale dictionary consisting of indicator functions of cubes at different locations in different scales. That estimator was shown to perform well in denoising and deconvolution, which we verify in simulations in Chapter 5. Here, we prove optimality for a general family of estimators constructed with multiscale systems satisfying Assumption 2.

Assumption 2. The system of functions $\Phi = \{\psi_{j,k} \,|\, (j,k) \in \Omega_n,\ n \in \mathbb{N}\}$ satisfies the following conditions:

a) for each $n \in \mathbb{N}$ the set Ω_n is defined as

$$\Omega_n = \{(j,k) \,|\, j = 0, \ldots, J-1,\ k \in \mathcal{D}_j\},$$
$$\mathcal{D}_j = \{k = (k_1, \cdots, k_d) \,|\, k_i = -2^{-j} + l_i\, 2^{-R}(1 + 2^{1-j}),\ \ l_i = 0, \ldots, 2^R - 1,\ i = 1, \ldots, d\},$$

where $J = \lceil \frac{1}{d} \log_2 n \rceil$ and $R = \lfloor J \,\max\{1, d/2\} \rfloor$;

b) there is a function $\psi \in C^\infty(\mathbb{R}^d)$ with $\operatorname{supp}\psi \subseteq [0,1]^d$, satisfying

$$|\mathcal{F}[\psi](\xi)| > 0 \ \text{ for } \ |\xi| \leq 2, \ \ \|\psi\|_{L^2} = 1, \ \ \|\psi\|_{L^\infty} \leq 2,$$

such that all functions $\psi_{j,k} \in \Phi$ are given by translation, dilation and rescaling of ψ, i.e.,

$$\psi_{j,k}(z) := 2^{jd/2}\, \psi(2^j(k - z))$$

for $j \geq 0$ and $k \in \mathcal{D}_j$.

In words, the dictionary Φ contains functions at scales $j = 0, \ldots, J-1$ and, for each scale, it contains shifted versions of the same function by a distance 2^{-R} in each coordinate, where $R = \lfloor J \,\max\{1, d/2\} \rfloor$. This choice of R gives an increased spatial resolution as compared with wavelets, which would have $R = J$. The reason for choosing this R is that, unlike wavelets, the functions $\psi_{j,k}$ from Assumption 2 do not enjoy any special approximation property. This forces us to choose a very redundant system in order to achieve a good approximation.

Remark 5.

a) An example of a function ψ satisfying the above assumptions is the (L^2-normalized) convolution of the indicator function of the cube $[\frac{1}{4}, \frac{3}{4}]^d$ with the standard mollifier. More generally, the Fourier transform of the indicator function of the cube $[a, b] \subset [0, 1]^d$ satisfies $|\mathcal{F}[1_{[a,b]}](\xi)| > 0$ if $|\xi_i (b - a)_i| < 2\pi$ for all $i = 1, \dots, d$. In particular, taking ψ to be a smooth approximation to the indicator function of a cube, the estimator (2.5) is similar to that proposed by Frick et al. (2012).

b) For $n \in \mathbb{N}$ we have $\#\Omega_n = J\, 2^{dR} = J\, 2^{d\lfloor J \max\{1,d/2\}\rfloor}$, whence

$$n^{\max\{1,d/2\}} \leq \#\Omega_n \leq n^{\max\{1,d/2\}} \log n.$$

Proposition 3. Let $\Phi = \{\psi_{j,k} \,\big|\, (j, k) \in \Omega_n,\ n \in \mathbb{N}\}$ satisfy Assumption 2. Then it satisfies Assumption 1 with $Q(x) = x^{\max\{1,d/2\}+1}$ and $\Gamma = \max\{1, d/2\}$.

See Section 7.3.3 for the proof of Proposition 3. We remark that part of the proof is based on the characterizations of Besov spaces via local means (see Section A.3 in the Appendix).

Shearlet and curvelet frames

Another relevant example of the multiscale TV-estimator in $d \geq 2$ corresponds to the case when Φ contains a directional multiscale dictionary, e.g. a frame of shearlets or curvelets. An estimator of that form was proposed by Candès and Guo (2002), and it was shown to perform well in simulations. We verify its good numerical performance in Chapter 5. In this example we show how Theorem 2 implies minimax optimality up to logarithms for that estimator.

In order to state our results, we first review some facts about directional dictionaries. The first directional multiscale systems to be introduced were curvelets (Candès and Donoho, 2000). They were proposed as an improvements over wavelets in dimension $d \geq 2$: while wavelets are parametrized by a scale and a position parameter, curvelets have an additional "orientation" parameter. This allows them to resolve directional information such as boundaries better than wavelets do. Following curvelets, many directional multiscale systems have been proposed. We refer to Grohs et al. (2013) and references therein for a unifying mathematical framework for these dictionaries.

There are several constructions of directional multiscale dictionaries, mostly based on partitions of frequency space (Candès and Donoho, 2000). We just mention here the original curvelet system by Candès and Donoho (2000), shearlets (Labate et al., 2013), and compactly supported shearlets (Kutyniok et al., 2012). An important remark is that these directional dictionaries can be constructed to be tight frames of $L^2(\mathbb{R}^d)$, meaning that we have

$$||g||_{L^2}^2 = \sum_{\omega \in \Omega} |\langle \varphi_\omega, g \rangle|^2 \quad \forall g \in L^2(\mathbb{R}^d).$$

Furthermore, the directional elements φ_ω can be taken to have unit norm in L^2. Moreover, the constructions of tight curvelet frames in Borup and Nielsen (2007) and of shearlet frames in Labate et al. (2013) yield smooth frame elements that are exponentially decaying in space.
In this example, our dictionary Φ consists of a basis of S-regular Daubechies wavelets together with a directional multiscale system. Let us fix the notation

$$\begin{aligned} \Phi &= \Phi^W \cup \Phi^D \\ \text{wavelets: } \Phi^W &= \{\psi_{j,k,e} \,|\, (j,k,e) \in \Theta^W\} \\ \text{directional: } \Phi^D &= \{\varphi_{j\tilde{\theta}} \,|\, (j,\tilde{\theta}) \in \Theta^D\}. \end{aligned}$$

As in Section 2.1, the index set Θ^W indexes the wavelets with nonzero overlap with the unit cube. Similarly, the index set Θ^D indexes the directional elements $\varphi_{j,\tilde{\theta}}$ whose overlap with the unit cube is larger than a small predefined threshold. We neglect elements with a small overlap with the unit cube, since they do not carry much information about functions supported there, and they are hence not crucial for reconstruction purposes. We index the directional dictionary Φ^D with a scale index $j \in \mathbb{N}_0$ and a position and orientation index $\tilde{\theta} \in \widetilde{\Theta}_{n,j}$.
For $n \geq 2$, we define a finite subset of Φ as

$$\begin{aligned} &\Phi_n = \Phi_n^W \cup \Phi_n^D \\ &\Phi_n^W = \{\psi_{j,k,e} \,|\, (j,k,e) \in \Theta_n^W\}, \quad \Theta_n^W = \Omega_n \quad \text{in the notation of Section 2.1} \\ &\Phi_n^D = \{\varphi_{j\tilde{\theta}} \,|\, (j,\tilde{\theta}) \in \Theta_n^D\}, \qquad \Theta_n^D = \{(j,\tilde{\theta}) \,|\, j = 0,\ldots,\tilde{J}, \;\; \tilde{\theta} \in \widetilde{\Theta}_{n,j}\}, \end{aligned}$$

where $\Theta_n^D \subset \Theta^D$ is such that $\#\Theta_n^D \asymp n$.

Assumption 3. Let Φ be a mixed system of S-regular Daubechies wavelets and a directional dictionary as constructed above. Assume that $S > \max\{1, d/2\}$, and choose the sets Θ_n^W and Θ_n^D as indicated above, such that

$$\#\Theta_n^W + \#\Theta_n^D \asymp n$$

for any $n \in \mathbb{N}$.

Proposition 4. Let Φ satisfy Assumption 3. Then it satisfies Assumption 1 with $\Gamma = 1$ and $Q(x) = C\,x$, for a constant $C > 0$.

The proof of Proposition 4 is given in Section 7.3.3. A direct consequence of the proposition is that the multiscale TV-estimator with a mixed dictionary of wavelets and curvelets is minimax optimal up to logarithms for the reconstruction of BV functions.

Remark 6. The assumption that Φ contains a wavelet basis in addition to a directional frame is crucial. Indeed, the wavelet basis allows us to upper-bound the Besov norm $B^{-d/2}_{\infty,\infty}$ by the maximum over the frame coefficients with respect to Φ, which we need in order to establish Assumption 1. Alternatively, if Φ consisted of a curvelet frame only, the embeddings in Lemma 9 in Borup and Nielsen (2007) together with classical embeddings of Besov spaces (see Remark 4 of Section 3.5.4 in Schmeisser and Triebel (1987)) would give the bound

$$\|g\|_{B^{-d/2}_{\infty,\infty}} \leq C \max_{(j,\tilde{\theta})\in\Theta^D} 2^{j\delta}|\langle\varphi_{j,\tilde{\theta}}, g\rangle|$$

for smooth enough and compactly supported functions g, and a $\delta > 0$ that depends on the dimension. Accordingly, step 3 in the sketch of the proof of Theorem 1 would deteriorate to

$$\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}} \leq C\frac{n^{\delta'}}{\sqrt{n}}\text{Polylog}_{d,\delta'}(n)$$

for some $\delta' > 0$, and a polylogarithmic factor that diverges as $\delta' \to 0$. This results in a polynomially suboptimal rate of convergence. We remark that this limitation arises from the suboptimal embeddings between Besov spaces and decomposition spaces associated with the curvelet frame (see Lemma 9 in Borup and Nielsen (2007)). The situation for the shearlet frame is analogous, as its associated decomposition space equals that of the curvelet frame (see Proposition 4.4 in Labate et al. (2013)).

Exceptions

We close this section presenting some dictionaries Φ that do not satisfy Assumption 1, so that Theorem 1 does not apply to them.

a) Wavelet systems of low smoothness do not satisfy Assumption 1. Our result relies crucially on the fact that the Besov spaces $B^{-d/2}_{\infty,\infty}$ and $B^1_{1,1}$ can be characterized by the size of wavelet coefficients. For that, wavelet bases with $S-1$ vanishing moments and smoothness S are needed, where $S > \max\{1, d/2\}$ (see Section 4.3 in Giné and Nickl (2015)).

b) Recall the multiscale TV-estimator with a general multiscale system: there we considered a dictionary Φ consisting of *smooth* functions supported on cubes in $[0,1]^d$. The smoothness part is essential, since we need enough regularity in order to bound the Besov $B^{-d/2}_{\infty,\infty}$-norm in terms of this dictionary, which is done by the characterization of Besov spaces by local means (see Section A.3 of the Appendix). In fact, if the kernel ψ in Assumption 2 was e.g. a discontinuous function, then the dictionary Φ would not satisfy Assumption 1.

c) As argued in Remark 6, a dictionary consisting solely of a curvelet frame or a shearlet frame does not suffice, since the decomposition spaces they generate (in the sense of Borup and Nielsen (2007)) do not match Besov spaces exactly, so Assumption 1 does not hold.

2.5 Regression in a discretized model

Until now we have considered the regression problem in a white noise model (1.1). In that model, we observe the full path of function values plus white noise. There are alternative models where one can pose the regression problem. One such model is the *nonparametric regression model with deterministic design*. There, we observe the values of a function contaminated with noise at a deterministic grid of points, i.e.,

$$Y_i = f(x_i) + \sigma\,\epsilon_i, \quad x_i \in \Gamma_n, \;\; i = 1, \dots, n, \tag{2.16}$$

where we assume that $n = m^d$ for some $m \in \mathbb{N}$, and

$$\Gamma_n := \left\{ \left(\frac{k_1}{m}, \cdots, \frac{k_d}{m} \right) \middle| \, k_i \in \{1, \dots, m\}, \; i = 1, \dots, d \right\} \tag{2.17}$$

is the observation grid. Of course, different grids may be used. In (2.16), ϵ_i are independent standard normal random variables, and $\sigma > 0$ plays the role of the standard deviation of the noise. Of course, for (2.16) to make sense we have to assume that f is defined on the grid Γ_n, i.e. that $f(x_i) \in \mathbb{R}$ is well-defined for all $x_i \in \Gamma_n$.

We remark that, while the white noise model (1.1) is convenient from a theoretical perspective (as it avoids discretization issues), the nonparametric regression model (2.16) is sometimes more realistic to model applications, where one observes discretely sampled data. A prominent example is image processing, where the grid Γ_n represents pixels. We employ this discretization in our simulations in Chapter 5.

Given observations (2.16), our goal is to estimate the function f. In this section we explain how to adapt the multiscale TV-estimator to this setting, and analyze its convergence properties. For that, we have to discretize the construction from Section 2.2. Let $\Phi_n = \{\phi_\omega^n \,|\, \omega \in \Omega_n\}$ be a dictionary of discretized elements, i.e., each ϕ_ω^n is a vector of n values

$$(\phi_\omega^n)_i = n^{-1/2}\, \phi_\omega(x_i) \quad \text{for } i = 1, \dots, n,$$

which are the evaluations of ϕ_ω at the grid points. The scaling factor $n^{-1/2}$ is chosen so that

$$\sum_{x_i \in \Gamma_n} \left| (\phi_\omega^n)_i \right|^2 \to \|\phi_\omega\|_{L^2}^2 = 1 \quad \text{as} \quad n \to \infty,$$

for any $\omega \in \Omega_n$, i.e., so that the vectors ϕ_ω^n have roughly unit norm in an ℓ^2 sense.

In this setting, the multiscale TV-estimator takes the form

$$\hat{f}_D \in \underset{g\in\mathcal{F}_n}{\operatorname{argmin}} |g|_{BV} \text{ subject to } \max_{\omega\in\Omega_n} \Big| \sum_{x_i\in\Gamma_n} (\phi^n_\omega)_i (g(x_i) - Y_i)\Big| \leq \kappa\sigma \sqrt{2\log \#\Omega_n}.$$

Here we show that the estimator $\hat{f}_D$ is subject to a discretization error that, for $d \geq 3$, dominates the minimax rate $n^{-\min\{\frac{1}{d+2},\frac{1}{dq}\}}$ of the multiscale TV-estimator in the white noise model.

Indeed, we would like to apply the strategy of Section 2.3 to bound the risk of the estimator $\hat{f}_D$. For that, we have to relate the multiscale constraint to the Besov norm $B^{-d/2}_{\infty,\infty}$, as explained in step 3 of the sketch of the proof of Theorem 1. And for that, we need to show that the coefficients of the residuals $\hat{f}_D - f$ with respect to the discretized dictionary Φ_n are similar to the coefficients with respect to the "continuous" dictionary Φ. In that sense, the discretization error

$$\delta_n := \max_{\omega\in\Omega_n} \left| \frac{1}{\sqrt{n}} \sum_{x_i\in\Gamma_n} (\phi^n_\omega)_i\, g(x_i) - \int_{[0,1]^d} \phi_\omega(y) g(y)\, dy \right|$$

for $g = \hat{f}_D - f$ will give an additional error term for the estimator $\hat{f}_D$: in particular, equation (2.15) would now be

$$\|\hat{f}_D - f\|_{B^{-d/2}_{\infty,\infty}} \leq C\sqrt{\frac{\log \#\Omega_n}{n}} + C\frac{L + \log n}{\sqrt{n}} + \delta_n. \tag{2.18}$$

Hence, the discretization error is not relevant as long as $\delta_n = O(n^{-1/2})$, but it dominates the error otherwise. As it turns out, the discretization error behaves as $\delta_n = O(n^{-1/d})$, which means that it dominates for $d \geq 3$.

Proposition 5. Assume that there is an $\omega \in \Omega_n$ such that $\phi_\omega(x) = 1_{[0,1]^d}(x)$ is the indicator function of the unit cube. Then there exist functions $h \in BV \cap L^\infty$ satisfying

$$\max_{\omega\in\Omega_n} \left| \frac{1}{\sqrt{n}} \sum_{x_i\in\Gamma_n} (\phi^n_\omega)_i\, h(x_i) - \int_{[0,1]^d} \phi_\omega(y) h(y)\, dy \right| \geq \frac{1}{2} n^{-1/d}$$

for infinitely many $n \in \mathbb{N}$ of the form $n = m^d$, $m \in \mathbb{N}$.

The proof of Proposition 5 is given in Section 7.3.4. It is a constructive proof: a function h with a discontinuity at a position $x^{(1)} = \alpha$ is constructed, where $x^{(1)}$ denotes the first coordinate of a vector $x \in [0,1]^d$. We lower bound the difference by using the difficulty of approximating an irrational number α by rationals.

Proposition 5 gives just one example in which the discretization error δ_n is of order $n^{-1/d}$. This is enough to conclude that, in general, $\hat{f}_D$ cannot be expected to satisfy a bound better than (2.18). In other words, $\|\hat{f}_D - f\|_{B^{-d/2}_{\infty,\infty}} = O(\max\{n^{-1/d}, n^{-1/2}\})$ (with high probability) cannot be improved in general. Following the proof of Theorem 1, this implies that

$$\|\hat{f}_D - f\|_{L^q} \leq C\, n^{-\min\{\frac{1}{d+2},\frac{1}{dq}\}\min\{1,2/d\}} (\log n)^{3-\min\{d,2\}} \tag{2.19}$$

with high probability. Observe that, for $d \geq 3$, the multiscale TV-estimator attains a strictly slower rate in this discretized model than in the white noise model.

Remark 7 (Improved rate for smoother functions)**.** As argued above, the slower convergence rate in the nonparametric regression model is a consequence of the low smoothness of functions of bounded variation. Alternatively, if g were a $C^S(\mathbb{R}^d)$ function and $\{\phi_\omega\}$ were an S-regular wavelet basis, then we would have $\delta_n = O(n^{-S/d})$. This can be easily verified by Taylor expansion and using the vanishing moments of the wavelet basis. Consequently, if $S > d/2$, the discretization error δ_n would be of the order $n^{-1/2}$, and its convergence rate would be $n^{-\min\{\frac{1}{d+2},\frac{1}{dq}\}}$: the same as in the white noise model. This is consistent with known equivalence results between the white noise and the regression models (Reiß, 2008), that state that both problems are equivalent in Le Cam's sense, provided that the regression function belongs to $C^{d/2}(\mathbb{R}^d)$.

At this point, we could ask the question: can the slower rate in (2.19) be improved in the discrete model, or is it the minimax rate for estimating a function $f \in BV_L$ from discrete observations (2.16)? We do not know the answer to this question, but some evidence indicates that the rate might be improvable. Indeed, in the discrete regression model with the *empirical* ℓ^2 risk, Sadhanala et al. (2016) showed that the minimax rate for estimating BV functions is $n^{-\min\{\frac{1}{d+2},\frac{1}{2d}\}}$ up to logarithmic factors, which matches the minimax rate in the white noise model for $q = 2$. By empirical ℓ^2 error we mean the quantity

$$\|\hat{f} - f\|_{\ell^2} := \Big(\frac{1}{n}\sum_{i=1}^{n} \big|\hat{f}(x_i) - f(x_i)\big|^2\Big)^{1/2}. \tag{2.20}$$

We remark, however, that it makes a big difference to consider the risk with respect to the empirical ℓ^2 error and not to the continuous L^2 error. Indeed, in the discretized model we only observe point evaluations of the function of interest, and it is comparably easier to bound the ℓ^2 error at those observations than to interpolate and bound the L^2 error. This is specially relevant for BV functions, which due to their roughness are not well approximated by interpolation. We do not pursue this topic any further in this thesis.

CHAPTER 3

Inverse problems in the white noise model

In this chapter we extend the analysis from Chapter 2 to statistical inverse problems, i.e., to the case where the operator T in (1.1) is not the identity. The main difference to the case $T = id$ concerns the dictionary used to construct the multiscale TV-estimator. In particular, using a dictionary Φ that merely satisfies Assumption 1 will not perform well: if we did so, we would constrain our estimator to satisfy

$$\max_{\omega \in \Omega_n} \left| \langle \phi_\omega, T\hat{f}_\Phi \rangle - \langle \phi_\omega, dY \rangle \right| \leq \gamma_n,$$

i.e., we would require the coefficients of $T\hat{f}_\Phi$ to be close to the coefficients of Tf, up to noise. But due to the ill-posedness of the inverse problems, we have no guaranty that this implies that the coefficients of $\hat{f}_\Phi$ are close to the coefficients of f, i.e.,

$$\left| \langle \phi_\omega, T\hat{f}_\Phi - Tf \rangle \right| \text{ "small" } \not\Longrightarrow \left| \langle \widetilde{\phi}_\omega, \hat{f}_\Phi - f \rangle \right| \text{ "small".} \tag{3.1}$$

We are interested in estimating f, so we actually want an implication of the form (3.1), since that would allow us to estimate e.g. the wavelet coefficients of f reliably, which would then let us estimate f. A way to do so is to use the wavelet-vaguelette decomposition (WVD) of T, provided that it admits one. In this section we show how to use the WVD of T to construct a multiscale TV-estimator for inverse problems, and prove that the corresponding estimator is minimax optimal up to logarithmic terms for estimating BV functions in any dimension. We also present examples of operators T that have a WVD, such as the Radon transform or a convolution operator.

3.1 Main results

We make the following assumptions on the operator T.

Assumption 4. Let $T : \mathcal{D}(T) \subseteq L^2(\mathbb{R}^d) \to L^2(\mathbb{M})$ denote a bounded, linear operator. For $\beta \geq 0$, assume that the following hold:

- there is a dictionary $\Phi = \{\psi_{j,\theta} \,|\, (j,\theta) \in \Omega\} \subset L^2(\mathbb{R}^d)$ satisfying Assumption 1 in Section 2.2 with $\Gamma > 0$, where the inequality there is replaced by

$$\|g\|_{B^{-d/2-\beta}_{\infty,\infty}} \leq C \max_{(j,\theta)\in\Omega_n} 2^{-\beta j} \left|\langle \psi_{j,\theta}, g\rangle\right| + C\, \|g\|_{L^\infty}\, n^{-1/2}$$

  for any $g \in L^\infty$ with $\operatorname{supp} g \subseteq [0,1]^d$;

- there is a set of functions $\{u_{j,\theta} \,\big|\, (j,\theta) \in \Omega\} \subset L^2(\mathbb{M})$, which we call *vaguelette system*, such that

$$T^* u_{j,\theta} = \kappa_j\, \psi_{j,\theta} \quad \forall (j,\theta) \in \Omega, \tag{3.2}$$

  with generalized singular values $\kappa_j = 2^{-j\beta}$. Furthermore, the vaguelettes satisfy

$$c_1 \leq \|u_{j,\theta}\|_{L^2} \leq c_2 \quad \forall (j,\theta) \in \Omega$$

  for some real constants $c_2 \geq c_1 > 0$.

Remark 8.

a) Assumption 4 is slightly weaker than assuming that the operator T has a wavelet-vaguelette decomposition (WVD) (Donoho, 1995). In particular, in a "proper" WVD the dictionary ψ_ω would be a wavelet basis. We nevertheless call $\{u_{j,\theta}\}$ a vaguelette system for simplicity.

b) As remarked in Section 2.1, we will only need the dictionary elements ψ_ω with nonzero overlap with the unit cube, which we index by the set Ω. We index the vaguelettes accordingly.

c) We recover the WVD of an operator if we choose the dictionary Φ to be a basis of Daubechies wavelets (Daubechies, 1992) in $L^2(\mathbb{R}^d)$ with D continuous partial derivatives and whose mother wavelet has R vanishing moments, such that $\min\{R, D\} > \max\{1, d/2 + \beta\}$. The condition $\min\{R, D\} > \max\{1, d/2 + \beta\}$ is necessary for ensuring that the norms of the Besov spaces $B^{-d/2-\beta}_{\infty,\infty}$ and $B^1_{p,q}$, $p, q \in [1, \infty]$, can be expressed in terms of wavelet coefficients with respect to the wavelet basis $\{\psi_{j,\theta}\}$ (see Section 4.3 in Giné and Nickl (2015)).

d) Let $\{\psi_{j,\theta}\}$ be a smooth enough wavelet basis. Then condition (3.2) implies that the inverse problem (1.1) is mildly ill-posed with degree of ill-posedness β. In particular, in this thesis we only consider finitely smoothing operators. See the Conclusion in Chapter 6 for a discussion of how to extend our construction to exponentially ill-posed problems.

Examples 2. We list here some examples of operators satisfying Assumption 4. For simplicity, we assume that $\{\psi_{j,\theta}\}$ is a smooth enough wavelet basis.

a) The integration operator

$$Tg(x) := \int_{-\infty}^{x} g(y)\, dy, \quad x \in \mathbb{R}.$$

Its domain consists of functions g such that $|\xi|^{-1}\mathcal{F}[g](\xi) \in L^2(\mathbb{R})$, where $\mathcal{F}$ denotes the Fourier transform. The vaguelettes are given by derivatives and integrals of the wavelets $\psi_{j,k,e}$, and the singular values are $\kappa_j = 2^{-j}$. Fractional integration, iterated integration and higher dimensional integrals also define operators satisfying Assumption 4. We refer to Donoho (1995) for more details.

b) The Radon transform, which maps a function g to

$$Tg(r, \theta) := \int_{\{x \cdot \theta = r\}} g(x)\, dx, \quad r \in \mathbb{R}, \quad \theta \in S^{d-1}, \tag{3.3}$$

where the integral is taken over the hyperplane defined by vectors x satisfying $x \cdot \theta = r$. See Section 3.3 for more details on how to apply the multiscale TV-estimator to Radon data.

c) The convolution operator

$$Tg(x) := \int_{\mathbb{R}^d} K(x - y) g(y)\, dy$$

for a smooth enough kernel $K \in L^1(\mathbb{R}^d)$ satisfies Assumption 4. See Section 3.3 for more details.

d) The identity operator, in which case we are in the white noise regression model. In that case we have $u_{j,\theta} = \psi_{j,\theta}$, and the estimator (3.4) reduces to the multiscale TV-estimator analyzed in Chapter 2.

More generally, operators satisfying a certain homogeneity condition with respect to dilations have a WVD (see Donoho (1995) for a general result). Finally, we remark in line with Donoho (1995) that Assumption 4 is in general not satisfied for operators T with a strong preference for a particular scale. An extreme example is convolution with a kernel whose Fourier transform has compact support. In that case, the equation $T^* u_{j,k,e} = \kappa_j \psi_{j,k,e}$ does not admit solutions $u_{j,k,e}$ for compactly supported wavelets $\psi_{j,k,e}$.

In this setting, we define our estimator as follows.

Definition 2. Let the observations dY follow the model (1.1), and let the operator T satisfy Assumption 4 with a vaguelette system $\{u_{j,\theta}\}$. We denote

$$\hat{f}_{\Phi,T} \in \underset{g \in \mathcal{F}_n \cap \mathcal{D}(T)}{\operatorname{argmin}} |g|_{BV} \text{ subject to } \max_{\omega \in \Omega_n} \left| \langle u_\omega, Tg \rangle - \langle u_\omega, dY \rangle \right| \leq \gamma_n, \tag{3.4}$$

as the *multiscale total variation estimator* for the operator T. In (3.4) we minimize over the set $\mathcal{F}_n$ defined in (2.6), intersected with the domain of T. We use the convention that, whenever the feasible set of the problem (3.4) is empty (which happens with vanishingly small probability as n grows, see Remark 9), the estimator $\hat{f}_{\Phi,T}$ is set to zero. ♣

Concerning the choice of the threshold γ_n, let $\sigma > 0$ be as in (1.1), and let c_2 be the constant in Assumption 4. For a constant $\kappa > 0$ to be specified later, we choose

$$\gamma_n = \kappa \, c_2 \, \sigma \sqrt{\frac{2 \log \# \Omega_n}{n}}. \tag{3.5}$$

As for the estimator in Chapter 2, this threshold is chosen so that the true regression function f satisfies the constraint in (3.4) with high probability (see Remark 9 below).

Example 2. In this example we illustrate the role played by the dictionaries $\{\psi_{j,\theta}\}$ and $\{u_{j,\theta}\}$ in the estimator (3.4). Following the logic of the multiscale TV-estimator from Chapter 2, we require the coefficients of $\hat{f}_{\Phi,T}$ with respect to a dictionary $\{u_\omega\}$ to be close to the observed coefficients. Ignoring for simplicity the noise terms, the constraint in (3.4) is

$$\max_{\omega \in \Omega_n} \left| \langle u_\omega, T\hat{f}_{\Phi,T} - Tf \rangle \right| \leq \gamma_n,$$

where f denotes the true regression function. Consider the following possibilities:

a) If $\{u_\omega\}$ were a wavelet basis, then its good approximation properties would imply that $T\hat{f}_{\Phi,T}$ is close to Tf. This is however no guaranty that $\hat{f}_{\Phi,T}$ is close to f. Let for instance T denote convolution by a rapidly decaying kernel: it acts by locally blurring the details of f, so Tf does not preserve the small details (high frequencies) of f. Consequently, if $\{u_\omega\}$ is a wavelet basis, the constraint does not force $\hat{f}_{\Phi,T}$ to match f in the high frequencies, but it may still give a good reconstruction for the low frequencies. This phenomenon affects the MIND estimator (Grasmair et al., 2018), which is also a variational multiscale estimator. We recall it and illustrate it in simulations in Chapter 5.

b) If $\{u_\omega\}$ is a vaguelette system associated with a wavelet basis $\{\psi_\omega\}$, the situation is more favorable. Again ignoring noise terms, the constraint on the estimator $\hat{f}_{\Phi,T}$ is

$$\max_{\omega\in\Omega_n} \left|\langle u_\omega, T\hat{f}_{\Phi,T} - Tf\rangle\right| = \max_{(j,\theta)\in\Omega_n} 2^{-\beta j} \left|\langle \psi_{j,\theta}, \hat{f}_{\Phi,T} - f\rangle\right| \leq \gamma_n$$

for singular values $\kappa_j = 2^{-\beta j}$. This constraint hence imposes similarity between $\hat{f}_{\Phi,T}$ and f directly in terms of their wavelet coefficients: this is good, since wavelets have strong approximation properties. Indeed, as in Chapter 2, we enforce similarity between $\hat{f}_{\Phi,T}$ and f at all scales simultaneously. There is however a crucial difference: the weight $2^{-\beta j}$ implies that our constraint becomes less strict for smaller scales (large j). We have illustrated the reason for this in the previous paragraph for a convolution operator: the high frequencies (small scales) of Tf are highly attenuated, so the high frequencies of our observations carry relatively little information about the high frequencies of f. Exactly how much information they carry is characterized by the degree of ill-posedness β and the factor $2^{-\beta j}$. Hence, using a vaguelette system $\{u_\omega\}$ allows the estimator (3.4) to extract as much information as possible about the high frequencies of f.

The performance of the estimators presented in points a) and b) is illustrated in simulations in Chapter 5, where we see the different levels of detail achieved by each of them.

Remark 9. Let us discuss the feasible set of the problem (3.4), which consists of the constraints

$$\max_{\omega\in\Omega_n} \left|\langle u_\omega, Tg\rangle - \langle u_\omega, dY\rangle\right| \leq \gamma_n, \quad \|g\|_{L^\infty} \leq \log n, \quad \operatorname{supp} g \subseteq [0,1]^d. \tag{3.6}$$

By Proposition 11 in Section 7.3.2 and the choice (3.5) for γ_n, the probability that the true regression function f satisfies the first constraint in (3.6) is not smaller than $1 - O((\#\Omega_n)^{1-\kappa^2})$. As long as $n \geq e^L$ and f satisfies the first constraint in (3.6), it also satisfies the others, since we assume that $f \in BV_L$. As a consequence, the feasible set of (3.4) is nonempty with probability of the order $1 - O((\#\Omega_n)^{1-\kappa^2})$. Hence, we will see that the caveat in Definition 2 about the feasible set does not play a decisive role for the convergence properties of $\hat{f}_{\Phi,T}$.

Proposition 6. In the setting of Definition 2, for each $n \in \mathbb{N}$ there exists almost surely a minimizer $\hat{f}_{\Phi,T} \in BV \cap L^\infty$ of (3.4).

The proof of Proposition 6 is given in Section 7.3.1. For given β, d and q, recall the definition of the exponent

$$\vartheta_{q,\beta} := \begin{cases} \frac{1}{d+2\beta+2} & \text{for } q < 1 + 2/(d+2\beta) \\ \frac{1}{q\,(d+2\beta)} & \text{for } q \geq 1 + 2/(d+2\beta). \end{cases} \tag{3.7}$$

Theorem 4. For $d \in \mathbb{N}$, let T satisfy Assumption 4 with $\beta \geq 0$. Assume the model (1.1) with $f \in BV_L$ for some $L > 0$. For $q \in [1, \infty)$, let $\vartheta_{q,\beta}$ be as in (3.7).

a) Let γ_n be as in (3.5) with $\kappa > 1$. Then for any $n \in \mathbb{N}$ with $n \geq e^L$, the estimator $\hat{f}_{\Phi,T}$ in (3.4) with parameter γ_n satisfies

$$\sup_{f \in BV_L} \|\hat{f}_{\Phi,T} - f\|_{L^q} \leq C\, n^{-\vartheta_{q,\beta}} (\log n)^{3-\min\{d,2\}} \tag{3.8}$$

with probability at least $1 - (\#\Omega_n)^{1-\kappa^2}$, for a constant $C > 0$ independent of n.

b) Under the assumptions of part a), if $\kappa^2 > 1 + \frac{1}{(d+2\beta+2)\Gamma}$, then

$$\sup_{f \in BV_L} \mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}] \leq C\, n^{-\vartheta_{q,\beta}} (\log n)^{3-\min\{d,2\}} \tag{3.9}$$

holds for n large enough and a constant $C > 0$ independent of n. The constant $\Gamma > 0$ in the condition on κ^2 is the one from Assumption 4.

The proof of Theorem 4 is given in Section 7.1. We have the following consequence of Theorem 4.

Theorem 5. Consider the setting of Theorem 4, and assume further that the operator T satisfies condition (3.11) below. Then the estimator (3.4) is asymptotically minimax optimal up to logarithmic factors for estimating functions $f \in BV_L$, $L > 0$, with respect to the L^q-risk, for any $q \in [1, \infty)$.

Proof. As in the regression setting, we show that the minimax risk over the smaller class $(B^1_{1,1} \cap L^\infty)_L \subset BV_L$ is lower bounded by $n^{-\vartheta_{q,\beta}}$. This matches the rate of convergence of the multiscale TV-estimator up to logarithmic factors, which gives the claim. And indeed, according to Theorem 6, the minimax rate of estimation in the inverse problem setting (1.1) over the class $(B^1_{1,1} \cap L^\infty)_L$ satisfies

$$\mathcal{R}(L^q, (B^1_{1,1} \cap L^\infty)_L) \geq C_{L,\sigma}\, n^{-\vartheta_{q,\beta}}$$

which completes the proof. □

Remark 10. In the same way as the multiscale TV-estimator for regression can be seen as a hybrid between wavelet thresholding and variational regularization, the multiscale TV-estimator for inverse problems is a mixture of wavelet-vaguelette thresholding and variational regularization. This analogy raises the question of how well thresholding of the WVD performs for estimating BV functions. This was answered by Donoho (1995), who proved that thresholding of the WVD is minimax optimal over a range of Besov spaces. His results cover the case of BV functions for $d = 1$ and β-smoothing operators with $\beta \in [0, 1/2)$. This is, to the best of our knowledge, the only available result for minimax optimal reconstructions of BV functions in inverse problems. In this sense, our result is an improvement in that the estimator (3.4) is nearly minimax optimal in any dimension $d \geq 1$ and for all $\beta \geq 0$.

Sketch of the proof of Theorem 4

The proof of Theorem 4 follows roughly the same ideas as that of Theorem 1, sketched in Section 2.3. The main differences concern the wavelet and vaguelette dictionaries. In this section we discuss how to deal with them.

1. Recall that in the regression setting in Section 2.3 we work conditionally on the event $\mathcal{A}_n$ in (2.11), which guaranties that the observational noise is not too large. In our present setting, the estimator $\hat{f}_{\Phi,T}$ is based on the projection of dY onto the vaguelette system $u_{j,\theta}$. We hence need to guarantee that the noise corrupting these observations is suitably small. The exact condition that we need is encoded in the event

$$\widetilde{\mathcal{A}}_n := \left\{ \max_{(j,\theta)\in\Omega_n} \left| \int_{\mathbb{M}} u_{j,\theta}(x)\, dW(x) \right| \leq \frac{\sqrt{n}}{\sigma}\, \gamma_n \right\}. \tag{3.10}$$

   As in the sketch of Theorem 1, our strategy is to show that $\hat{f}_{\Phi,T}$ converges at the optimal rate conditionally on the event $\widetilde{\mathcal{A}}_n$. Further, we show that this event happens with probability approaching 1 as $n \to \infty$

2. In order to bound the L^q-risk, we also use here an interpolation inequality derived from Theorem 3. However, for reasons to become clear soon, we need to relate it to the BV and Besov $B^{-d/2-\beta}_{\infty,\infty}$ norms, i.e.,

$$\|\hat{f}_{\Phi,T} - f\|_{L^q} \leq C\, \|\hat{f}_{\Phi,T} - f\|_{B^{-d/2-\beta}_{\infty,\infty}}^{\frac{2}{d+2\beta+2}} \|\hat{f}_{\Phi,T} - f\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}} \quad \text{for } d \geq 2 \text{ and } q \in [1, 1 + 2/(d+2\beta)].$$

   While the term $\|\hat{f}_{\Phi,T} - f\|_{BV}$ can be bounded as in Section 2.3, the term $\|\hat{f}_{\Phi,T} - f\|_{B^{-d/2-\beta}_{\infty,\infty}}$ requires a special analysis, which we sketch now. First, since the dictionary Φ satisfies Assumption 4, we can bound the Besov norm as

$$\|\hat{f}_{\Phi,T} - f\|_{B^{-d/2-\beta}_{\infty,\infty}} \leq \max_{(j,\theta)\in\Omega_n} 2^{-\beta j} |\langle \psi_{j,\theta}, \hat{f}_{\Phi,T} - f\rangle| + C\, \|\hat{f}_{\Phi,T} - f\|_{L^\infty}\, n^{-1/2}.$$

   The second term can be handled as in Section 2.3. For the first term, we have

$$\begin{aligned} \max_{(j,\theta)\in\Omega_n} 2^{-\beta j} |\langle \psi_{j,\theta}, \hat{f}_{\Phi,T} - f\rangle| &= \max_{(j,\theta)\in\Omega_n} 2^{-\beta j} |\kappa_j^{-1} \langle T^* u_{j,\theta}, \hat{f}_{\Phi,T} - f\rangle| \\ &\leq \max_{(j,\theta)\in\Omega_n} |\langle u_{j,\theta}, T\hat{f}_{\Phi,T}\rangle - \langle u_{j,\theta}, dY\rangle| + \frac{\sigma}{\sqrt{n}} \max_{(j,\theta)\in\Omega_n} |\langle u_{j,\theta}, dW\rangle|, \end{aligned}$$

using the definition of vaguelettes and $\kappa_j = 2^{-\beta j}$. The first term in the right-hand side is bounded by γ_n by construction of the estimator $\hat{f}_{\Phi,T}$ in (3.4), while the second term is bounded by γ_n conditionally on the event $\widetilde{\mathcal{A}}_n$. Plugging the value (3.5) of γ_n, we get altogether the bound

$$\|\hat{f}_{\Phi,T} - f\|_{B^{-d/2-\beta}_{\infty,\infty}} \leq C\, n^{-1/2} \log n$$

conditionally on $\widetilde{\mathcal{A}}_n$. Inserting this in the interpolation inequality gives part a) of Theorem 4 for $q \leq 1 + 2/(d + 2\beta)$.

3. For $q > 1 + 2/(d + 2\beta)$, we use Hölder's inequality as in the sketch of the proof of Theorem 1. Finally, the claim in part b) of convergence in expectation follows easily from that.

Remark 11. We have sketched the proof for $d \geq 2$. As in the proof of Theorem 1, the case $d = 1$ requires a slightly different treatment. We refer to the proof in Section 7 for the details.

3.2 Minimax lower bounds

Here we prove a lower bound for the minimax risk over Besov spaces $B^s_{p,t}$, $s > 0$, $p, t \in [1, \infty]$, for observations (1.1) from a β-smoothing operator, with respect to the L^q-risk. By the embedding $B^1_{1,1} \subset BV$ this provides a lower bound on the minimax risk over BV.

In order to state the result, we make the following assumption. For $\beta \geq 0$, the linear operator $T : L^2(\mathbb{R}^d) \to L^2(\mathbb{M})$ satisfies

$$\|T\psi_{j,k,e}\|_{L^2} \leq c\, 2^{-j\beta} \quad \forall (j,k,e) \in \Omega \tag{3.11}$$

for a constant $c > 0$, where $\{\psi_{j,k,e}\}$ is a wavelet basis of compactly supported wavelets. We remark that any operator T that admits a WVD satisfies this condition (Donoho, 1995). Notice that the case where $T = id$ with $\beta = 0$ is allowed, and gives a lower bound for the minimax rates in a regression setting. For inverse problems, the Radon transform and a convolution operator with suitable kernel satisfy (3.11).

Theorem 6. Let $1 \leq p, t \leq \infty$, $1 \leq q < \infty$, $s > 0$ and $L > 0$. Let T satisfy (3.11) for $\beta \geq 0$. Then there is a constant $C_{L,\sigma} > 0$ such that the minimax risk over the set $(B^s_{p,t} \cap L^\infty)_L$ in (1.17) for observations from the inverse problem (1.1) satisfies

$$\mathcal{R}(L^q, (B^s_{p,t} \cap L^\infty)_L) \geq C_{L,\sigma}\, r_n(s,p,t,q)$$

for all $n \in \mathbb{N}$, where

$$r_n(s,p,t,q) = \begin{cases} n^{-\frac{s}{2s+2\beta+d}} & \text{if } q < p\frac{d+2s+2\beta}{d+2\beta} \\ \left(\frac{\log n}{n}\right)^{\frac{s+d(\frac{1}{q}-\frac{1}{p})}{2s+2\beta+2d(\frac{1}{2}-\frac{1}{p})}} & \text{if } q \geq p\frac{d+2s+2\beta}{d+2\beta} \text{ and } s > d/p \\ n^{-\frac{sp}{(d+2\beta)q}} & \text{if } q \geq p\frac{d+2s+2\beta}{d+2\beta} \text{ and } s \leq d/p. \end{cases} \tag{3.12}$$

As stated in Theorem 5 above, the first consequence of Theorem 6 is that the multiscale TV-estimator is minimax optimal over BV_L up to logarithmic factors for any $q \in [1, \infty)$ and any $\beta \geq 0$.

More generally, Theorem 6 gives insight about the difficulty of estimating quite general functions. The parameter regimes $q < p\frac{d+2s+2\beta}{d+2\beta}$ (dense case) and $q \geq p\frac{d+2s+2\beta}{d+2\beta}$ and $s > d/p$ (sparse case) are well understood, and the associated minimax rates have been known for a while for Besov spaces if $T = id$ (see Chapter 10 in Härdle et al. (2012)), and for Sobolev spaces with $p = 2$ for some inverse problems (see Cavalier and Tsybakov (2002)). Their proofs follow the classical strategy of constructing a set of alternatives in $(B^s_{p,t} \cap L^\infty)_L$ that are well separated in the L^q-norm, and applying an information inequality (e.g. Fano's inequality).

On the other hand, the regime $q \geq p\frac{d+2s+2\beta}{d+2\beta}$ and $s \leq d/p$ is far less popular. For regression ($\beta = 0$), this regime was observed by Goldenshluger and Lepskii (2014) and Lepskii (2015) for anisotropic Nikolslkii classes, which in the isotropic case correspond to $B^s_{p,\infty}$, and in general allows for different smoothness and integrability indices in different spatial directions.

Regarding the boundaries between regimes, Donoho et al. (1997) showed that at the boundary $q = p\frac{d+2s}{d}$, $s > d/p$, the lower bound can be tightened by an additional logarithmic factor. At that boundary for $s < d/p$ and at the boundary $s = d/p$ we do not know whether the bounds can be tightened, since the only estimators known to converge there (the one in Lepskii (2015) and in the present thesis) attain the lower bound up to logarithmic factors.

Notice that for $s < d/p$, functions in $B^s_{p,t}$ are not continuous. The presence of discontinuities then precludes consistency in the L^∞-risk, a phenomenon that was stressed by Lepskii (2015) and that is well-known in dimension $d = 1$ for change-point estimation (Li et al., 2017). We remark that the L^∞ inconsistency is responsible for the slower rate in Theorem 6. To see that, consider the opposite case, i.e., $s > d/p$. It is well-known since Nemirovski's work (Nemirovski,

1985) that the minimax risk over Sobolev $W^{s,p}$ spaces w.r.t. the L^q-risk is determined by its values at $q = p(1 + 2s/d)$ and at $q = \infty$. For a suitable estimator (e.g. the window estimator in Nemirovski (1985)) one has the bounds

$$\|\hat{f} - f\|_{L^{p(1+2s/d)}} \leq C\,(n^{-1}\ \log n)^{\frac{s}{2s+d}}$$
$$\|\hat{f} - f\|_{L^\infty} \leq C\,(n^{-1}\ \log n)^{\frac{s-d/p}{2s+d-2d/p}}$$

with high probability. The L^q-risk for all other $q \in [1, \infty]$ follows by domination and interpolation, i.e.,

$$\|\hat{f} - f\|_{L^q} \leq \begin{cases} \|\hat{f} - f\|_{L^{p(1+2s/d)}} \leq C\,(n^{-1}\ \log n)^{\frac{s}{2s+d}} & \text{if } q \leq p(1 + 2s/d) \\ \|\hat{f} - f\|_{L^{p(1+2s/d)}}^{\frac{p(2s+d)}{qd}} \|\hat{f} - f\|_{L^\infty}^{1-\frac{p(2s+d)}{qd}} \leq C\,(n^{-1}\ \log n)^{\frac{s+d/q-d/p}{2s+d-2d/p}} & \text{else.} \end{cases}$$

This argument holds for $s > d/p$, since then we can guarantee that the L^∞ risk tends to zero. However, for $s \leq d/p$ the L^∞ risk does not tend to zero, whence the risk bound for $q \geq p(1+2s/d)$ deteriorates to

$$\|\hat{f} - f\|_{L^q} \leq \|\hat{f} - f\|_{L^{p(1+2s/d)}}^{\frac{p(2s+d)}{qd}} \|\hat{f} - f\|_{L^\infty}^{1-\frac{p(2s+d)}{qd}} \leq C\,(n^{-1}\ \log n)^{\frac{sp}{dq}},$$

which matches the lower bound in Theorem 6.

Remark 12 ("Ideal" risk)**.** A practical implication of this result is that, if $s \leq d/p$, using the L^q-risk for large q comes at the cost of a slower convergence. On the other hand, in many applications one wants to take q as large as possible. Our result suggests that the *ideal* q is given by $q = p(1 + 2/(d + 2\beta))$, as it is the largest index that still achieves the "fast" rate $n^{-1/(d+2\beta+2)}$. In particular, for BV functions and $\beta = 0$, this implies that the $q = 1 + 2/d$ risk is better suited than, say, the L^2-risk for $d \geq 3$.

Remark 13 (Multiscale alternatives)**.** The proof of the lower bound on the minimax rate in the regime $s \leq d/p$, which we give in Section 7.2 is based on the classical reduction to testing: we construct a set of alternatives separated by a distance δ and show that no statistical testing procedure can distinguish them perfectly. This construction is then "inverted", and implies that any estimation procedure makes an expected error of at least δ. The largest possible lower bound is then achieved by looking for the largest distance δ as a function of n such that no testing procedure can distinguish the alternatives perfectly (see e.g. Tsybakov (2009) for more details). As in the dense regime, our construction of the alternatives is based on Assouad's cube applied to a wavelet basis $\{\psi_{j,k,e}\}$ (Assouad, 1983). For simplicity of the notation, we sketch the construction here for $\beta = 0$.

Recall that in the dense regime ($q < p(1 + 2/d)$), the set of alternatives that determines the lower bound is

$$\{g_\epsilon = g_0 + \gamma \sum_{(k,e)\in P_j^d \times E_j} \epsilon_{k,e}\psi_{j,k,e} \,\big|\, \epsilon_{k,e} \in \{-1, +1\}, \ \ (k, e) \in \{0, \dots, 2^j - 1\}^d \times E_j\},$$

where $\gamma > 0$ parametrizes the "signal strength" of the alternatives. This regime is called *dense* because the difficulty of estimating is driven by functions supported everywhere. On the other hand, the minimax lower bound in the sparse regime ($q \geq p(1 + 2/d)$, $s > d/p$) is determined by the set of alternatives

$$\{g_{k,e} = g_0 + \gamma \psi_{j,k,e} \,\big|\, (k, e) \in \{0, \dots, 2^j - 1\}^d \times E_j\}.$$

In other words, the functions that are most difficult to estimate in this regime are localized spikes. Finally, in the multiscale regime ($q \geq p(1 + 2/d)$, $s \leq d/p$) the minimax lower bound is driven by alternatives of the form

$$\{g_\epsilon = g_0 + \gamma \sum_{(k,e)\in R_j} \epsilon_{k,e}\psi_{j,k,e} \,\big|\, \epsilon_{k,e} \in \{-1, +1\}, \ \ (k, e) \in R_j\},$$

for a set $R_J \subset \{0, \dots, 2^j - 1\}^d \times E_j$ of cardinality $\#R_j = \lfloor 2^{j(d-sp)} \rfloor$. These functions distribute their mass among $\#R_j$ spikes whose locations can vary. Interestingly, this suggests that multiscale estimators like (2.5), which enforce a local fitting at different locations and scales, may be optimal in this regime. In the present work we have verified this for BV functions, and in Lepskii (2015) a kernel estimator with spatially varying bandwidth was shown to be optimal over Nikolskii classes in this regime.

3.3 Examples

For the following examples we choose the dictionary $\Phi = \{\psi_{j,k,e}\}$ to consist of sufficiently regular Daubechies wavelets. The precise regularity depends on the ill-posedness of the operator T: for a β-smoothing operator (i.e. with singular values $\kappa_j = 2^{-j\beta}$), we choose Φ to be an S-regular basis with S times continuously differentiable elements, where $S > \max\{1, d/2 + \beta\}$.

Radon transform

Due to its application in nondestructive imaging, in particular in medial applications, tomography is a very relevant inverse problem. While there are plenty of mathematical models for tomography, which mainly depend on the type of tomography and the geometry of the detector (see e.g. Chapter

1 in Scherzer et al. (2009)), in this section we will exemplarily consider tomography modeled by the Radon transform. For simplicity we consider here the two dimensional case, in which the Radon transform of a function g is given by its line integrals along different directions, see (3.3).

Functions in the range of T are supported on cylindrical sets of the form $\mathbb{M} = \mathbb{R} \times [0, 2\pi)$. Moreover, the domain of T consists of functions $g \in L^2(\mathbb{R}^2)$ whose Fourier transform satisfies $|\xi|^{-1/2}\mathcal{F}[g](\xi) \in L^2$. This is a condition on the low frequencies which essentially ensures that local averages remain reasonably small.

In this section we will show how to apply the estimation framework developed above to this type of inverse problems. For that, let $\{\psi_{j,k,e}\}$ denote a basis of Daubechies wavelets as described in Section 2.1. For $(j, k, e) \in \Omega$, the vaguelettes are functions on the radial and angular coordinates (r, θ) defined by

$$u_{j,k,e}(r, \theta) = \frac{2^{-j/2}}{(2\pi)^2} \int_{\mathbb{R}} |\rho| \, \mathcal{F}[\psi_{j,k,e}](\rho \cos\theta, \rho \sin\theta) \, e^{ir\rho} \, d\rho. \tag{3.13}$$

It is easy to verify directly (see e.g. Chapter 2 in Natterer (1986)) that the vaguelettes satisfy the equation

$$T^* u_{j,k,e} = \kappa_j \, \psi_{j,k,e}$$

for singular values $\kappa_j = 2^{-j/2}$. Moreover,

$$c_1 \leq \|u_{j,k,e}\|_{L^2} \leq c_2 \quad \forall (j, k, e) \in \Lambda,$$

for explicit constants c_1, c_2 depending on $\psi_{0,0,e}$, see Section 3.3 in Donoho (1995) for a proof of this claim. Let us remark that the system $\{u_{j,k,e}\}$ is part of a WVD for T (see Donoho (1995) for the details). In particular, it satisfies condition (3.11).

Altogether, the observations above imply that the Radon transform satisfies Assumption 4 with $\beta = 1/2$ in dimension $d = 2$. By Theorem 4, the multiscale total variation estimator (3.4) is nearly minimax optimal for recovering a function $f \in BV_L$ from noisy Radon observations. We remark that the same analysis can be performed for the Radon transform in higher dimensions, in which case $\beta = (d-1)/2$, for the X-ray transform, with $\beta = 1/2$ for any dimension (Natterer, 1986), as well as for other tomography operators, such as photoacoustic and thermoacoustic tomography (see e.g. Haltmeier (2013)).

Convolution

Let T denote the convolution operator with a kernel $K \in L^1(\mathbb{R}^d)$, i.e.,

$$Tg(x) := \int_{\mathbb{R}^d} K(x-y)g(y)\,dy.$$

We let $\mathbb{M} = \mathbb{R}^d$, and by Young's inequality T is a bounded operator from $\mathcal{D}(T) = L^2(\mathbb{R}^d)$ to itself whose operator norm equals $\|K\|_{L^1}$. The inverse problem (1.1) with a convolution operator T is a model for a myriad of applications in image and signal processing, including microscopy and astronomy models (see e.g. Bertero et al. (2009)). The problem of recovering a signal f from noisy measurement of its convolution Tf is hence of extreme practical relevance. In this section we show that the multiscale TV-estimator (3.4) solves this problem in a minimax optimal sense. For that, we need to impose regularity conditions on T, which naturally have the form of a decay condition on the Fourier transform of K. In particular, we assume that the kernel K satisfies

$$a_1\,(1+|\xi|^2)^{-\beta/2} \leq |\mathcal{F}[K](\xi)| \leq a_2\,(1+|\xi|^2)^{-\beta/2} \quad \forall \xi \in \mathbb{R}^d \tag{3.14}$$

for constants $a_1, a_2 \geq 0$ and some $\beta \geq 0$. Given a basis of Daubechies wavelets $\{\psi_{j,k,e}\}$ like that in Section 2.1 with $S > \max\{1, d/2+\beta\}$, define the system of functions

$$u_{j,k,e}(x) := 2^{j(d/2-\beta)}\mathcal{F}^{-1}\Big[\frac{\mathcal{F}[\psi_{0,0,e}](\cdot)}{\mathcal{F}[K](-2^j\cdot)}\Big](2^jx-k) \tag{3.15}$$

indexed by the set Ω in (2.2). These functions satisfy the following relations

$$T^*u_{j,k,e} = \kappa_j\,\psi_{j,k,e} \quad \text{where } \kappa_j = 2^{-j\beta},$$
$$c_1 \leq \|u_{j,k,e}\|_{L^2} \leq c_2 \quad \forall (j,k,e) \in \Omega,$$

where we can choose $c_1 = \min_{e\in\{0,1\}^d} \|(-\Delta)^{\beta/2}\psi_{0,0,e}\|_{L^2}$ and $c_2 = \max_{e\in\{0,1\}^d} \|\psi_{0,0,e}\|_{H^\beta}$ (see Proposition 13 in Section 7.3.5 for the proof). Further, it is easily verified that such a convolution operator satisfies (3.11).

These results show that the convolution operator T under the assumptions above satisfies Assumption 4. By Theorem 4 we conclude that the multiscale TV-estimator is minimax optimal for estimating functions $f \in BV_L$, up to logarithmic factors. Finally, in Section 5.2 of Chapter 5 we analyze the numerical performance of the multiscale TV-estimator for deconvolution problems.

Here we have considered convolution kernels that decay polynomially in Fourier domain, which correspond to mildly ill-posed inverse problems. We discuss the extension to exponentially ill-posed inverse problems in the Conclusion in Chapter 6.

CHAPTER 4

Computation

In this chapter we discuss how to implement and efficiently compute the multiscale estimator

$$\hat{f}_\Phi \in \underset{g \in \mathcal{F}_n}{\operatorname{argmin}} \; |g|_{BV} \;\text{ s.t. }\; \max_{\omega \in \Omega_n} \left| \langle \phi_\omega, g \rangle - Y_\omega \right| \leq \gamma_n. \tag{4.1}$$

Notice that this estimator covers the settings of regression and of inverse problems, in which case the dictionary ϕ_ω would be chosen as indicated in Chapter 3. The optimization problem in (4.1) presents two challenges:

a) The objective function $|\cdot|_{BV}$ is a *non-smooth* functional. This is a difficulty, since it yields standard optimization methods such as gradient descent inapplicable. We are left with two alternatives: either use techniques from non-smooth optimization, or find a smooth surrogate for our problem and apply standard techniques to it. We pursue both approaches: on one hand, we use the primal-dual Chambolle-Pock algorithm (Chambolle and Pock, 2011) for non-smooth optimization (see Section 4.1), and on the other hand, we use the Moreau-Yosida regularization and a Newton-type algorithm in combination with the path-following technique (see Section 4.3). Each method has advantages and disadvantages, which we discuss below.

b) The constraint in (4.1) involves the maximum over the set Ω_n. This is the index set of the dictionary Φ. As argued in the Introduction, the estimator (4.1) performs best for very redundant dictionaries Φ. This implies in particular that the set Ω_n will be quite numerous (typical numbers are $\#\Omega_n \sim 10^4$ for $d = 1$ and $n = 256$, and $\#\Omega_n \sim 10^7$ for $d = 2$ and images of size 256×256), thus making the evaluation of the constraint in (4.1) a computationally demanding task. An efficient way of dealing with this problem is to solve (4.1) by a primal-dual approach, as presented in Section 4.1. In that situation, the constraint appears only via the proximal mapping of its Fenchel conjugate, which in our case turns out to be simply the soft-thresholding operator. Soft-thresholding can be implemented efficiently, which makes the primal-dual approach very successful.

In Section 4.1 below we briefly describe the Chambolle-Pock primal-dual algorithm, and in Section 4.2 we explain how to apply it to our problem. In Section 4.3 we present an alternative algorithm that uses the Moreau-Yosida regularization and a Newton-type method to solve (4.1). Finally, we briefly comment on further alternative algorithms in Section 4.4, and discuss the advantages and disadvantages of the different approaches.

Remark 14. In this section we will work with the Fenchel transform of a functional F, denoted by $F^\star$. We warn the reader of the similarity between this and the notation for the adjoint of an operator K, denoted by K^*.

4.1 The Chambolle-Pock algorithm

The Chambolle-Pock algorithm, introduced by Chambolle and Pock (2011), is an algorithm for *non-smooth* convex optimization. Roughly speaking, it operates by solving the optimality conditions for the primal and the dual problems alternatively. In order to explain it, we need some notation. Most of the general notation and standard claims in this section can be found in Rockafellar (2015). As customary in the optimization literature, we consider all real-valued functionals in this section to map to the *extended real line* $\mathbb{R} \cup \{\pm\infty\}$. This has the consequence that infima and suprema are always attained, so we write min and max instead of inf and sup.
Let us introduce the following notation:

a) V and W are finite dimensional vector spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_W$ arising from inner products $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_W$. When it is clear from the context, we drop the dependence on V and W from the notation and write simply $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$. The topological dual W^* of W is identified with W as a vector space, and the same is done for V.

b) Let $K : V \to W^*$ be a continuous linear mapping with operator norm $\|K\|_{op} := \max_{\|v\|\leq 1} \|Kv\|$. Recall that we denote by K^* the dual of K with respect to the inner products of V and W, i.e.

$$\langle w, Kv\rangle_W = \langle K^* w, v\rangle_V \ \ \forall v \in V, \forall w \in W.$$

c) Define the Fenchel transform (also called convex conjugate) of a convex, lower-semicontinuous functional $H : W \to [0, +\infty)$ as the mapping

$$H^\star(z) := \max_{w\in W} \langle z, w\rangle - H(w), \quad z \in W^*.$$

The functional $H^\star$ is then convex and lower-semicontinuous (see Section 12 in Rockafellar (2015)).

d) Let $F^\star : W \to [0, +\infty)$ and $G : V \to [0, +\infty)$ be convex, lower-semicontinuous functionals. Assume that $F^\star$ is the Fenchel conjugate of a convex, lower-semicontinuous functional F. Then $F^{\star\star} = F$ (see Rockafellar (2015)).

e) We say that a functional F is *simple* if the resolvent mapping

$$v = (I + \tau\,\partial F)^{-1}(w) := \operatorname*{argmin}_{z \in W^*} \frac{\|z - w\|^2}{2\tau} + F(z)$$

can be computed efficiently for any $\tau > 0$. Here ∂F denotes the subdifferential of F. We remark that if F is simple, then $F^\star$ is simple as well by Moreau's identity (see the remark after Theorem 31.5 in Section 31 in Rockafellar (2015)). That something can be "computed efficiently" is an admittedly vague statement, but we shall give it a more concrete meaning in Section 4.2 when we consider particular functionals.

In this setting, the Chambolle-Pock algorithm solves problems of the form

$$\min_{v \in V} F(Kv) + G(v). \tag{4.2}$$

Expressing F as the convex conjugate of $F^\star$, this can be written as

$$\min_{v \in V} \max_{w \in W} \langle Kv, w \rangle - F^\star(w) + G(v). \tag{4.3}$$

Assume that a solution $(\overline{v}, \overline{w}) \in V \times W$ to (4.3) exists. It follows (see Theorem 31.3 in Section 31 in Rockafellar (2015)) that it satisfies the conditions

$$\begin{cases} K\overline{v} \in \partial F^\star(\overline{w}) \\ -K^*\overline{w} \in \partial G(\overline{v}), \end{cases}$$

where we write inclusions because the subgradient of a non-smooth functional is in general set-valued. Adding the identity mappings I_V and I_W of the spaces V and W, these equations can be written as

$$\begin{cases} \overline{w} + \sigma\,K\overline{v} \in (I_W + \sigma\,\partial F^\star)(\overline{w}) \\ \overline{v} - \tau\,K^*\overline{w} \in (I_V + \tau\,\partial G)(\overline{v}), \end{cases} \tag{4.4}$$

for any $\sigma, \tau > 0$. In this setting, the Chambolle-Pock algorithm finds a solution $(\overline{v}, \overline{w})$ to (4.3) by iteratively solving the two equations in (4.4), as show in Algorithm 1. Note that this can be done efficiently under the assumption that F and G are simple, which is the case for the problem we are interested in (see Section 4.2 below).

Algorithm 1 Chambolle-Pock algorithm

```
Require: $\sigma, \tau > 0$, $\theta \in (0,1]$, $N = 0$, $(v_0, w_0) \in X \times Y$, stopping criterion
while stopping criterion not satisfied do
    $w_{N+1} \leftarrow (I_W + \sigma\,\partial F^\star)^{-1}(w_N + \sigma\,K\,\widetilde{v}_N)$
    $v_{N+1} \leftarrow (I_V + \tau\,\partial G)^{-1}(v_N - \tau\,K^*\,w_{N+1})$
    $\widetilde{v}_{N+1} \leftarrow v_N + \theta(v_{N+1} - v_N)$
    $N \leftarrow N + 1$
end while
Return $(v_N, w_N)$
```

The stopping criterion in Algorithm 1 typically involves a maximum number of iterations $N_{\max}$ and a convergence criterion of the form $\max\{\|v_N - v_{N+1}\|, \|w_N - w_{N+1}\|\} < \epsilon$.

Theorem 1 in Chambolle and Pock (2011) guarantees that, if the step sizes τ and σ in Algorithm 1 satisfy $\tau\,\sigma < \|K\|_{op}^{-2}$, then there is a saddle-point $(\overline{v}, \overline{w})$ of (4.3) towards which the sequence (v_N, w_N) converges as $N \to \infty$. Recall that $\overline{v}$ is the solution to our original problem (4.2). Moreover, there is a variant of Algorithm 1 that uses so-called *acceleration* and guarantees that

$$\|v_N - \overline{v}\| \leq C\,N^{-1} \tag{4.5}$$

holds for N large enough, where $C > 0$ is a constant depending on the initialization and the parameters of the algorithm. In a nutshell, the idea behind acceleration is to choose the parameters θ, σ and τ in the algorithm to depend on the iteration number N. We refer to Theorem 2 of Chambolle and Pock (2011) for a proof of this result.

4.2 Implementation of the estimator

In order to compute the estimator in (4.1), we first need to discretize it. We do so by representing a function g by its values at a regular grid of n points in $[0,1]^d$, where $n = m^d$ for $m \in \mathbb{N}$. More precisely, if Γ_n denotes the equidistant grid in (2.17), we denote the discretization of a function g in $[0,1]^d$ by $g_n := \{g(x_i)\}_{x_i \in \Gamma_n}$. We see g_n as a d-dimensional array of size m^d: e.g. $g_n \in \mathbb{R}^{m\times m}$ is a matrix if $d = 2$, and a vector $g \in \mathbb{R}^n$ if $d = 1$. We denote the set of all such arrays by $\mathbb{R}^{\Gamma_n}$. Denote by $\|D\,g_n\|_1$ the bounded variation seminorm of the array g_n, defined as

$$\|D\,g_n\|_1 := \sum_{x\in\Gamma_n} |Dg_n(x)|,$$

where

$$|Dg_n(x)| := \sqrt{\sum_{y \leq x} \left|g_n(x) - g_n(y)\right|^2}$$

and the sum is over all neighbors y of x in the grid with coordinates not smaller than those of x, written $x \leq y$. This restriction is immaterial, but it simplifies the work with the finite difference operator, defined below. We use the convention that two points $x, y \in \Gamma_n$ are neighbors if, when seen as vectors $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d)$, they differ by m^{-1} in exactly one coordinate, and are equal in the others. Finally, we define the *finite difference* operator D by

$$Dg_n(x,y) = \begin{cases} (g_n(x) - g_n(y)) & \text{if } x \leq y \in \Gamma_n, \\ 0 & \text{else.} \end{cases}$$

Given the dictionary $\Phi = \{\phi_\omega \,|\, \omega \in \Omega\}$ of functions, we denote by ϕ_ω^n the discretization of the function ϕ_ω in the grid Γ_n. This is again a d-dimensional array of size m^d.

With this notation, the discretization of the minimization problem in (4.1) can be written as

$$\min_{g_n \in \mathbb{R}^{\Gamma_n}} \|D g_n\|_1 \;\text{ s.t. }\; \max_{\omega \in \Omega_n} \left|\langle \phi_\omega^n, g_n \rangle_{\Gamma_n} - Y_\omega\right| \leq \gamma_n, \tag{4.6}$$

where

$$\langle \phi_\omega^n, g_n \rangle_{\Gamma_n} := \frac{1}{n} \sum_{x \in \Gamma_n} \phi_\omega^n(x) g_n(x)$$

is a discretization of the L^2-inner product between the functions ϕ_ω and g. We solve the discretized problem (4.6) by formulating it in the form (4.2) and using the Chambolle-Pock algorithm. For that, we turn the constraint minimization in (4.6) into a penalized minimization problem by means of the indicator function

$$1_{\leq 0}(z) := \begin{cases} 0 & \text{if } \max_{\omega \in \Omega_n} z_\omega \leq 0 \\ +\infty & \text{else} \end{cases} \quad \text{for } z \in \mathbb{R}^{\#\Omega_n}.$$

Hence, we can write (4.6) as

$$\min_{g_n \in \mathbb{R}^{\Gamma_n}} \|D g_n\|_1 + 1_{\leq 0}(K g_n - Y - \gamma_n) + 1_{\leq 0}(-K g_n + Y - \gamma_n), \tag{4.7}$$

where $K : \mathbb{R}^{\Gamma_n} \to \mathbb{R}^{\#\Omega_n}$ is the linear operator that maps an array g_n to the vector of its $\#\Omega_n$ coefficients with respect to the discretized dictionary $\{\phi_\omega^n \,|\, \omega \in \Omega_n\}$, i.e.,

$$[K g_n]_\omega := \langle \phi_\omega^n, g_n \rangle_{\Gamma_n} \quad \text{for } \omega \in \Omega_n.$$

Notice that (4.7) indeed has the form of (4.2) with $V = \mathbb{R}^{\Gamma_n}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\Gamma_n}$, $W = \mathbb{R}^{\#\Omega_n}$ equipped with the standard inner product, the operator K as above, and functionals

$$\begin{aligned} F(v) &:= 1_{\leq 0}(v - Y - \gamma_n) + 1_{\leq 0}(-v + Y - \gamma_n) \ \text{ for } v \in \mathbb{R}^{\#\Omega_n}, \\ G(w) &:= \|Dw\|_1 \ \text{ for } w \in \mathbb{R}^{\Gamma_n}. \end{aligned}$$

Notice that G is proper, convex and lower-semicontinuous, while F is lower-semicontinuous and convex due to the fact that the constraint set in the definition of the indicator function $1_{\leq 0}$ is convex. Furthermore, the convex conjugate of F is given by

$$F^*(z) := \sup_{\substack{\max |x_\omega - Y_\omega| \leq \gamma_n \\ \omega \in \Omega_n}} \langle x, z \rangle = \langle z, Y \rangle + \gamma_n \|z\|_1 = \sum_{\omega \in \Omega_n} z_\omega Y_\omega + \gamma_n |z_\omega|.$$

Moreover, the mappings G and F^* are *simple* in the sense of Section 4.1, since their proximal mappings can be computed as the minimization problems

$$\min_{x \in \mathbb{R}^{\Gamma_n}} \frac{\|x - v\|^2}{2\tau} + \|Dx\|_1$$

for $v \in \mathbb{R}^{\Gamma_n}$ and

$$\min_{z \in \mathbb{R}^{\#\Omega_n}} \frac{\|z - w\|^2}{2\tau} + \gamma_n \sum_{\omega \in \Omega_n} \left| z_\omega - \tau Y_\omega \right| \tag{4.8}$$

for $w \in \mathbb{R}^{\#\Omega_n}$, which can be solved efficiently for any $\tau > 0$: the former by Chambolle's algorithm (Chambolle, 2004) or by quadratic programming (Nesterov and Nemirovsky, 1994), and the latter has an exact solution in terms of soft-thresholding. We conclude that the Chambolle-Pock algorithm can be used to solve the minimization problem (4.7).

Discretization of the *BV* seminorm

Remark 15 (Discretization of the BV seminorm)**.** Since we have discretized the BV seminorm in order to apply the Chambolle-Pock algorithm, one could ask how much we lose by discretizing the original problem. This was answered by Chambolle (2004), who showed that the properly rescaled discretized functional $\|D g_n\|_1$ converges to the BV seminorm in the sense of Γ-convergence. He also showed that Chambolle's algorithm in the discretized model produces reconstructions that converge to the minimizer of the continuous model in the limit $n \to \infty$.
While these results imply that one can rely on Chambolle's algorithm, some authors have shown that the discretization of the BV seminorm can be unstable in general. In the setting of Bayesian inverse problems, Lassas and Siltanen (2004) and Lassas et al. (2009) proved that imposing a

discretized BV prior (analogous to regularizing with the BV seminorm) shows the following phenomenon: as the level of discretization grows, the posterior mean estimator converges to the posterior mean corresponding to a *Sobolev* H^1 prior (Theorem 5.1 in Lassas and Siltanen (2004)). Further, Lassas et al. (2009) show that Besov $B^1_{1,1}$ priors do not show this effect. This is one of the main computational differences between the BV and the Besov $B^1_{1,1}$ or Sobolev seminorms: the former is not discretization invariant, while the latter are. We refer to Section 1.4 in the Introduction for other results concerning the discretization of the BV seminorm.

4.3 Semismooth Newton approach

Here we present an alternative approach for solving (4.1) that is based on smoothing the original problem and applying a Newton-type method to solve it. Of course, this yields the solution to a smoothed problem, and not to the original one. This issue is mitigated by the technique of path-following (see e.g. Hintermüller (2010) and Hintermüller and Rasch (2015)), which essentially amounts to iteratively solving the smoothed problem with a *decreasing* amount of regularization. Schematically, let F denote the original functional we want to minimize, and let F_ϵ denote the functional "regularized at level ϵ", whatever this means (we will see below an explicit example of regularization). The path-following schema is sketched in Algorithm 2, and is based on the following assumptions:

a) it is more difficult to minimize the unregularized functional F than its regularized version F_ϵ;

b) the smaller ϵ, the more "similar" F_ϵ and F are, and the more computationally demanding it is to minimize F_ϵ;

c) the computational cost for minimizing F_ϵ depends crucially on the initialization.

With these ideas in mind, the path-following schema would ideally start with a large parameter ϵ_0, for which the minimizer x_0 of F_{ϵ_0} is easily computed. In each iteration ϵ will get smaller, which means that F_ϵ will be more difficult to minimize, but we will also have better initialization points, which makes minimization easier.

So far we have only talked about "regularizing" the original problem in a broad sense. In the following we will consider the Moreau-Yosida regularization of the subdifferential of the functional. The reason for using it is that the semismooth Newton method applied to the Moreau-Yosida regularization of a functional is known to achieve superlinear convergence (see Section 5 of Hintermüller (2010)). One of the inspirations to use this approach is the work of Clason et al. (2018), who used these techniques to solve an optimization problem involving a BV-penalty.

Algorithm 2 Path-following schema

```
Require: $\epsilon_0 > 0$, $r \in (0,1)$, $N = 0$, $v_{-1} \in V$, mapping $\epsilon \mapsto F_\epsilon(\cdot)$, stopping criterion
while stopping criterion not satisfied do
    $v_N \leftarrow \operatorname*{argmin}_u F_{\epsilon_N}(u)$   using $v_{N-1}$ as initialization
    $\epsilon_{N+1} \leftarrow r \cdot \epsilon_N$
    $N \leftarrow N + 1$
end while
Return $v_N$
```

Let us explain this approach in more detail. We consider for simplicity the case $d = 1$, since the mappings D and D^* are then easier to handle. The optimality condition for the minimization problem (4.7) is given by the set inclusion

$$0 \in D^*(\partial\|\cdot\|_{L^1})(Du) + K^*(\partial 1_{\leq 0})(Ku - Y - \gamma_n) - K^*(\partial 1_{\leq 0})(-Ku + Y - \gamma_n), \tag{4.9}$$

where $\partial\|\cdot\|_{L^1}$ denotes the subdifferential of the L^1-norm, and $\partial 1_{\leq 0}$ denotes the subdifferential of the indicator function $1_{\leq 0}$. In $d \geq 2$, the subdifferential of the BV seminorm is slightly different, since then we have the L^1 norm of the *Euclidean norm* of the gradient (see Section 5.2 in Clason et al. (2018) for the details).

Our goal is to find a function u such that (4.9) holds, but the fact that the subdifferentials are set-valued complicates matters. Our approach here is to replace them by their Moreau-Yosida regularization, which is a single-valued Lipschitz-continuous functional. The Moreau-Yosida regularization of the subdifferential ∂F of a convex, lower-semicontinuous functional F is defined as

$$(\partial F)_\delta(v) := \frac{1}{\delta}\Big(v - (I + \delta\partial F)^{-1}(v)\Big) \quad \text{for} \quad \delta > 0.$$

We refer to Section 3 of Parikh and Boyd (2014) for further details on this regularization technique. The Moreau-Yosida regularizations of the two subdifferentials appearing in (4.9) are given in $d = 1$ by

$$(\partial\|\cdot\|_{L^1})_\delta(v) = \begin{cases} 1 & \text{if } v \geq \delta \\ \frac{v}{\delta} & \text{if } v \in (-\delta, \delta) \\ -1 & \text{else,} \end{cases}$$

$$(\partial 1_{\leq 0})_\delta(v) = \frac{1}{\delta}\max\{0, v\},$$

where the maximum is applied component-wise to the vector $v \in \mathbb{R}^{\#\Omega_n}$. Substituting the subdifferentials in (4.9) by their regularized counterparts yields the equation

$$0 = D^*(\partial\|\cdot\|_{L^1})_{\delta_1}(Du) + \frac{1}{\delta_2}K^*(\max\{Ku - Y - \gamma_n, 0\} - \max\{-Ku + Y - \gamma_n, 0\}) \tag{4.10}$$

for regularization parameters $\delta_1, \delta_2 > 0$. This is now an equation of the form $F_{\delta_1,\delta_2}(u) = 0$ for a Lipschitz-continuous functional $F_{\delta_1,\delta_2}(\cdot)$. Actually, this functional is semismooth (see Definition 2.5 in Hintermüller (2010)), which means that the semismooth Newton method can be used, and it converges superlinearly to a solution $\overline{u}$ of $F_{\delta_1,\delta_2}(u) = 0$ (see Theorem 2.14 in Hintermüller (2010)). The semismooth Newton method for this problem can be readily implemented. Denote by $\mathcal{D}_N[F_{\delta_1,\delta_2}]$ the Newton derivative of the functional at the position u_N. We initialize the iteration at u_0 and solve the linear equations

$$\mathcal{D}_N[F_{\delta_1,\delta_2}]u_{N+1} = \mathcal{D}_N[F_{\delta_1,\delta_2}]u_N - F_{\delta_1,\delta_2}(u_N) \quad \text{for } N \geq 0$$

iteratively until a stopping criterion is satisfied.

We have just described how to use the path-following technique for approximating a "difficult" optimization problem by a sequence of "easier" problems. Then we have discussed how to construct the easier problems with the Moreau-Yosida regularization, and how to solve them with the semismooth Newton method. The question now is: do we have convergence guarantees for this approach? The answer is yes, the combination of path-following and the semismooth Newton method achieves local superlinear convergence (see Section 5 of Hintermüller (2010)), i.e.,

$$|u_{N+1} - \overline{u}| \leq C\,|u_N - \overline{u}|^q \quad \text{for } \; N \in \mathbb{N}$$

for some $q > 1$, a constant $C > 0$ depending on the derivatives of F_{δ_1,δ_2}, and $\overline{u}$ being a solution of $F_{\delta_1,\delta_2}(\overline{u}) = 0$. Given a good initialization u_0, the error tends to zero considerably faster than the error of the Chambolle-Pock algorithm (4.5) does. In this sense, the semismooth Newton approach is preferable over the Chambolle-Pock algorithm.

4.4 Alternative methods and comparison

Linear programming

We remark that the problem (4.7) can be solved with other methods too. It is for instance straightforward to cast (4.7) as a *linear program* (LP):

$$\min_{(g_n,h_n)\in\mathbb{R}^{\Gamma_n}\times\mathbb{R}^{\Gamma_n\times\Gamma_n}} \sum_{(x,y)\in\Gamma_n^2} h_n(x,y) \text{ s.t. } \begin{cases} Dg_n(x,y) \leq h_n(x,y) & \forall x,y\in\Gamma_n \\ -Dg_n(x,y) \leq h_n(x,y) & \forall x,y\in\Gamma_n \\ [Kg_n]_\omega \leq Y_\omega + \gamma_n & \forall\omega\in\Omega_n \\ -[Kg_n]_\omega \leq -Y_\omega + \gamma_n & \forall\omega\in\Omega_n. \end{cases}$$

We can use this observation to solve the problem (4.7) by some standard method, e.g., the simplex algorithm or an interior point method (Nesterov and Nemirovsky, 1994). In spite of its conceptual and technical simplicity compared to the Chambolle-Pock algorithm or the semismooth Newton method presented above, the approach to (4.7) via linear programming is feasible in dimension $d = 1$ only. Its complexity scales polynomially in $n^2 = \#\Gamma_n^2$ and in $\#\Omega_n$, and since the set Ω_n is bigger in higher dimensions, the linear programming approach becomes impractical already in $d = 2$.

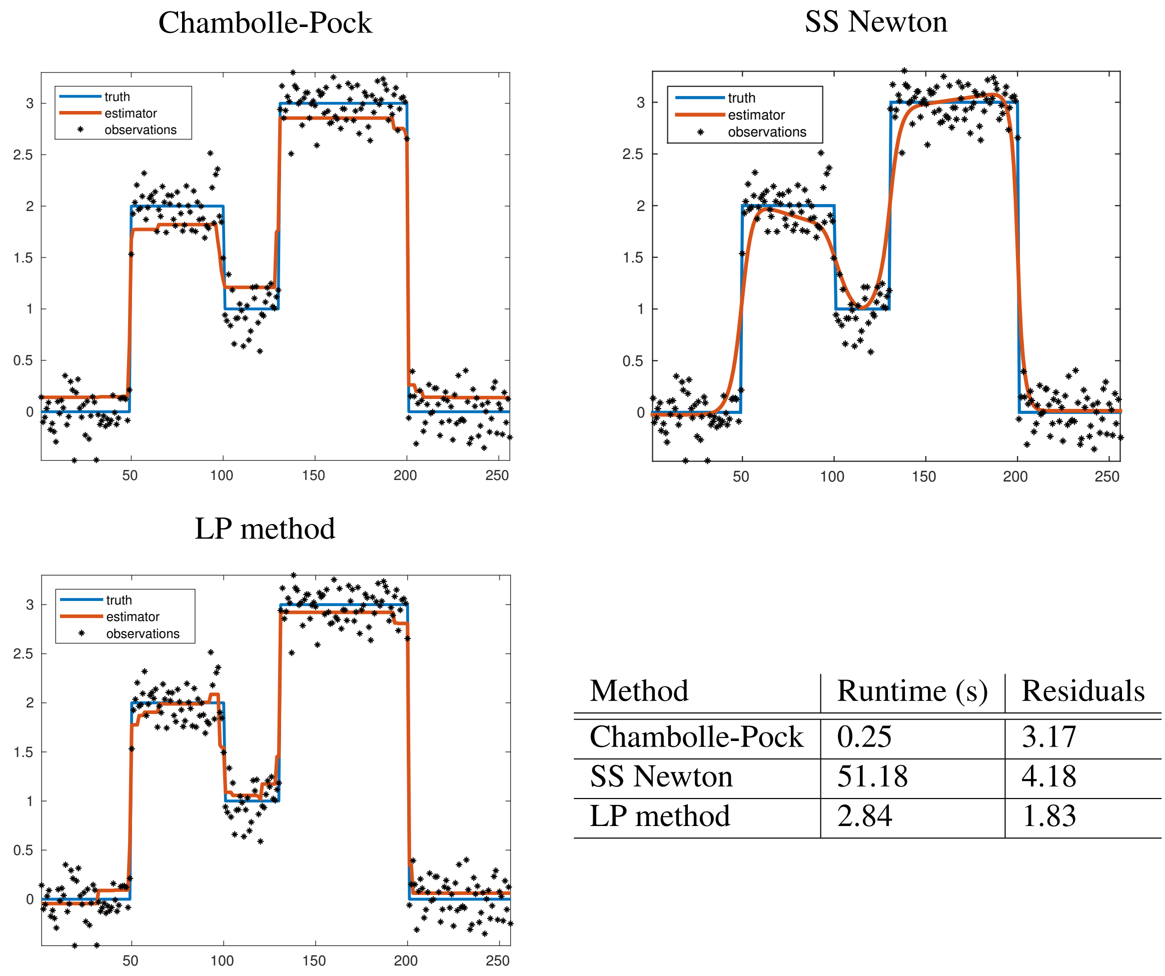
Algorithm	Advantages	Disadvantages
Chambolle-Pock	computationally efficient	poor convergence guaranty
Semismooth Newton	local superlinear convergence	computationally demanding
LP method	local quadratic convergence	feasible in $d = 1$ only

Table 4.1: Comparison of advantages and disadvantages of computation methods available for the problem (4.1). The Chambolle-Pock algorithm and the semismooth Newton approaches produce good results and are feasible in dimensions $d = 1, 2$. The iterations in the Chambolle-Pock algorithm can be computed faster, but this method enjoys a slower theoretical convergence guaranty than the semismooth Newton and the LP methods (see the table in Figure 4.1 for an illustration).

ADMM algorithm with orthogonal projections

An alternative approach for computing the estimator in (4.7) uses a variant of the *alternating direction method of multipliers* (ADMM) algorithm (Boyd et al., 2011), which was employed by Frick et al. (2012), Frick et al. (2013) and Grasmair et al. (2018) to solve minimization problems with a multiscale constraint of the form (4.7). It proceeds by splitting the problem into two subproblems: a smoothing step (using the BV-seminorm in our case), and a projection to the constraint set. Since the constraint set is the intersection of half-spaces, the projection can be computed e.g. with Dykstra's algorithm (Dykstra, 1983) or some alternative method (Bauschke et al., 2006).

We remark that the approach using the ADMM algorithm with a projection step typically has a longer runtime than the other algorithms presented here. The reason for that is that the splitting into a smoothing and a projection steps is highly asymmetric: the smoothing step can be solved very efficiently, while the projection onto the intersection of many half-spaces may be quite time consuming. The projection step is bypassed in the Chambolle-Pock algorithm by solving the dual problem instead, which has the form of soft-thresholding (4.8).


Chambolle-Pock

SS Newton

LP method

Method	Runtime (s)	Residuals
Chambolle-Pock	0.25	3.17
SS Newton	51.18	4.18
LP method	2.84	1.83

Figure 4.1: Comparison of the Chambolle-Pock algorithm, the semismooth (SS) Newton method and a LP method for solving the problem (4.1) in $d = 1$ for sample size $n = 256$, corrupted with Gaussian noise with standard deviation $\sigma = 0.1 \, \|f\|_{L^\infty}$. The runtimes in seconds and the L^2-norm of the residuals are given in the table.

Comparison

In Table 4.1 we give a summary of the main advantages and disadvantages of the methods we used. We illustrate their performance in Figures 4.1 and 4.2 for examples in $d = 1$ and $d = 2$, respectively. Notice that the semismooth Newton method produces smooth results in $d = 1$: this is natural, since it smoothes the original problem in order to apply gradient methods. On the other hand, the Chambolle-Pock algorithm and the linear programming approach produce solutions with sharp jumps, since they do not smooth the BV-functional. Moreover, as shown in the table in Figure 4.1, the Chambolle-Pock algorithm is two orders of magnitude faster than the semismooth Newton method, and they achieve comparable errors.

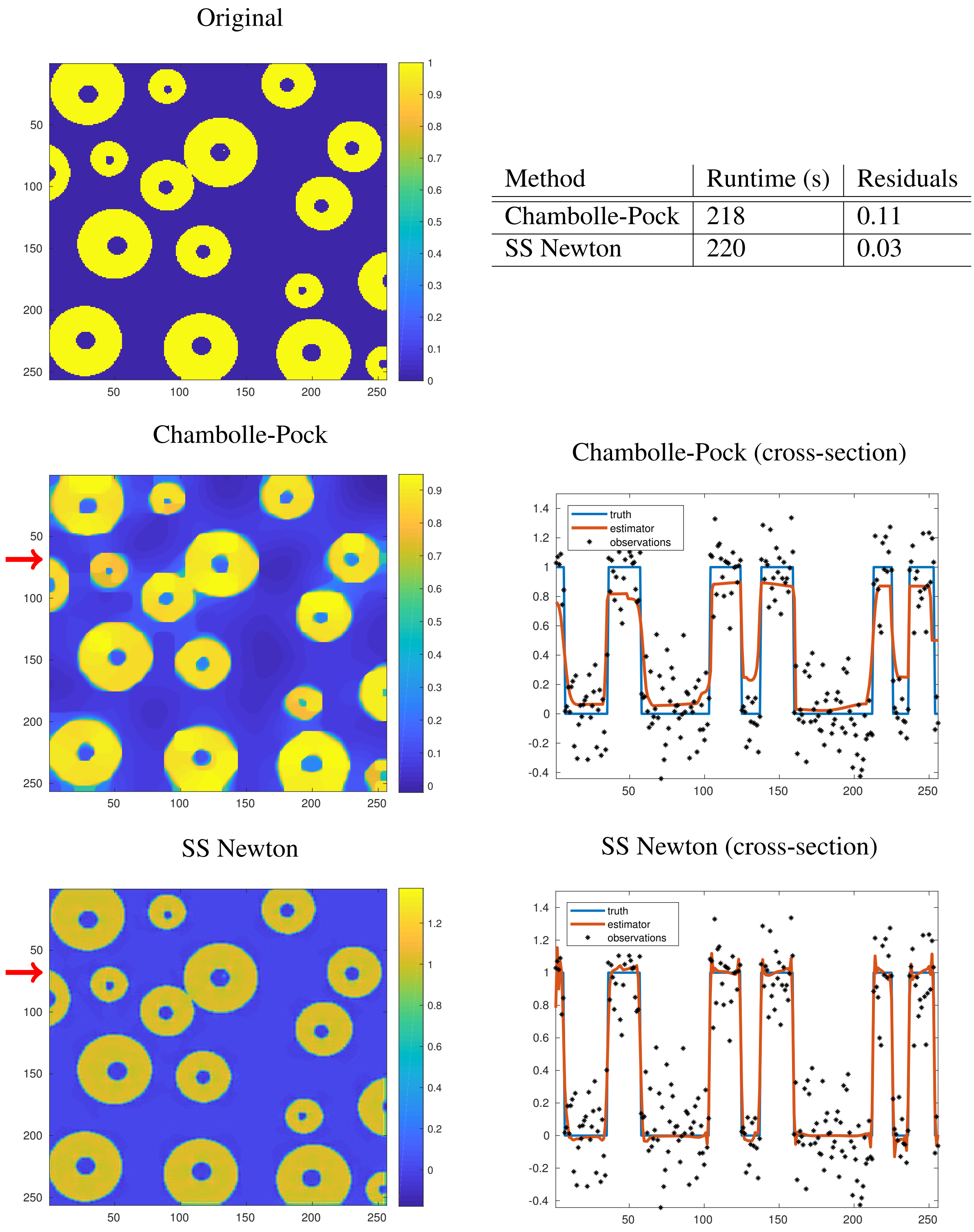
The situation in dimension $d = 2$ is slightly different: here we only compute the Chambolle-Pock and the semismooth Newton reconstructions, since the linear programming approach would be very time-consuming. In the table in Figure 4.2 we see that the two methods have similar runtime, but the relative error of the semismooth Newton method is one order of magnitude smaller than that of the Chambolle-Pock algorithm. This is also visually seen in the plots in Figure 4.2, where the semismooth Newton method provides a more satisfactory reconstruction.

Software

We implemented the estimators presented in this and the next section in MATLAB. The implementation of the Chambolle-Pock algorithm is based on the Multiscale OPtimization package (MOP), developed by Dr. Housen Li and available at `http://stochastik.math.uni-goettingen.de/mop`.

The implementation of the semismooth Newton method is based on unpublished code by Dr. Frank Werner. In his code, the objective function to be minimized is the L^2-norm instead of the BV-seminorm. This difference added an additional difficulty, since this functional is not smooth and required additional regularization, as sketched above.

In Figure 4.1 we solved the linear program with the dual-simplex algorithm from the Matlab function `linprog`, see `https://de.mathworks.com/help/optim/ug/linprog.html`.



Original

Chambolle-Pock

Chambolle-Pock (cross-section)

SS Newton

SS Newton (cross-section)

Method	Runtime (s)	Residuals
Chambolle-Pock	218	0.11
SS Newton	220	0.03

Figure 4.2: Comparison of the Chambolle-Pock algorithm and the semismooth Newton method for solving the problem (4.1) in $d = 2$ for an image of size $n = 256 \times 256$ corrupted with normal noise with standard deviation $\sigma = 0.2 \|f\|_{L^\infty}$. The cross-sections correspond to the positions marked by the red arrows. The runtimes in seconds and the relative L^2-error $\|f - \hat{f}\|_{L^2} / \|f\|_{L^2}$ are given in the table. Notice the different color scales, which are chosen to show the local variations in each image.

CHAPTER 5

Simulations

In this chapter we analyze the numerical performance of the multiscale TV-estimator on one-dimensional signals and two-dimensional images. We consider both the regression setting ($T = id$) and deconvolution inverse problems.

5.1 Simulations for regression

5.1.1 Practical considerations

In this section we show the performance of the following multiscale TV-estimators:

1) the multiscale TV-estimator with a system of dyadic intervals (in $d = 1$) or squares (in $d = 2$). We take the dictionary Φ to consist of indicator functions of a dyadic partition down to the lowest resolution scale of the image. We implemented it with the methods described in Chapter 4.

2) the multiscale TV-estimator with a curvelet frame, used on images ($d = 2$). The curvelets are computed with the package fdct_wrapping_matlab from CurveLab-2.1.3 (`http://www.curvelet.org/download.html`). The resulting estimator is a variant of the estimator proposed by Candès and Guo (2002).

In our simulation study we also considered the multiscale TV-estimator with a wavelet dictionary of symmlets with 6 vanishing moments (see e.g. Section 7.2.3 in Mallat (2008)). The basis is implemented using the package Wavelab850/Orthogonal, available in `http://statweb.stanford.edu/~wavelab/Wavelab_850/download.html`. This estimator performed similarly to the multiscale and the curvelet constrained estimators presented below, so we do not include it for the sake of conciseness.

Discretization

We evaluate the multiscale TV-estimator on observations from the nonparametric regression model, presented in Section 2.5, i.e.,

$$Y_i = f(x_i) + \sigma\,\epsilon_i, \quad x_i \in \Gamma_n, \;\; i = 1, \dots, n, \tag{5.1}$$

where ϵ_i are independent standard normal random variables, and Γ_n is an equidistant grid of n points in $[0, 1]^d$ (see (2.17)). The reason for using this model and not the white noise model is that the nonparametric regression model is arguably a more realistic model for the signal and image denoising problems that we consider in this section. For instance, in image processing one typically observes pixel values, which are properly modeled by the discrete regression model (5.1).

Besides, as shown in Section 2.5, the multiscale TV-estimator can be applied to discrete observations (5.1), yielding a discretization error of order $O(n^{-1/d})$. We showed in Section 2.5 that this error does not affect the overall convergence rate for $d = 1$ and $d = 2$, which justifies the use of the discretized model (5.1) in those cases.

Choice of γ_n

We test the estimators on several one-dimensional ($d = 1$) signals of lengths $n = 256$ and $n = 512$, and on images ($d = 2$) with $n = 256 \times 256$ pixels. The theory developed in Chapters 2 and 3 states that, asymptotically as $n \to \infty$, γ_n should be chosen as $\kappa\sigma\,\sqrt{2\log \#\Omega_n/n}$, $\kappa > 1$, in the regression setting, and correspondingly for inverse problems. For finite n, however, another choice of γ_n is possible, which gives the multiscale TV-estimator statistical interpretability. We choose a threshold of the form $\gamma_n = \sigma\, q_{1-\alpha}/\sqrt{n}$, where $q_{1-\alpha}$ is the $1-\alpha$-quantile of the statistic $\max_{\omega\in\Omega_n} |\langle\psi_\omega, dW\rangle|$, that is

$$\mathbb{P}(\max_{\omega\in\Omega_n} |\langle\psi_\omega, dW\rangle| \leq q_{1-\alpha}) = 1 - \alpha \tag{5.2}$$

for some fixed $\alpha \in (0, 1)$. This implies that the true regression function f satisfies the constraint in (1) with probability $1 - \alpha$. In practice, we compute $q_{1-\alpha}$ through Monte Carlo simulations, that is, as the empirical $1 - \alpha$-quantile of a sample of 5000 realizations of the statistic $\max_{\omega\in\Omega_n} |\langle\psi_\omega, dW\rangle|$. The quantile $q_{1-\alpha}$ can be computed independently of the observations Y_i, and is in particular independent of the true regression function f.

Finally, we remark that for some dictionaries ψ_ω, such as orthonormal wavelet bases, the distribution of $\max_{\omega\in\Omega_n} |\langle\psi_\omega, dW\rangle|$ equals that of the maximum of $\#\Omega_n$ independent normal random variables. Its $1-\alpha$ quantile is then given by

$$q_{1-\alpha} = \sqrt{2\log \#\Omega_n} + \frac{2\log\log \#\Omega_n - \log\log(1-\alpha)^{-1} + O(1)}{\sqrt{2\log \#\Omega_n}},$$

which for $\#\Omega_n \to \infty$ and $\alpha \to 0$ slowly enough is of the same order as $\sigma^{-1}\sqrt{n}\gamma_n$ in (1.9). An analogous result holds for more general dictionaries (see Kabluchko (2011)).

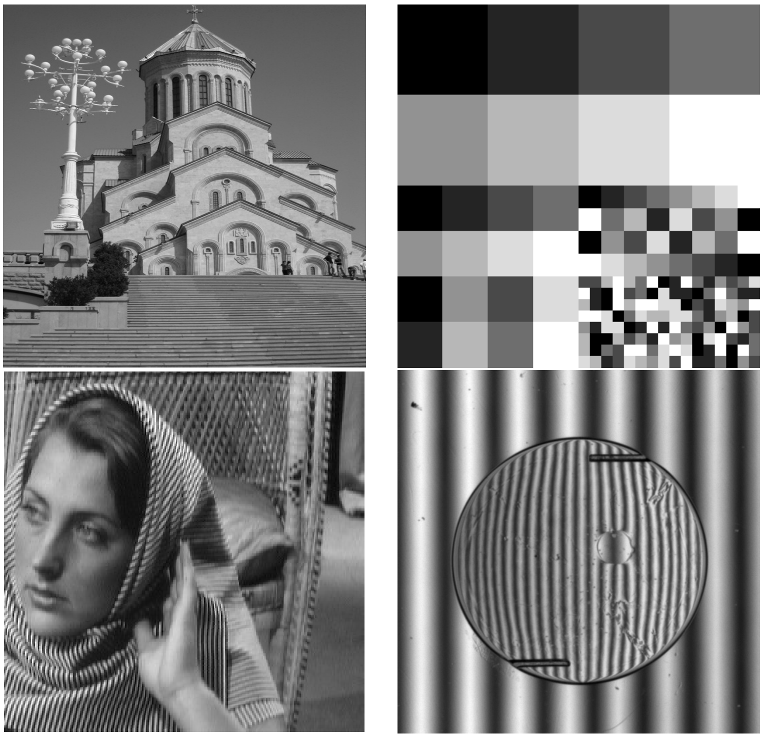


Figure 5.1: Test images used for the simulations in this section. The results are presented in Table 5.1. From top left, clockwise: 'Building', 'Board', 'Lens', and 'Barbara'.

Methods for comparison

We compare the multiscale TV-estimator with the following methods.

L^2-TV regularization. For comparison, we compute the classical TV-regularized least squares estimator

$$\hat{f}_\lambda = \operatorname*{argmin}_g \|g - Y\|_{\ell^2}^2 + \lambda |g|_{BV}. \tag{5.3}$$

Here, the data fidelity term is the empirical ℓ^2 norm, defined as

$$\|g - Y\|_{\ell^2}^2 := \frac{1}{n}\sum_{i=1}^{n} (g(x_i) - Y_i)^2.$$

The minimizer in (5.3) is computed using the well-known Chambolle algorithm (Chambolle, 2004). The regularization parameter λ in (5.3) is chosen in an oracle way so as to minimize the distance to the true regression function. In particular, we consider the two parameter choices

$$\lambda_{MSE} = \operatorname*{argmin}_{\lambda>0} \|f - \hat{f}_\lambda\|_{\ell^2},$$
$$\lambda_{Breg} = \operatorname*{argmin}_{\lambda>0} D_{BV}(f, \hat{f}_\lambda),$$

where D_{BV} denotes the symmetrized Bregman divergence of the BV seminorm:

$$\begin{aligned} D_{BV}(f,g) &:= \int (|\nabla f(x)| - |\nabla g(x)|)\Big(\frac{\nabla f}{|\nabla f|}(x) - \frac{\nabla g}{|\nabla g|}(x)\Big)\, dx \\ &= \int (|\nabla f(x)| + |\nabla g(x)|)\Big(1 - \frac{\nabla f \cdot \nabla g}{|\nabla f||\nabla g|}(x)\Big)\, dx, \end{aligned}$$

where for functions of bounded variation, the ratio $\frac{\nabla f}{|\nabla f|}$ has to be interpreted as the *sign* of the measure ∇f. The Bregman divergence associated with a convex functional is attractive because it provides a measure of similarity that matches the regularity enforced by the functional. Indeed, notice that $D_{BV}(f, g)$ is small if either f and g are constant, or if their gradients are parallel; this encourages the estimator (5.3) with λ_{Breg} to be locally constant with discontinuities close to those of the true function f.

We remark that these choices of λ are oracles in the sense that they need knowledge of the truth f for their computation. The estimators computed with these oracles are hence idealizations not accessible in practice, where λ has to be chosen in a data driven way, e.g. by Lepskii's balancing principle (Lepskii, 1991) or by cross-validation (Wahba, 1977). In particular, the comparison of these oracle estimators is not fair for the multiscale TV-estimator, which does not have knowledge of the truth f.

Wavelet or curvelet thresholding. We also compute the hard-thresholding estimator (Starck et al., 2002). In $d = 1$, we use the wavelet thresholding estimator with a basis of symmlets with 6 vanishing moments, and in $d = 2$ we employ a curvelet frame. Thresholding estimators proceed as follows: if $\{\psi_\omega \,|\, \omega \in \Omega_n\}$ denotes the dictionary in which we want to threshold, we project the observations onto ψ_ω and apply hard-thresholding to them, i.e.,

$$Y_\omega := \frac{1}{n} \sum_{x_i \in \Gamma_n} Y_i \, \psi_\omega(x_i), \quad \text{and} \quad \mathrm{Thr}(Y_\omega, \tau) := \begin{cases} Y_\omega & \text{if } \left|Y_\omega\right| \geq \tau \\ 0 & \text{if } \left|Y_\omega\right| < \tau \end{cases}$$

for a threshold τ. We observe that the choice $\tau = 3\,\sigma$ yields good results in practice. Notice that Y_ω are roughly equal to the coefficients $\langle \psi_\omega, f\rangle$ of the true function, plus normal noise. Hence, thresholding has the effect of suppressing the noise, and leaving an approximation to the true coefficients. The thresholded coefficients are then put back together with the dictionary ψ_ω (or a suitable dual frame, see e.g. Christensen (2003)), giving the estimator

$$\hat{f}_{Thr}(x) := \sum_{\omega \in \Omega_n} \psi_\omega(x) \, \mathrm{Thr}(Y_\omega, \tau).$$

Thresholding in a multiscale frame is known to give very good empirical results in image denoising, and it also enjoys optimality guaranties (Donoho and Johnstone (1998), Candès and Donoho (2000)). This method is nevertheless known to present oscillatory artifacts in its reconstructions, which arise when frequencies are wrongly suppressed by thresholding. The curvelet and wavelet transforms are implemented using the software in CurveLab-2.1.3 (`http://www.curvelet.org/download.html`) and Wavelab850/Orthogonal (`http://statweb.stanford.edu/~wavelab/Wavelab_850/download.html`), respectively.

MIND estimator. Finally, we also compare our estimator with the Multiscale Nemorovski-Dantzig estimator (MIND) proposed by Grasmair et al. (2018) (see also Li (2016)). The MIND uses a multiscale penalty akin to the one we use here, but where the test functions ψ_ω are indicator functions of dyadic intervals or squares. Moreover, the MIND minimizes a Sobolev penalty instead of the BV seminorm. In formulas, the MIND is defined as

$$\hat{f}_{\mathrm{MIND}} \in \operatorname*{argmin}_g \|D^k g\|_{L^2}^2 \quad \text{such that} \quad \max_{\omega \in \Omega_n} \left| \langle \psi_\omega, g - Y \rangle \right| \le \gamma_n.$$

The threshold γ_n is chosen by a quantile construction as explained above, and $k \in \mathbb{N}$ is a tuning parameter. In simulations for $d = 1$ we observe that $k = 1$ yields very irregular reconstructions, while $k \geq 3$ gives heavily oversmoothed results (see e.g. Figure 5.2). We therefore use $k = 2$ in the following. However, for $d = 2$ we found that $k = 1$ gives the best performance. The MIND is computed using the MATLAB package MOP from Grasmair et al. (2018), which is available at `http://stochastik.math.uni-goettingen.de/mop`.

Quality measures

Besides the qualitative comparison of the reconstructions, we also evaluate the performance of the estimators quantitatively. For that, we consider the risk with respect to the L^q norm, $q \in \{1, 2, \infty\}$, and with respect to the BV seminorm. Clearly, the L^∞ norm measures the largest deviation, while the L^1 and L^2 risks measure the averaged deviation in different ways. The BV seminorm is also of interest as a risk functional, since it measures how much noise or artifacts are still present in the reconstruction.

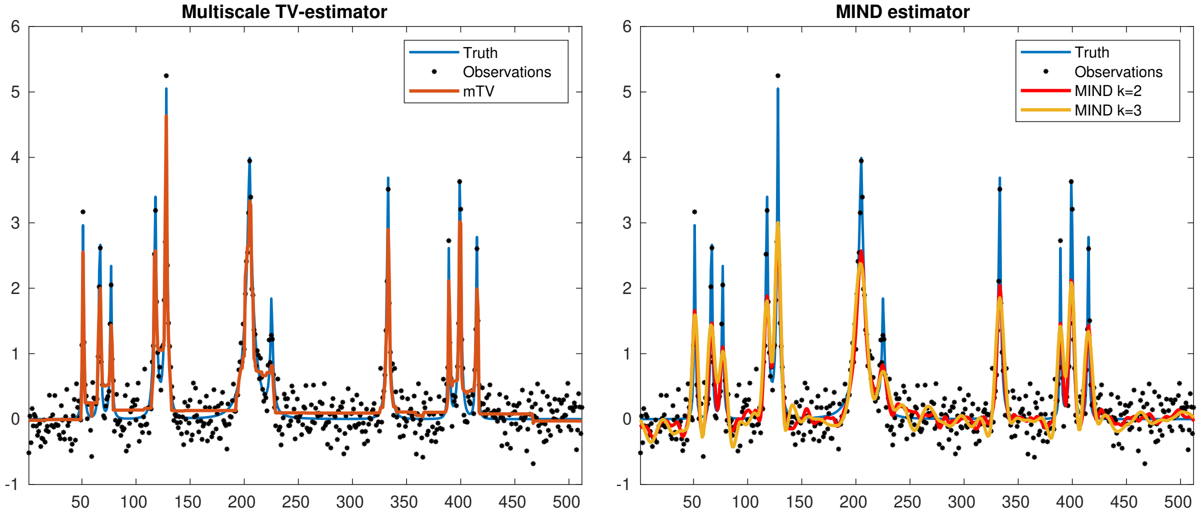


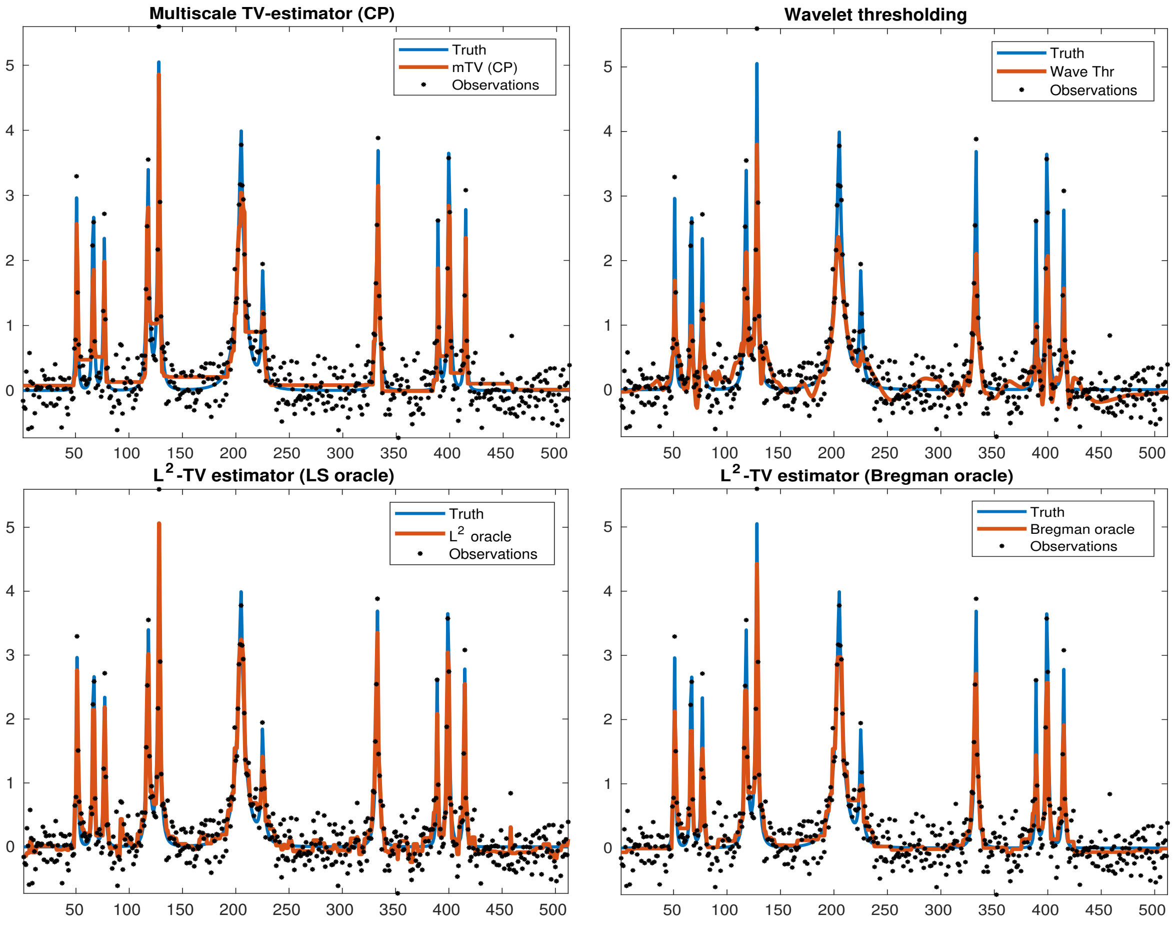
Figure 5.2: Multiscale TV-estimator and MIND estimator on Bumps signal, with $n = 512$ and $\sigma = 0.05\,\|f\|_{L^\infty}$.

For $d = 2$ we also consider the structural similarity index (SSIM). The SSIM was introduced by Wang et al. (2004), and it measures the (dis)similarity between two images taking into account the luminance (i.e. magnitude), contrast (i.e. variance) and structure (i.e. covariance) of the images (i.e. of their pixel values). Given two images F and G, the SSIM between them is defined as

$$\mathrm{SSIM}(F,G) := \frac{(2\mu_F\,\mu_G + c_1)(2\sigma_{FG} + c_2)}{(\mu_F^2 + \mu_G^2 + c_1)(\sigma_F^2 + \sigma_G^2 + c_2)},$$

where μ_F is the average of the pixel values of image F, σ_F^2 is their variance, and $\sigma_{FG} = (N-1)^{-1}\sum_{i=1}^{N}(F_i - \mu_F)(G_i - \mu_G)$ is a sort of "covariance" between pixel values of F and G. The constants $c_1, c_2 > 0$ are independent of the images, and are chosen to avoid division by a small number. The SSIM is thought to be a more sensitive quality measure than the mean square error, the peak signal-to-noise ratio, or L^q-risks in general. It takes values in the interval $[-1, 1]$, and larger values indicate more similar images.

Finally, we remark how difficult a problem it is to conceive a quality measure that matches human perception: neither the L^q norms nor the BV seminorm do so. While the SSIM seems to be a good candidate, we are not aware of any theoretical result proving that a certain method performs well with respect to the SSIM. A promising alternative that has been proposed recently involves the Wasserstein metric (Weed and Berthet, 2019). We discuss this and other possibilities in the Conclusion in Chapter 6.



Error / Method	Time (s)	L^2 error	L^1 error	L^∞ error	BV error
Multiscale TV (CP)	0.33	4.72	69.46	1.06	**25.48**
MIND ($k = 2$)	0.72	6.01	91.72	1.84	63.55
Wavelet thresholding	**0.03**	6.78	88.43	1.73	36.72
L^2-TV with λ_{MSE}	0.09	**3.22**	53.48	**0.74**	25.73
L^2-TV with λ_{Breg}	0.48	4.09	**51.10**	1.16	25.86

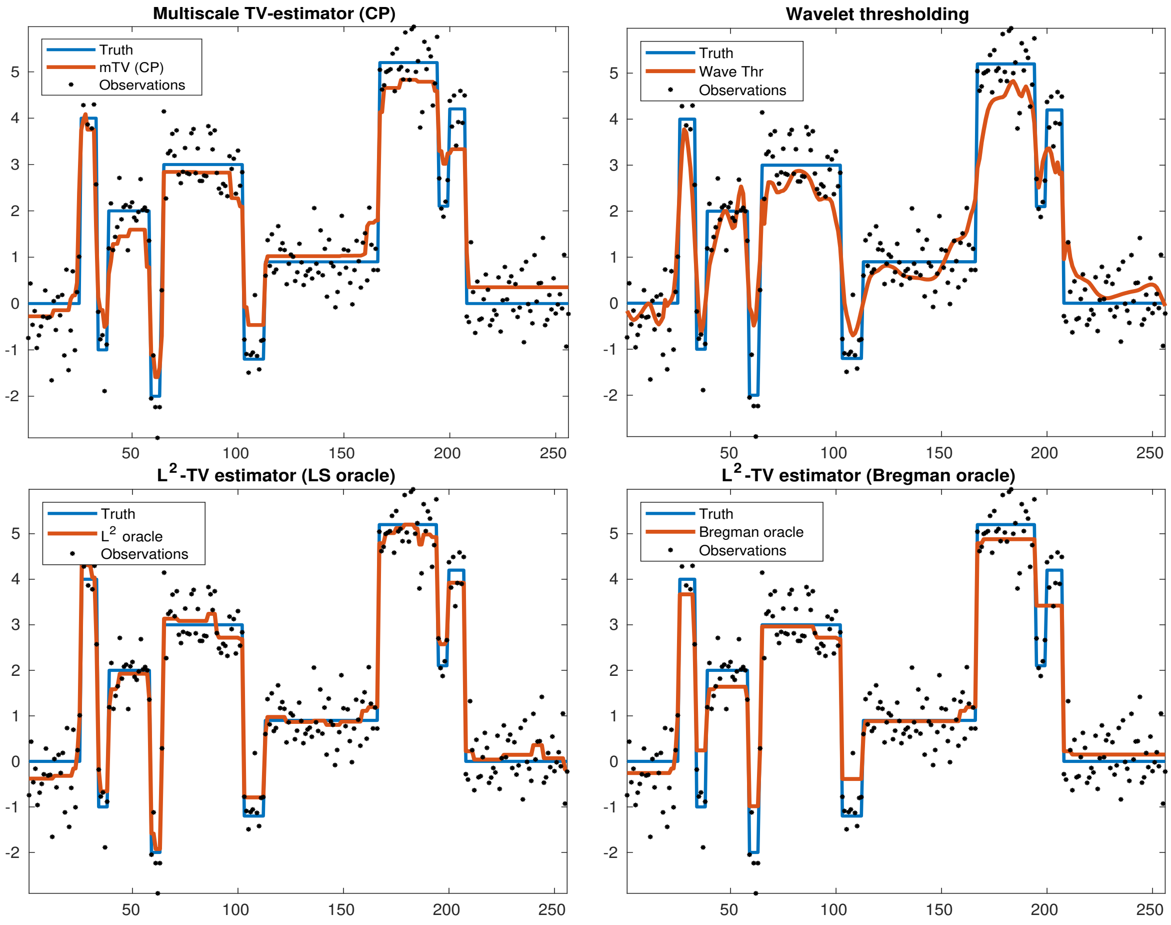
Figure 5.3: Bumps function, with $n = 512$ observations, $\sigma = 0.05\,\|f\|_{L^\infty}$. In the table: runtimes and risks of these estimators plus that of the MIND estimator with $k = 2$.

5.1.2 Simulation results

Simulations in one dimension

We simulate in one dimension the performance of the following estimators:

1) The multiscale TV-estimator, constructed with a set of dyadic intervals. We show the estimator computed with the Chambolle-Pock algorithm presented in Chapter 4 (the LP and the semismooth Newton approaches lead to roughly the same reconstructions and risks). The threshold γ_n is chosen as in Section 5.1.1 with the empirical $\alpha = 0.05$ quantile.



Error / Methods	Time (s)	L^2 error	L^1 error	L^∞ error	BV error
Multiscale TV (CP)	0.31	6.29	74.58	1.51	**25.91**
MIND ($k = 2$)	0.17	14.05	190.85	2.56	78.78
Wavelet thresholding	**0.01**	8.84	64.87	3.32	34.56
L^2-TV with λ_{MSE}	0.05	**5.17**	59.43	**1.45**	31.59
L^2-TV with λ_{Breg}	0.45	5.36	**57.05**	1.65	29.25

Figure 5.4: Blocks function, with $n = 256$ observations, $\sigma = 0.1 \, \|f\|_{L^\infty}$. In the table: runtimes and risks of these estimators plus that of the MIND estimator with $k = 2$.

2) The L^2-TV estimator with L^2 oracle λ_{MSE} and with Bregman oracle λ_{Breg}.

3) The wavelet hard-thresholding estimator with threshold $\tau = 3\sigma$. The wavelets are symmlets with 6 vanishing moments, as described in Section 5.1.1.

4) The MIND estimator with $k = 2$, and threshold γ_n chosen as in Section 5.1.1 with the empirical $\alpha = 0.05$ quantile.

We present the performance of these estimators in two standard signals with different sample size n and variance σ^2. The signals are "Blocks" and "Bumps" (Donoho and Johnstone, 1994).

The performance in these two signals is representative of what we have observed in others. For each method, we compute its error with respect to the L^1, L^2, L^∞ norms, as well as the BV seminorm. We also record the runtime of each method. The results for the two signals are shown in Figures 5.3 and 5.4. In each caption, one reconstruction method together with the observations and the ground truth is shown. A table presents the runtimes and risks of the different methods. In Figure 5.2 we compare the multiscale TV and the MIND estimators. The results can be summarized as follows:

a) Concerning runtime, wavelet thresholding is clearly superior to the other methods. However, even though it captures the main features of the signals, such as modes, it presents too many oscillatory artifacts. Consequently, the BV error of wavelet thresholding is specially high.

b) Concerning the L^q-risks, the L^2-TV estimator with L^2 oracle performs better than the multiscale TV-estimator. The reason for that is clear: the L^2-TV estimator is tuned in order to minimize the L^2-risk, which helps in minimizing the other risks. On the other hand, the multiscale TV-estimator has the smallest BV risk. This indicates that it does not include many noisy artifacts.

c) Concerning the presence of artifacts, both the multiscale TV-estimator and the Bregman oracle perform well, while the MIND, the L^2 oracle and the wavelet thresholding develop oscillatory artifacts.

d) Concerning the level of detail of the reconstruction, the multiscale TV-estimator, the MIND and the L^2 oracle capture the main features of the signals, such as modes and valleys. On the other hand, the Bregman oracle seems to miss some features, possibly due to oversmoothing.

Simulations in two dimensions

In two dimensional images we simulate the performance of the following estimators:

1) The multiscale TV-estimator, constructed either with a set of indicator functions of dyadic squares (Figure 5.6) or with a curvelet frame (Figure 5.7). In both cases, we choose γ_n with the quantile construction from Section 5.1.1 with $\alpha = 0.05$. The estimators are computed with the Chambolle-Pock algorithm presented in Chapter 4. We remark that the linear programming approach is not competitive in two dimensions for $\#\Omega_n$ large (here we have $\#\Omega_n \sim 10^7$), and the semismooth Newton approach performs essentially like the Chambolle-Pock method here.

2) The L^2-TV estimator with L^2 oracle λ_{MSE} and with Bregman oracle λ_{Breg}.

3) The curvelet hard-thresholding estimator with threshold $\tau = 3\sigma$.

4) The MIND estimator with $k = 1$, and threshold γ_n chosen as in Section 5.1.1 with the empirical $\alpha = 0.05$ quantile.

We compare these estimators on images of size 256 × 256 corrupted with normal noise with standard deviation $\sigma = 0.2\,\|f\|_{L^\infty}$, where $\|f\|_{L^\infty}$ denotes the maximal pixel value of the uncorrupted image. For each method, we record its runtime and its *relative* risk with respect to the L^q-loss, $q \in \{1, 2, \infty\}$, and with respect to the *BV* seminorm. Further, we also show their SSIM (see Section 5.1.1). The use of the relative risk, i.e., $\|\hat{f} - f\|_{L^p}/\|f\|_{L^p}$, has the effect of making the risk for different images comparable.

We evaluate the estimators on test images from the *Digital Image Processing, 3er edition* (DIP3/e) database, available under `http://imageprocessingplace.com/DIP-3E/dip3e_book_images_downloads.htm`. In Table 5.1 we present the SSIM values achieved by the methods in four representative images.

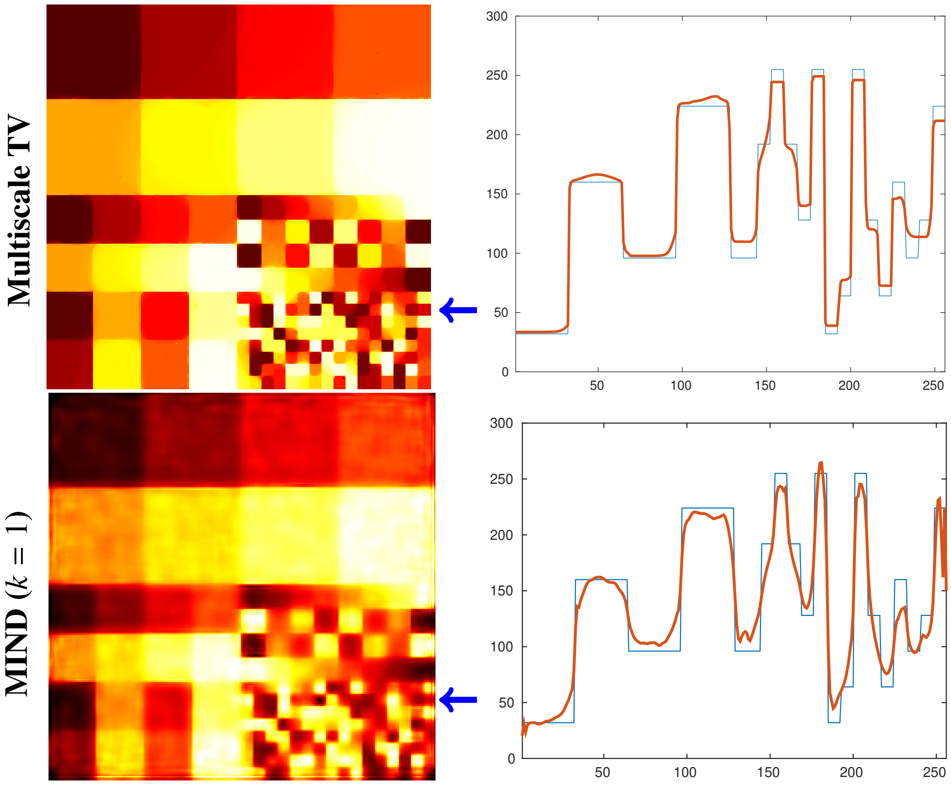


Figure 5.5: Multiscale TV-estimator and MIND estimator with $k = 1$ on "Board" image of size $n = 256 \times 256$ with $\sigma = 0.2\,\|f\|_{L^\infty}$. The positions of the cross-sections are marked with an arrow.

Images Methods	Building	Board	Barbara	Lens
Multiscale TV	**0.81**	**0.96**	0.80	0.85
MIND ($k = 1$)	0.67	0.64	0.70	0.88
Curvelet thresholding	0.79	0.69	**0.82**	**0.92**
L^2-TV with λ_{MSE}	0.59	0.53	0.75	0.70
L^2-TV with λ_{Breg}	0.74	0.94	0.77	0.76

Table 5.1: Comparison of the different methods in terms of the structural similarity index (SSIM, see Section 5.1.1) in four representative images (see Figure 5.1). The "Building" image is shown in Figure 1.2 in the Introduction.

Error Method		Time (s)	L^2 error	L^1 error	L^∞ error	BV error
Board	Multiscale TV	470.95	$2.06 \cdot 10^{-2}$	$\mathbf{2.90 \cdot 10^{-2}}$	**0.38**	**0.48**
	MIND ($k = 1$)	175.44	$7.72 \cdot 10^{-2}$	$9.72 \cdot 10^{-2}$	0.64	1.47
	Curvelet thresholding	**2.03**	$3.13 \cdot 10^{-2}$	$6.66 \cdot 10^{-2}$	0.57	1.98
	L^2-TV with λ_{MSE}	39.48	$\mathbf{1.49 \cdot 10^{-2}}$	$6.90 \cdot 10^{-2}$	0.41	2.41
	L^2-TV with λ_{Breg}	54.29	$3.56 \cdot 10^{-2}$	$4.77 \cdot 10^{-2}$	0.45	0.56
Lens	Multiscale TV	504.23	$3.92 \cdot 10^{-2}$	$7.40 \cdot 10^{-2}$	0.34	0.44
	MIND ($k = 1$)	163.78	$6.19 \cdot 10^{-2}$	$8.85 \cdot 10^{-2}$	0.49	0.43
	Curvelet thresholding	**1.25**	$2.38 \cdot 10^{-2}$	$5.03 \cdot 10^{-2}$	0.45	**0.32**
	L^2-TV with λ_{MSE}	16.20	$\mathbf{1.92 \cdot 10^{-2}}$	$\mathbf{1.08 \cdot 10^{-2}}$	0.45	0.70
	L^2-TV with λ_{Breg}	19.60	$6.68 \cdot 10^{-2}$	$9.54 \cdot 10^{-2}$	**0.30**	0.58

Table 5.2: Runtime and risks of the different methods on the "Board" and "Lens" images. The reconstructions are shown in Figures 5.6 and 5.7, respectively.

In Figures 5.6 and 5.7 we show the reconstructions of the different methods in the "Board" and "Lens" images, while the reconstruction for the "Building" image is shown in Figure 1.2 in the Introduction, and the "Barbara" image is omitted for conciseness. In Table 5.2, the runtime and BV and L^q-risks, $q \in \{1, 2, \infty\}$, of the different methods on the "Board" and "Lens" images are presented.

The results of the simulations can be summarized as follows:

a) Concerning the SSIM, the results in Table 5.1 show that curvelet thresholding and the multiscale TV-estimator outperform TV-regularization in all the examples. This is in agreement with the visual impression of the reconstructions in Figures 1.2, 5.6 and 5.7.

b) Concerning runtime, curvelet thresholding is unsurprisingly superior, while the multiscale TV-estimator is slower than L^2-TV by an order of magnitude. The risks in Table 5.2 present a more complex scenario. On one hand, the multiscale TV-estimator with a set of dyadic cubes is clearly superior to the others in the "Board" image (Figure 5.6). This is not surprising, since that is a locally constant image, where both total variation and a dictionary of indicator functions are bound to perform well. On the other hand, TV-regularization and curvelet thresholding have the best risks in the "Lens" image (Figure 5.7).

c) Concerning the presence of artifacts or noise in the reconstruction, we see in Figures 5.6 and 5.7 that the L^2-TV estimator with L^2 oracle still presents noise, while the curvelet thresholding estimator shows artifacts, which are specially prominent in the "Board" image. On the other hand, the Bregman oracle and the multiscale-TV estimator rightly denoise the image without developing artifacts or leaving noise.

d) Concerning the level of details of the reconstruction, curvelet thresholding and the multiscale TV-estimator perform best, as they identify essentially all features of the image. The L^2 oracle also does so, but some details are lost due to the noise, while the Bregman oracle smoothes out some details.

e) Concerning the comparison between the MIND and the multiscale TV-estimator, we see in Figure 5.5 that the MIND with $k = 1$ tends to oversmoothing. This is not surprising, as it penalizes the Sobolev H^1 seminorm, which is smoother than the BV seminorm. In terms of the risk, we see that the MIND is not competitive with the multiscale TV-estimator.

Summarizing, these results support the intuition that the multiscale TV-estimators combine desirable properties of TV-regularization and of multiscale dictionaries. Indeed, TV-regularization enforces locally constant reconstructions and suppresses Gibbs oscillations, and the multiscale dictionaries impose proximity to the data at all scales simultaneously. This is best seen in Figure 5.7: the TV-regularizer with λ_{Breg} removes the noise and gives a good locally constant reconstruction at the big scales only, and over-regularizes the small details; on the other hand, curvelet thresholding reconstructs the image very well down to the smallest scales, at the prize of including artifacts. The multiscale TV-estimator has the ability to perform well in the small scales, and it avoids artifacts due to the smoothing effect of the bounded variation penalty.

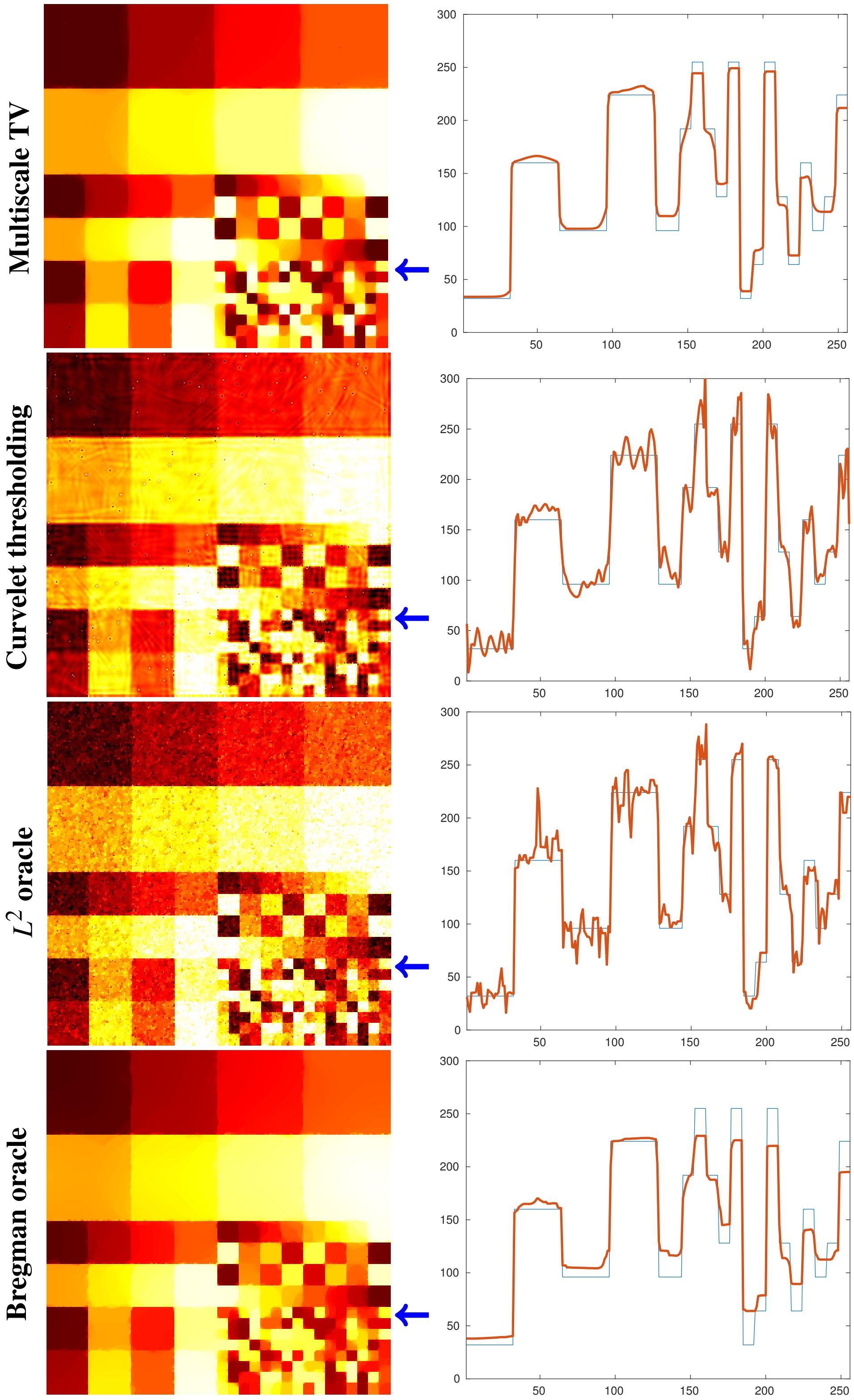


Figure 5.6: Reconstruction and cross-section (marked with the arrow) of the noisy 256 × 256 "Board" image with $\sigma = 0.2\,\|f\|_{L^\infty}$.

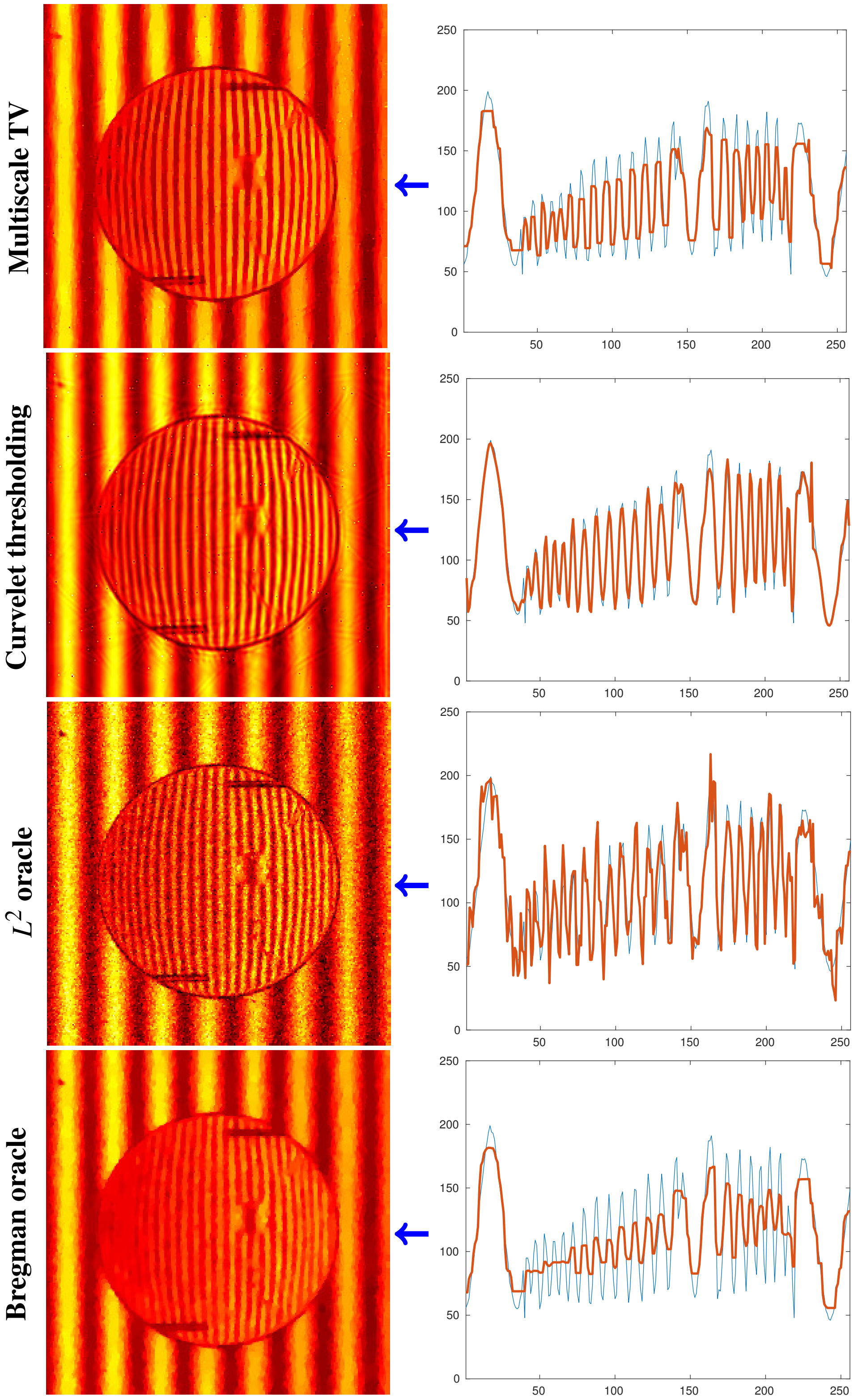


Figure 5.7: Reconstruction and cross-section (marked by the arrow) of the noisy 256 × 256 "Lens" image with $\sigma = 0.2\,\|f\|_{L^\infty}$.

5.2 Simulations for deconvolution

5.2.1 Practical considerations

In this section we analyze the performance of the multiscale TV-estimator for deconvolution. In this case, we observe noisy samples of the convolved regression function

$$Y_i = Tf(x_i) + \sigma\epsilon_i, \quad x_i \in \Gamma_n, \quad i = 1, \dots, n, \tag{5.4}$$

where

$$Tf(x) := \int_{\mathbb{R}^d} K(x-y)f(y)\,dy \quad \text{for } x \in [0,1]^d.$$

Here, $\epsilon_i \sim \mathcal{N}(0,1)$ are independent random variables, $x_i \in \Gamma_n$ are points in an equidistant grid of n points (see (2.17)), and $K \in L^1(\mathbb{R}^d)$ is a known kernel. As in the example for deconvolution in Section 3.3, we assume that the Fourier transform of the kernel K decays polynomially at infinity. In particular, we assume in the following that

$$\mathcal{F}[K](\xi) = (1 + b^2\,|\xi|^2)^{-\beta/2} \quad \forall \xi \in \mathbb{R}^d \tag{5.5}$$

for constants $b, \beta > 0$. We remark that a convolution operator with kernel K as above has singular values decaying as $\kappa_j = 2^{-j\beta}$. The left caption of Figure 5.8 shows such a kernel K in $d = 1$ with $\beta = 2$ and $b = 6.4$, centered at $x = 1/2$.

Multiscale TV-estimator

In order to compute the multiscale TV-estimator as presented in Section 3.3, we need a dictionary $\{\psi_\omega\}$ and the corresponding vaguelette system $\{u_\omega\}$ (recall Assumption 4). In this section we choose the dictionary elements ψ_ω to be dilations and translations of a symmetric beta density, e.g. in $d = 2$ we have

$$\psi_{0,0}(x,y) := x^\rho(1-x)^\rho y^\rho(1-y)^\rho \, 1_{[0,1]^2}(x,y) \tag{5.6}$$

for some $\rho > 0$. In the following we choose $\rho = 4$. This choice of the dictionary ψ_ω is motivated by previous work on the problem of testing qualitative features of a function f from convolved and noisy observations. For that, statistical tests are performed on the empirical coefficients $\langle u_\omega, Y\rangle$. The work by Proksch et al. (2018) (see also Schmidt-Hieber et al. (2013)) showed that choosing the dictionary ψ_ω as in (5.6) minimizes the variance of the test statistics among all tensor-type probe functionals, provided that ρ in (5.6) matches the order of decay β of the Fourier transform of the convolution kernel.

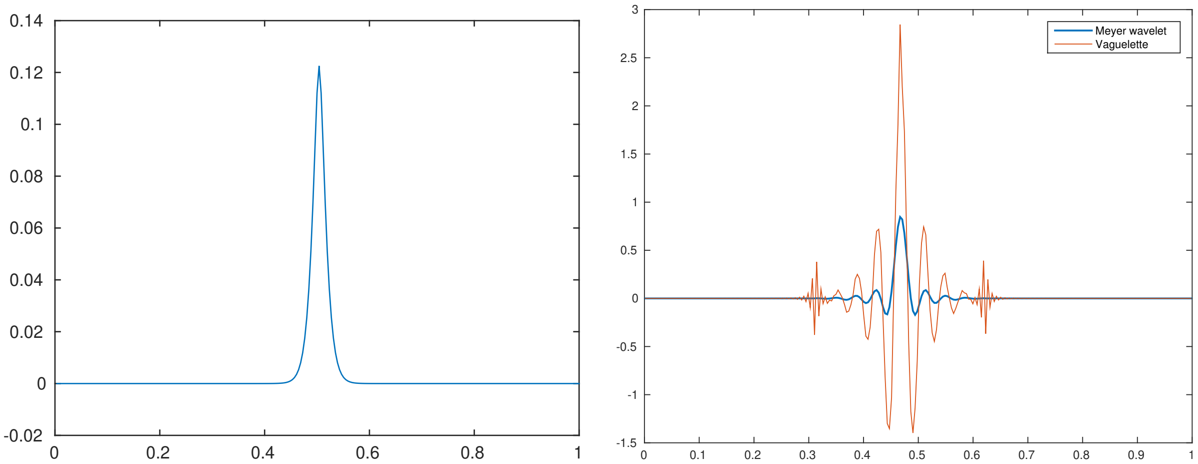


Figure 5.8: Left: convolution kernel K defined by (5.5) with $\beta = 2$ and $b = 6.4$. Right: Meyer father wavelet (blue) and corresponding vaguelette (orange).

Given the test function (5.6), our dictionary consists of dilations and translation of it. We consider a fixed set of dilations satisfying

$$\operatorname{supp} \psi_{j,k} = k + [0, L_1] \times [0, L_2], \quad \text{for} \quad L_1, L_2 \in \{\ell_1, \dots, \ell_q\}$$

for all dilations and translations (j, k) considered. Written in units of pixels, the set of allowed sizes $\{\ell_1, \dots, \ell_q\}$ that we employ in the following is either $\{5, 10, \dots, 40\}$ (see Figures 5.12 and 5.13) or $\{2, 4, \dots, 20\}$ (Figures 5.10, 5.11 and 5.14). This means that the smallest support of our dictionary elements is 5×5 or 2×2, respectively. We employ *all* possible translations of the dictionary elements at each scale. The minimal possible translation length is of course one pixel. With this choice of the multiscale dictionary $\{\psi_\omega\}$, we choose the associated vaguelette system $\{u_\omega\}$ as in Assumption 4 in Chapter 3.

The threshold $\gamma_n = \sigma\, q_{1-\alpha}/\sqrt{n}$ is chosen as in Section 5.1.1, that is: $q_{1-\alpha}$ is the empirical $1-\alpha$ quantile of 5000 realizations of the random variable $\max_{\omega\in\Omega_n} |\langle u_\omega, dW\rangle|$. In practice we choose $\alpha = 0.05$.

Methods for comparison

We compare the multiscale TV-estimator with the following methods:

L^2-TV regularization. As in Section 5.1.1, we compute the L^2-TV estimator for inverse problems

$$\hat{f}_\lambda = \operatorname*{argmin}_g \|Tg - Y\|_{\ell^2}^2 + \lambda |g|_{BV}, \tag{5.7}$$

where Tg denotes the convolution of K with g. As in Section 5.1.1, we choose the Lagrange multiplier λ in an oracle way so as to minimize the risk of the estimator $\hat{f}_\lambda$: λ_{MSE} gives the estimator with smallest L^2 error, while λ_{Breg} gives the estimator with smallest BV-Bregman divergence to the truth f.

Wavelet-vaguelette thresholding

We also compute the wavelet-vaguelette thresholding estimator proposed by Donoho (1995), which is defined as follows. Let $\{\psi_{j,k,e}\}$ be a basis of Meyer wavelets (Meyer, 1990), and $\{u_{j,k,e}\}$ be the associated vaguelette system (see Figure 5.8 for an illustration of the Meyer wavelet and the associated vaguelette). The observations Y_i are first projected onto the vaguelette system,

$$Y_{j,k,e} := \frac{1}{n} \sum_{x_i \in \Gamma_n} Y_i \, u_{j,k,e}(x_i).$$

By construction, these discretized vaguelette coefficients are discrete approximations to rescaled and noisy wavelet coefficients of the true regression function $\kappa_j \langle f, \psi_{j,k,e} \rangle + \sigma \langle \epsilon, u^n_{j,k,e} \rangle$. We hence divide $Y_{j,k,e}$ by the singular value $\kappa_j = 2^{-j\beta}$, and do scale dependent thresholding in order to remove the noise. We use the scale dependent threshold

$$\tau_j = \frac{3\,\sigma}{\kappa_j} \frac{1}{\#\{(k,e)\,|\,(j,k,e) \in \Omega_n\}} \sum_{(k,e)\,|\,(j,k,e) \in \Omega_n} \|u_{j,k,e}\|_{L^2},$$

where we average the L^2 norm of all vaguelettes at scale j. The rescaled and thresholded observations are given by

$$\mathrm{Thr}(Y_{j,k,e}/\kappa_j, \tau_j) := \begin{cases} \frac{Y_{j,k,e}}{\kappa_j} & \text{if } \left|\frac{Y_{j,k,e}}{\kappa_j}\right| \geq \tau_j \\ 0 & \text{if } \left|\frac{Y_{j,k,e}}{\kappa_j}\right| < \tau_j. \end{cases}$$

The final estimator is constructed by putting these coefficients back together with the wavelet basis $\psi_{j,k,e}$,

$$\hat{f}_{WV}(x) = \sum_{(j,k,e) \in \Omega_n} \psi_{j,k,e}(x) \, \mathrm{Thr}(Y_{j,k,e}/\kappa_j, \tau_j).$$

The rationale behind this approach is that $Y_{j,k,e}/\kappa_j$ is roughly $\langle f, \psi_{j,k,e} \rangle$ plus noise, so performing thresholding should help to remove the noise. We illustrate this procedure in Figure 5.9. In the main plot we see the empirical rescaled vaguelette coefficients $Y_{j,k,e}/\kappa_j$ in blue, the thresholded rescaled coefficients $\mathrm{Thr}(Y_{j,k,e}/\kappa_j, \tau_j)$ in red, and the true wavelet coefficients of $\langle f, \psi_{j,k,e} \rangle$ in yellow. The x axis shows the index (j, k, e): to the left are the indices corresponding to small scales, i.e., to large j.

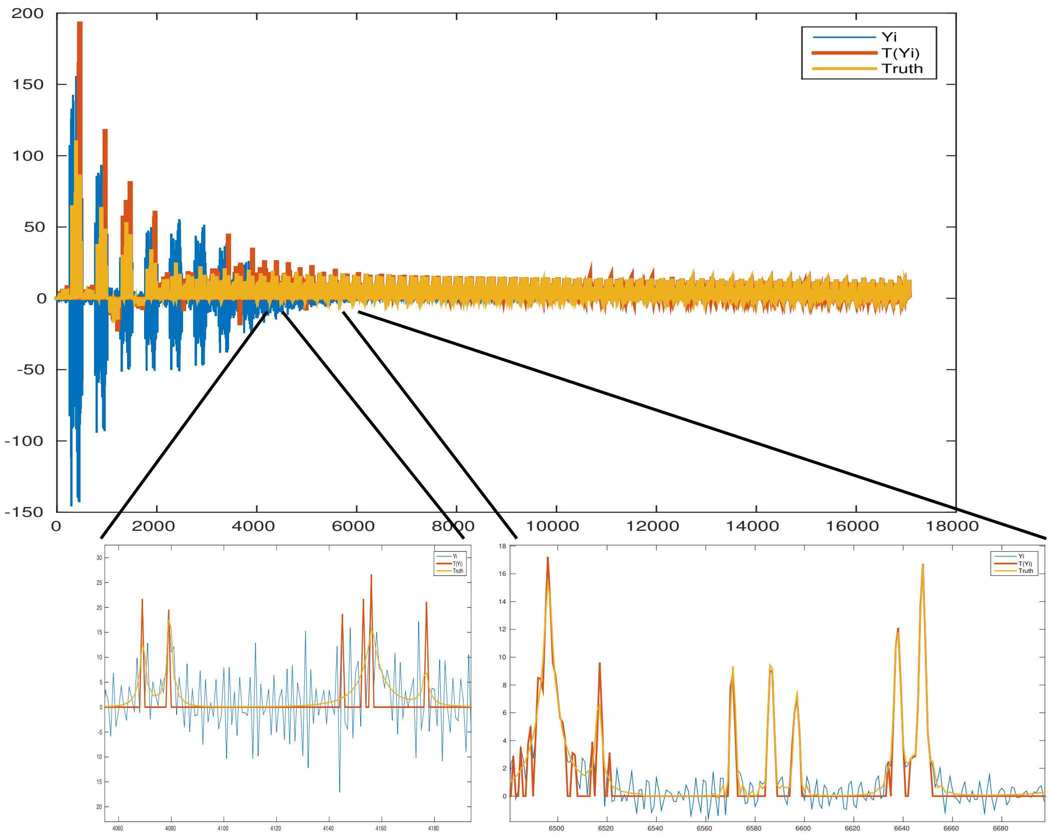


Figure 5.9: Vaguelette coefficients of the data (blue), thresholded and rescaled coefficients (red), and wavelet coefficients of the true signal (yellow). The x axis represents the indices j,k,e, where the left means large j, i.e., small scales.

Notice that the magnitude of the coefficients and the variance of the noise increases with j: this is the obvious consequence of dividing by the singular value κ_j. The explanation of this phenomenon is that the convolution operator distorts the information in the small scales strongly: this results in a larger variance for j large, which makes the coefficients $Y_{j,k,e}/\kappa_j$ for large j (small scales) close to useless for the reconstruction. That is the reason for choosing a scale dependent threshold τ_j. We refer to Example 2 in Chapter 3 for a discussion of this effect.

In the zoom in of Figure 5.9 we see that, while the empirical coefficients are quite noisy, the thresholded coefficients do resemble the true ones. Notice, however, that the coefficients in the small scales are poorly estimated. At this point we want to recall a phenomenon that concerns dictionary thresholding methods: if one chooses too large a threshold, one losses part of the signal and the final reconstruction will have less *mass* than the true function, i.e. it will have smaller magnitude. This can be seen for instance in Figure 5.10, where the WV thresholding estimator reconstructs all modes of the signal correctly, but its magnitude is smaller by a factor of 3 with respect to the true signal.

MIND estimator. We also compute the MIND estimator, which for the deconvolution inverse problem (5.4) is defined as

$$\hat{f}_{\text{MIND}} \in \underset{g}{\operatorname{argmin}} \, \|D^k g\|_{L^2}^2 \quad \text{such that} \quad \max_{\omega \in \Omega_n} \left| \langle \widetilde{\psi}_\omega, Tg - Y \rangle \right| \leq \gamma_n,$$

where the functions $\widetilde{\psi}_\omega$ are indicator functions of intervals ($d = 1$) or rectangles ($d = 2$). For the dictionary of indicator functions we use the same set of scales as indicated above for the multiscale TV-estimator. In the following we show the MIND with $k = 1$, since this gave the best results in simulations. One difference with the multiscale TV-estimator is that here we compute γ_n using the quantiles of the statistic $\max_{\omega \in \Omega_n} |\langle \widetilde{\psi}_\omega, dW \rangle|$.

Risk measures

As in Section 5.1.1, we measure the risk of the estimators with respect to the L^p norm, $p \in \{1, 2, \infty\}$, as well as with respect to the BV seminorm. For images, we also compute their SSIM index (see Section 5.1.1).

5.2.2 Simulation results

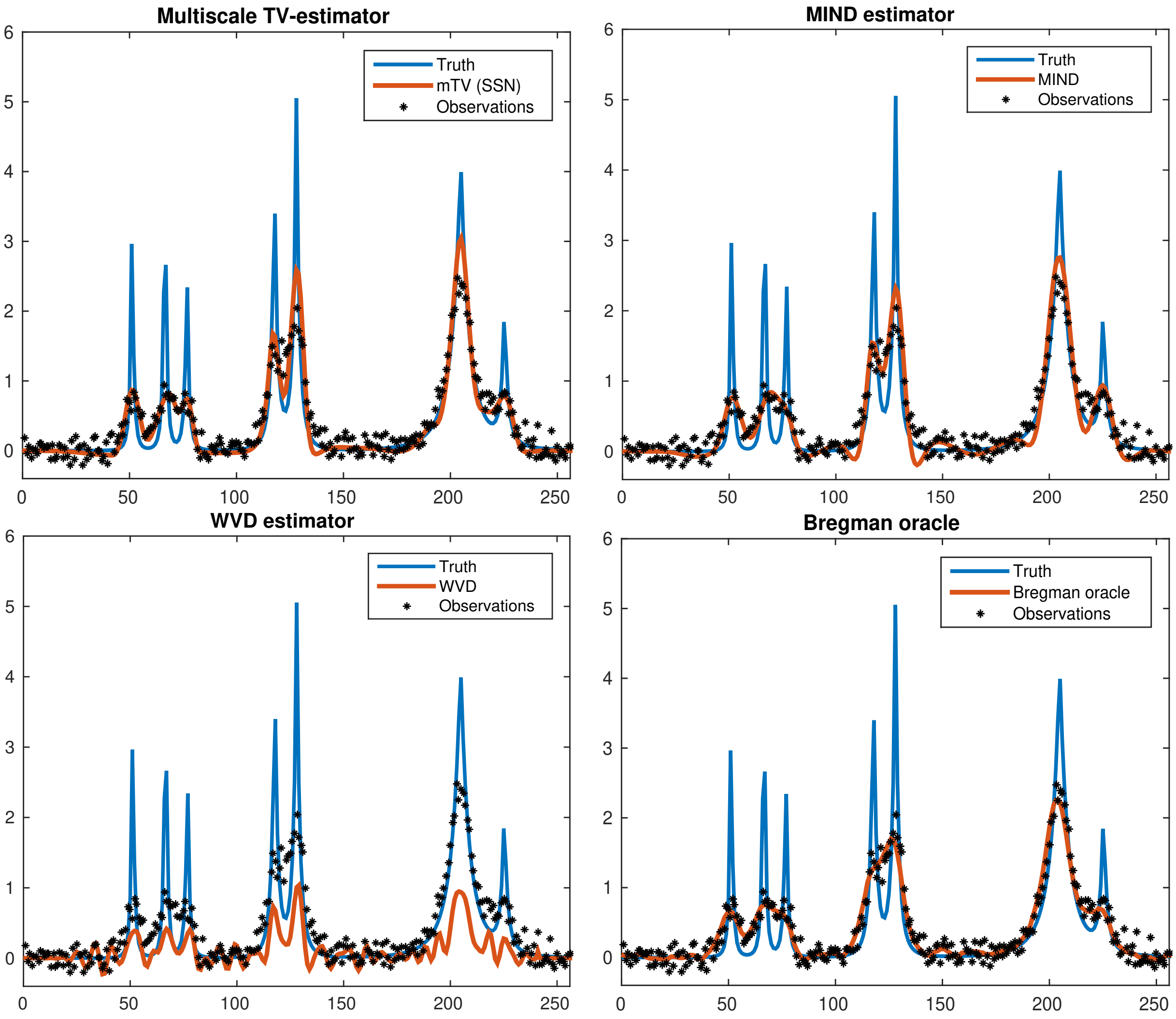
In the following we present severals plots and tables with the reconstructions and risks of different estimators. In one dimensional signals, we compute the multiscale TV-estimator, the MIND estimator with $k = 1$, the wavelet-vaguelette thresholding estimator, and the L^2-TV estimator with oracles λ_{MSE} and λ_{Breg}. See Figures 5.10 and 5.11 for the plots and tables. In two-dimensional images we compute the multiscale TV-estimator, the MIND estimator with $k = 1$, and the L^2-TV estimator with L^2 and Bregman oracles. For the MIND and the multiscale TV-estimator we use an overcomplete system as described above, with the scales j corresponding to blocks of pixels of lengths $\{5, 10, \dots, 40\}$ (Figures 5.12 and 5.13) or of lengths $\{2, 4, \dots, 20\}$ (Figures 5.10,5.11 and 5.14). Notice the corresponding increase in computation time in Figure 5.14.
The results of our simulations, shown in Figures 5.10 to 5.14, can be summarized as follows:

a) Concerning runtime, WV thresholding is the fastest method in $d = 1$, and the L^2-TV and multiscale methods are about one order of magnitude slower. In $d = 2$, however, multiscale methods are between one and two orders of magnitude slower, see Figures 5.13 and 5.14.

b) Concerning the risk measures, the multiscale TV-estimator has the smallest risk in an L^1 and L^2 sense, as well as in term of the SSIM (see Section 5.1.1). The situation is less clear for the L^∞ and BV risks, since here the multiscale TV-estimator, the MIND and the Bregman oracle attain comparable results.

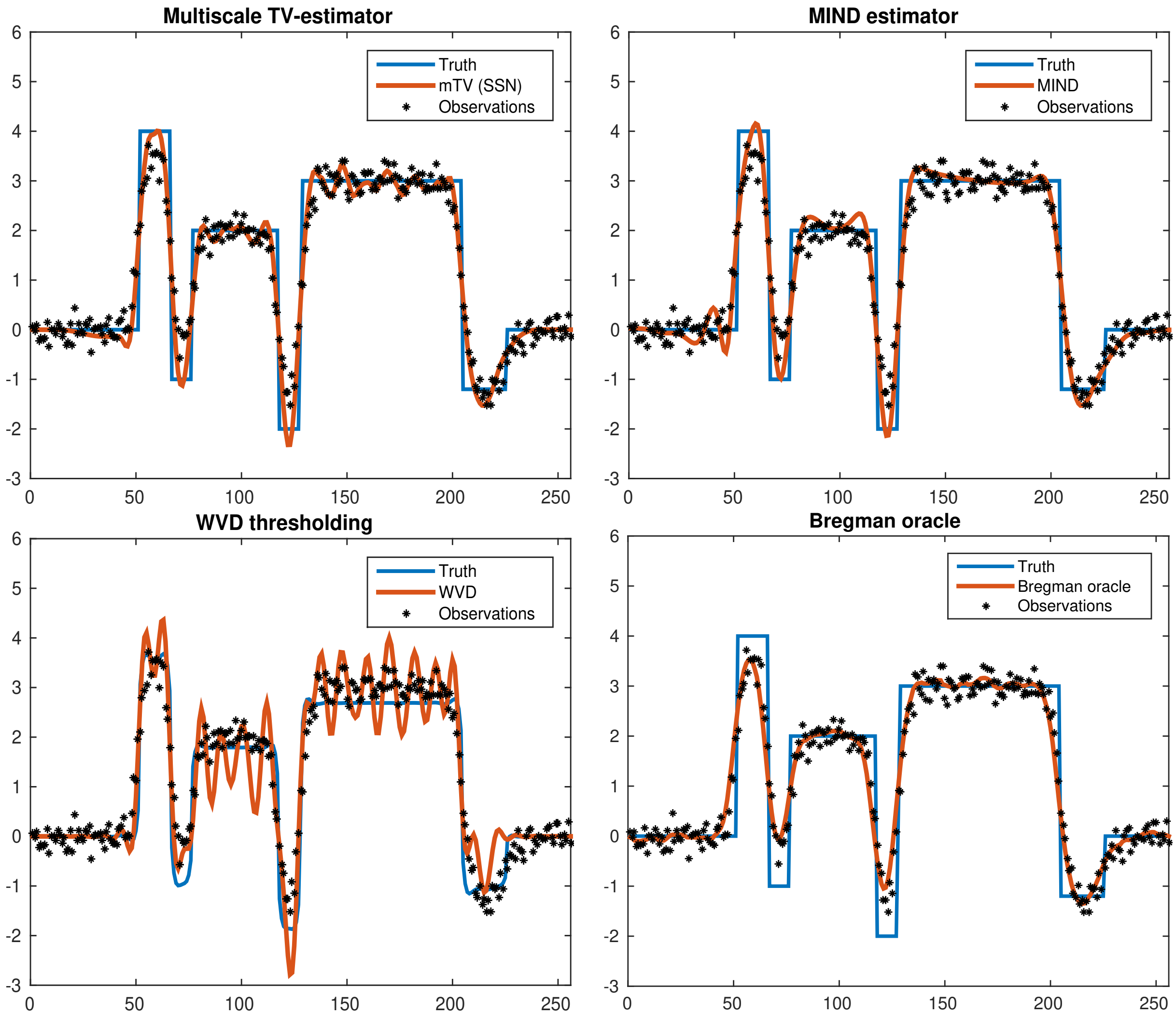
c) In terms of noise, we see that the wavelet-vaguelette thresholding estimator and the L^2 oracle give quite noisy reconstructions with artifacts, while the MIND, the multiscale TV-estimator and the Bregman oracle suppress the noise properly. On the other hand, the L^2 oracle resolves most details in Figure 5.13, while the other methods miss them.

d) Concerning the level of detail in the reconstruction, it is apparent that the MIND does not deconvolve the images properly, thus missing the information in the small scales. On the other extreme, the L^2 oracle gives a quite noisy reconstruction (see e.g. Figure 5.14), while the multiscale-TV and the Bregman oracle present a compromise between regularization, denoising and reconstruction, see e.g. the multiscale TV-reconstructions in Figures 5.12 and 5.14.



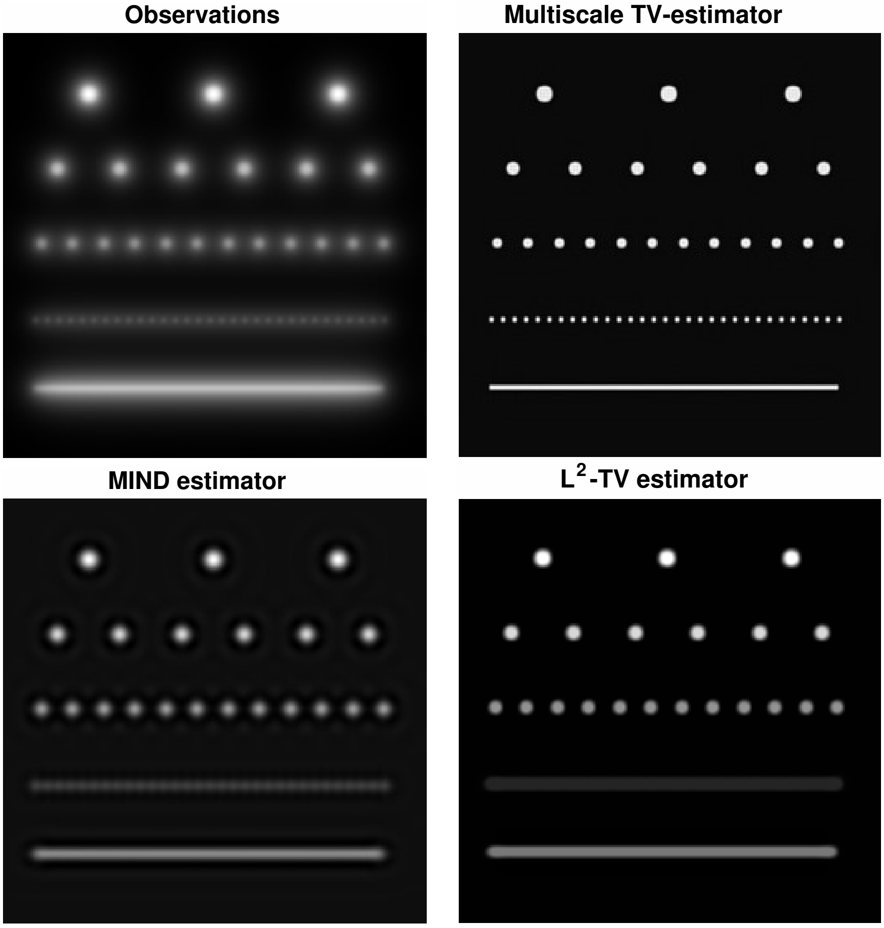
Error Methods	Time (s)	L^2 error	L^1 error	L^∞ error	BV error
Multiscale TV (SSN)	25.84	**3.00**	**18.11**	**1.53**	**15.59**
MIND ($k = 1$)	34.65	6.71	55.96	3.17	31.43
VWD threshold.	**3.59**	10.24	81.66	4.22	35.22
L^2-TV with λ_{MSE}	23.86	7.43	56.22	3.62	30.43
L^2-TV with λ_{Breg}	25.07	7.96	62.36	3.75	31.30

Figure 5.10: Bumps function convolved with kernel with $b = 6.4$ and $\beta = 2$, with $n = 256$ observations and $\sigma = 0.05 \|f\|_{L^\infty}$.



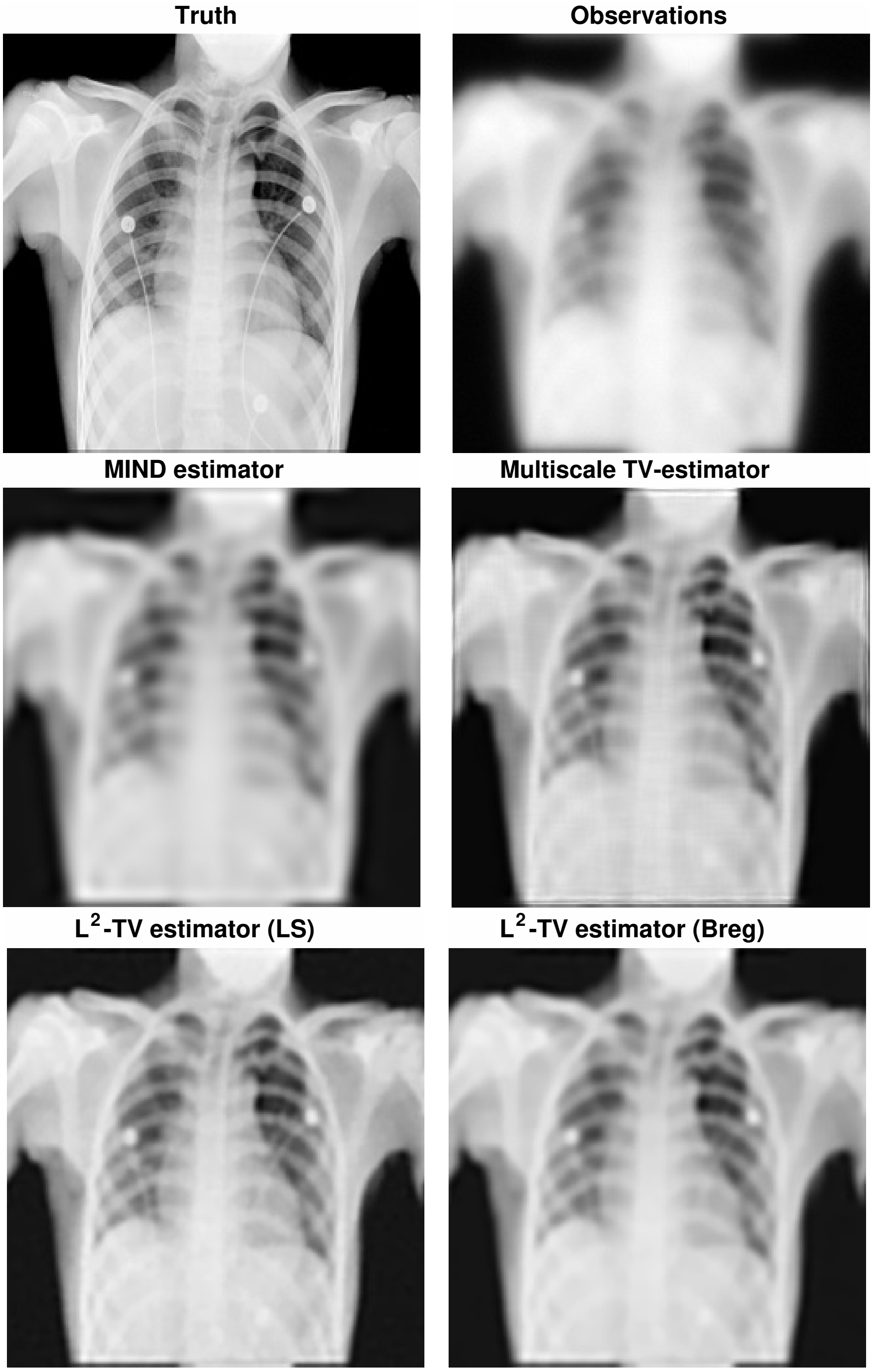
Error / Methods	Time (s)	L^2 error	L^1 error	L^∞ error	BV error
Multiscale TV (SSN)	26.30	**8.21**	**68.43**	2.18	41.84
MIND ($k = 1$)	30.07	8.53	69.97	**2.15**	42.14
VWD threshold.	**4.88**	9.53	99.85	2.19	64.35
L^2-TV with λ_{MSE}	9.45	12.33	119.43	2.69	39.51
L^2-TV with λ_{Breg}	22.62	10.86	100.09	2.61	**37.95**

Figure 5.11: Blocks function convolve with kernel with $b = 6.4$ and $\beta = 2$, with $n = 256$ observations and $\sigma = 0.05\,\|f\|_{L^\infty}$.



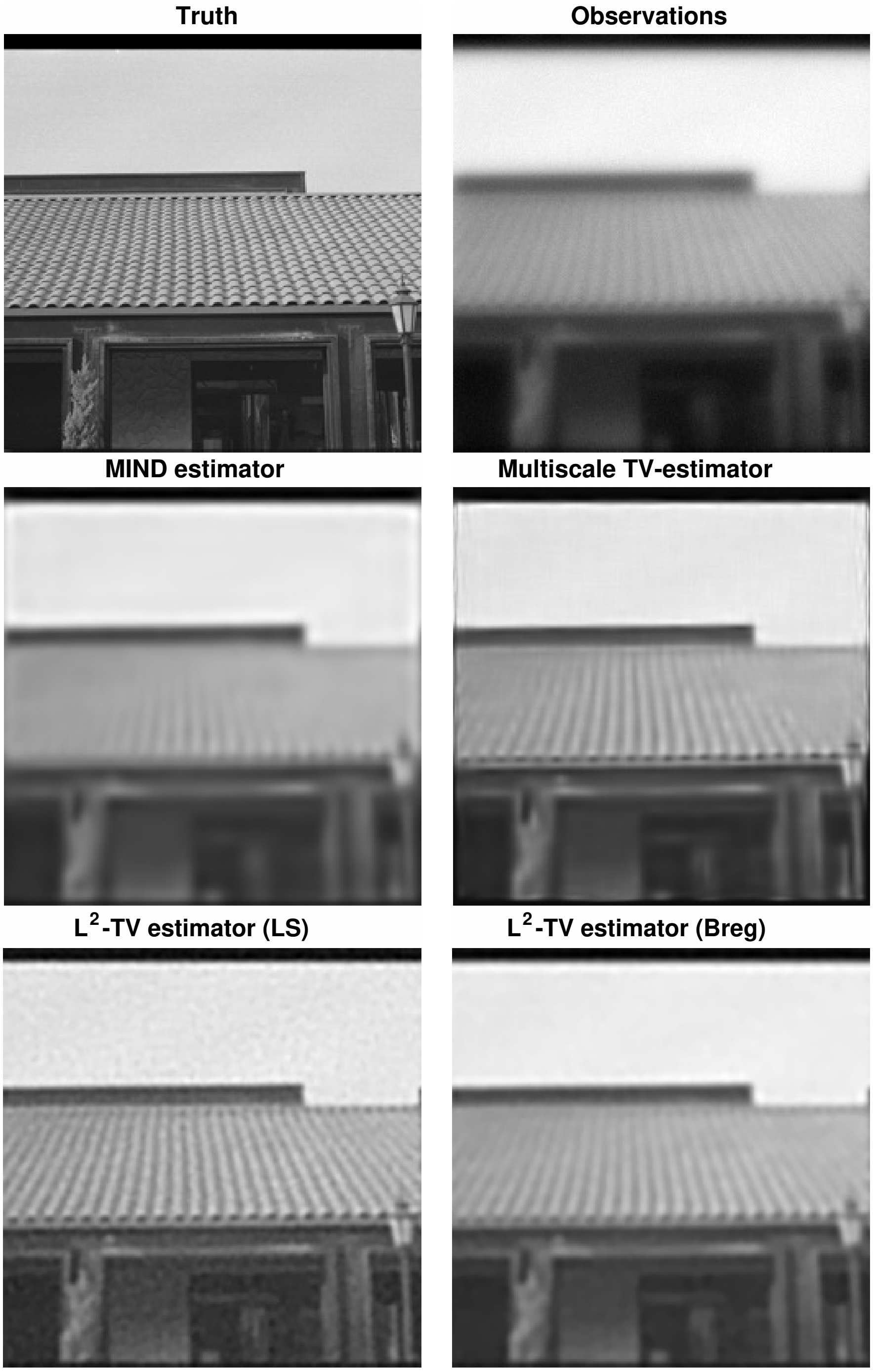
Error / Methods	Time (s)	L^2 error	L^1 error	L^∞ error	BV error	SSIM
Multiscale TV	571.45	**$4.0 \cdot 10^{-2}$**	**$9.8 \cdot 10^{-2}$**	**0.13**	**0.21**	**0.99**
MIND estimator	589.45	0.50	1.03	0.79	1.30	0.86
L^2-TV with λ_{MSE}	**42.90**	0.62	0.85	0.87	1.03	0.88

Figure 5.12: Test image of size $n = 256 \times 256$ convolved with kernel of width $b = 2.56$ and $\beta = 2$ with noise of standard deviation $\sigma = 2^{-8} \|f\|_{L^\infty}$. We show relative errors, i.e., $\|\hat{f} - f\|_{L^p} / \|f\|_{L^p}$.



Error Methods	Time (s)	L^2 error	L^1 error	L^∞ error	BV error	SSIM
Multiscale TV	406.27	$\mathbf{2.8 \cdot 10^{-2}}$	$\mathbf{3.5 \cdot 10^{-2}}$	**0.36**	**0.73**	**0.92**
MIND estimator	172.83	$4.4 \cdot 10^{-2}$	$5.2 \cdot 10^{-2}$	0.47	0.81	0.86
L^2-TV with λ_{MSE}	**41.23**	$3.8 \cdot 10^{-2}$	$4.6 \cdot 10^{-2}$	0.49	**0.73**	0.90
L^2-TV with λ_{Breg}	44.61	$3.9 \cdot 10^{-2}$	$4.7 \cdot 10^{-2}$	0.50	**0.73**	0.89

Figure 5.13: Test image of size $n = 256 \times 256$ convolved with kernel of width $b = 12.8$ and $\beta = 2$ with noise of standard deviation $\sigma = 2^{-8}\, \|f\|_{L^\infty}$. We show relative errors, i.e., $\|\hat{f} - f\|_{L^p} / \|f\|_{L^p}$.



Error Methods	Time (s)	L^2 error	L^1 error	L^∞ error	BV error	SSIM
Multiscale TV	2672	$\mathbf{5.4 \cdot 10^{-2}}$	$\mathbf{9.0 \cdot 10^{-2}}$	0.67	1.02	**0.72**
MIND estimator	1416	$6.4 \cdot 10^{-2}$	$10.3 \cdot 10^{-2}$	**0.46**	0.99	0.65
L^2-TV with λ_{MSE}	50.14	$7.6 \cdot 10^{-2}$	$9.4 \cdot 10^{-2}$	0.52	0.99	0.71
L^2-TV with λ_{Breg}	**35.99**	$7.9 \cdot 10^{-2}$	$9.8 \cdot 10^{-2}$	0.48	**0.95**	0.68

Figure 5.14: Test image of size $n = 256 \times 256$ convolved with kernel of width $b = 12.8$ and $\beta = 2$ with noise of standard deviation $\sigma = 2^{-7}\,\|f\|_{L^\infty}$, where $\|f\|_{L^\infty} = 245$. We show relative errors, i.e., $\|\hat{f} - f\|_{L^p}/\|f\|_{L^p}$.

CHAPTER 6

Conclusion and outlook

In this thesis we have considered the estimation of functions of bounded variation in white noise regression and inverse problems. For that, we have constructed estimators that combine variational regularization with multiscale dictionaries. This type of estimators have been highly successful in practice, but they lack theoretical guarantees. Under suitable assumptions on the dictionaries, which essentially amount to certain approximation properties and a compatibility with the forward operator, we have shown that these estimators attain the optimal rates of convergence in a minimax sense up to logarithmic factors for estimating BV functions. We have also presented two numerical methods for computing the proposed estimators, which we have illustrated in simulations.

The main theoretical contribution of this work is the proof that the proposed estimators are nearly minimax optimal for estimating BV functions in any dimension. Indeed, until now only results in dimension $d = 1$ were known. Our contribution is hence of practical relevance, since BV functions are routinely used in applications from medical imaging to geology and astronomy, and we now give the first theoretical guaranties for using BV functions in such applications. Furthermore, our analysis covers variational estimators based on multiscale dictionaries, which have been proposed before and shown to perform excellently in applications, but for which no theoretical guaranty was available for estimation over BV. We have now closed this gap, proving that these estimators are nearly minimax optimal.

At a technical level, our main contribution is to relate the multiscale data fidelity associated to a multiscale dictionary to a Besov norm of negative smoothness. Even though multiscale constraints have been used in the past, this connection is novel and opens the door to a number of tools from harmonic analysis that enable an analysis of our estimators. Moreover, this link allows us to precisely characterize the conditions on the multiscale dictionaries that guarantee minimax optimality.

In the context of inverse problems, the multiscale dictionary has to be adapted to the forward operator, and the right concept to consider is the wavelet-vaguelette decomposition of the operator, introduced by Donoho (1995). Combining these dictionaries with bounded variation regularization yields nearly minimax optimal estimators for a range of mildly ill-posed inverse problems of any finite degree of ill-posedness in any dimension.

A second theoretical contribution of this work is the analysis of the minimax risk over BV and, more generally, over Besov spaces $B^s_{p,t}$ with $s \leq d/p$. For BV we have shown that the L^q minimax risk in the regression setting undergoes a sharp transition: for $q \leq 1 + 2/d$, the polynomial risk is $n^{-1/(d+2)}$, while for $q \geq 1 + 2/d$ it behaves as $n^{-1/(dq)}$. In particular, the rate deteriorates for larger q, and there is no L^∞-consistent estimator of BV functions. Analogous results have been proven before for estimation over anisotropic Nikolskii classes, that in the isotropic case correspond to $B^s_{p,\infty}$ (Lepskii, 2015). Our results now describe the minimax rates over the whole scale of Besov spaces. Interestingly, the proof of the minimax lower bound in the regime $s \leq d/p$ shows that, in that setting, the most challenging functions to estimate are of "multiscale" type, consisting of blocks of signals at different locations and scales. This suggests that, in that regime, only estimators that incorporate multiscale information can be optimal. Indeed, the only estimators known to be optimal there are a kernel estimator with spatially varying bandwidth proposed by Lepskii (2015), and the multiscale TV-estimator studied in this thesis.

Finally, a practical contribution of this thesis concerns the efficient numerical computation of the multiscale TV-estimator. We propose two methods: one is based on the Chambolle-Pock primal-dual algorithm, which takes advantage of the fact that, in the dual formulation, the multiscale constraint has the form of soft-thresholding, which can be applied efficiently even in high dimensions. The other method is based on a semismooth Newton iteration applied to a sequence of regularized problems with decreasing amount of regularization, using the path-following technique. Both methods perform well in practice, giving locally constant and spatially adaptive reconstructions. With these computation methods, we compared the multiscale TV-estimator with other estimation methods in simulations in regression and deconvolution. The overall conclusion is that the multiscale TV-estimator gives good results both in terms of quantitative risk measures and of visual quality.

We conclude with the discussion of several directions in which the work of this thesis can be extended.

(a) **Dependent data:** We have developed a theory for estimation in a Gaussian white noise model. However, correlated noise is present in many applications. From a modeling perspective, one could consider the observational model (1.1) with fractional Gaussian noise instead of white noise. In that case, the construction of the multiscale TV-estimator could be modified in two ways: either by changing the multiscale constraint to take the correlations into account (here the multiple testing interpretation discussed in the Introduction could be useful), or by adapting the threshold γ_n. For the threshold, the interpretation of γ_n as a quantile, as discussed in the Introduction, could be needed.

(b) **Other noise models:** More generally, one might be willing to drop the assumption that the noise is Gaussian. A quite general class of noise models that could be considered is Lévy processes, which contain Gaussian and Poisson processes as special cases. In this case we would also have to adapt the multiscale constraint in the way indicated above. We expect the estimation rate in these models to be in general different from the rate under Gaussian noise, as it is known that the rate of convergence depends on the tail decay of the noise distribution (see e.g. Han and Wellner (2019) and Section 3.4 in van der Vaart and Wellner (1996)). Finally, staying in the Gaussian setting, the extension to SDE-based models (see e.g. Gobet et al. (2004)) is also of interest.

(c) **Variants of the multiscale constraint:** From the perspective of multiple hypothesis testing, our multiscale data-fidelity term corresponds to a test statistic. However, in order to perform optimally in testing nonparametric hypotheses, it is typically necessary to introduce additional weights, which would modify our constraint to

$$\max_{\omega\in\Omega_n}\Big(a_\omega\big|\langle\phi_\omega, g\rangle - \langle\phi_\omega, dY\rangle\big| - b_\omega\Big)$$

for certain weights $\{a_\omega, b_\omega\}$ (see Dümbgen and Spokoiny (2001)). The purpose of these weights is to prevent the maximum to be overly driven by particular terms. We expect that, in our setting, this finer weighting would help us get rid of some of the additional logarithmic factors in the risk bound. On the other hand, the modified data-fidelity term would no longer match a Besov $B^{-d/2}_{\infty,\infty}$ norm exactly, and a different analysis would be needed. We remark that there are modifications of Besov spaces that could be suitable for this setting (see Section 5.2.2 in Giné and Nickl (2015)).

(d) **Exponentially ill-posed inverse problems:** In our analysis of ill-posed inverse problems we have seen that the multiscale constraint on the observations essentially corresponds to a constraint on the Besov $B^{-d/2-\beta}_{\infty,\infty}$ distance between the estimator and the truth, where $\beta \geq 0$ is the degree of ill-posedness of the operator. Exponentially ill-posed inverse problems correspond formally to $\beta = \infty$, and it is not clear how to extend the present approach to that case. A common strategy to deal with exponentially ill-posed inverse problems is to introduce source conditions of logarithmic type (Hohage, 2000). Alternatively, an approach closer to ours was developed by Petsa and Sapatinas (2009), who consider wavelet thresholding estimators for deconvolution with exponentially decaying kernels.

(e) **Alternative risk functionals:** We have proven convergence rates with respect to the L^q-risk, which measures the *global* error made by the estimator. Admittedly, the use of these risk functionals is driven by technical convenience, as they are relatively easy to handle. On the other hand, these risks only take pointwise differences into account and are arguably not well suited to measure similarity between, say, images. In this sense, an interesting research direction would be to construct risk functionals that measure similarity in a more faithful way. In that sense, multiscale risk functionals have been proposed as an alternative quality measure which takes spatial adaptation into account (see e.g. Cai and Low (2005) and Li (2016)). Alternatively, the use of a Wasserstein distance as a risk functional has been proposed recently in the context of density estimation (Weed and Berthet, 2019).

(f) **Computational speed-up:** We have proposed two methods for computing the multiscale TV-estimator: a semismooth Newton method combined with the path-following technique, and the Chambolle-Pock primal-dual method. The semismooth Newton method seems to perform better than the Chambolle-Pock algorithm in images ($d = 2$) in terms of reconstruction error, but it results in a slight oversmoothing. For that reason, it might be preferable to use a primal-dual method (that does not smooth the problem) modified to yield better error control in a shorter runtime. Many such accelerating variants have been proposed, see e.g. Malitsky and Pock (2018) and Luke and Malitsky (2018), and it would be of interest to apply them to our problem.

CHAPTER 7

Proofs

7.1 Proof of the main theorems

We begin with an auxiliary result for the proof of Theorem 4.

Proposition 7. Let $\{\psi_{j,\theta}\}$ and $\{u_{j,\theta}\}$ denote the dictionary and vaguelette system from Assumption 4. For $L > 0$, $n \in \mathbb{N}$, $n \geq e^L$, let $\hat{f}_{\Phi,T}$ denote the estimator (3.4) with parameter γ_n given by (1.9). Then conditionally on the event $\widetilde{\mathcal{A}}_n$ in (3.10) we have

$$(i)\ \ \|\hat{f}_{\Phi,T} - f\|_{B^{-d/2-\beta}_{\infty,\infty}} \leq C\,\gamma_n + C\frac{\|f\|_{L^\infty} + \log n}{\sqrt{n}},$$
$$(ii)\ \ \|\hat{f}_{\Phi,T} - f\|_{BV} \leq \|f\|_{L^\infty} + 2|f|_{BV} + \log n,$$

for any $f \in BV_L$, and a constant $C > 0$ independent of n, f and $\hat{f}_{\Phi,T}$.

Proof. For part (i), recall that the dictionary $\{\psi_{j,\theta}\}$ satisfies Assumption 4, so we can bound the Besov $B^{-d/2-\beta}_{\infty,\infty}$ norm as

$$\begin{aligned}\|\hat{f}_{\Phi,T} - f\|_{B^{-d/2-\beta}_{\infty,\infty}} &\leq \max_{(j,\theta)\in\Omega_n} 2^{-\beta j}|\langle\psi_{j,\theta}, \hat{f}_{\Phi,T} - f\rangle| + C\,\|\hat{f}_{\Phi,T} - f\|_{L^\infty}\, n^{-1/2}\\ &= \max_{(j,\theta)\in\Omega_n} 2^{-\beta j}|\kappa_j^{-1}\langle T^* u_{j,\theta}, \hat{f}_{\Phi,T} - f\rangle| + C\,\|\hat{f}_{\Phi,T} - f\|_{L^\infty}\, n^{-1/2}\\ &= \max_{(j,\theta)\in\Omega_n} |\langle u_{j,\theta}, T\hat{f}_{\Phi,T} - Tf\rangle| + C\,\|\hat{f}_{\Phi,T} - f\|_{L^\infty}\, n^{-1/2},\end{aligned}$$

using $\kappa_j = 2^{-j\beta}$. The first term can be bounded as

$$\begin{aligned}\max_{(j,\theta)\in\Omega_n}\left|\langle u_{j,\theta}, T\hat{f}_{\Phi,T} - Tf\rangle\right| &\leq \underbrace{\max_{(j,\theta)\in\Omega_n}\left|\langle u_{j,\theta}, T\hat{f}_{\Phi,T}\rangle - \langle u_{j,\theta}, dY\rangle\right|}_{\leq\gamma_n} + \max_{(j,\theta)\in\Omega_n}\left|\langle u_{j,\theta}, Tf\rangle - \langle u_{j,\theta}, dY\rangle\right|\\ &\leq \gamma_n + \max_{(j,\theta)\in\Omega_n}\frac{\sigma}{\sqrt{n}}\left|\int_{\mathbb{M}} u_{j,\theta}(x)\,dW(x)\right| \leq 2\gamma_n\end{aligned}$$

conditionally on $\widetilde{\mathcal{A}}_n$, where in the second inequality we used the definition of $\hat{f}_{\Phi,T}$. And the supremum norm $\|\hat{f}_{\Phi,T} - f\|_{L^\infty}$ can be bounded by $\|f\|_{L^\infty} + \log n$ by construction of $\hat{f}_{\Phi,T}$. This completes the proof of (*i*). For (*ii*), we have

$$\|\hat{f}_{\Phi,T} - f\|_{BV} \leq \|\hat{f}_{\Phi,T} - f\|_{L^1} + |\hat{f}_{\Phi,T} - f|_{BV} \leq \|\hat{f}_{\Phi,T} - f\|_{L^\infty} + |\hat{f}_{\Phi,T} - f|_{BV},$$

where we used that $\hat{f}_{\Phi,T}$ and f are supported inside the unit hypercube, whence their L^1 norm is dominated by their L^∞ norm. The first term in the right-hand side is bounded by $\|f\|_{L^\infty} + \log n$, while the second is bounded by $|\hat{f}_{\Phi,T}|_{BV} + |f|_{BV}$. Finally, conditionally on $\widetilde{\mathcal{A}}_n$ we have $|\hat{f}_{\Phi,T}|_{BV} \leq |f|_{BV}$. This is so because $\hat{f}_{\Phi,T}$ is defined as the minimizer of the BV seminorm among the functions satisfying $\max_{(j,\theta)\in\Omega_n} |\langle u_{j,\theta}, Tg\rangle - \langle u_{j,\theta}, dY\rangle| \leq \gamma_n$ and $\|\hat{f}_{\Phi,T}\|_{L^\infty} \leq \log n$. Note that, conditionally on $\widetilde{\mathcal{A}}_n$ and for $n \geq e^L$, the function f satisfies this constraint, and hence f is an admissible function for the minimization problem defining $\hat{f}_{\Phi,T}$, whence $|\hat{f}_{\Phi,T}|_{BV} \leq |f|_{BV}$. This completes the proof. □

We will also need a variant of the interpolation inequality in Theorem 3.

Proposition 8. For $d \in \mathbb{N}$ and $\beta \geq 0$, let $q^* := 1 + 2/(d + 2\beta)$.

a) If $q^* \leq 2$, there is a constant $C > 0$ such that

$$\|g\|_{L^q} \leq C\, \|g\|_{B^{-d/2-\beta}_{\infty,\infty}}^{\frac{2}{d+2\beta+2}} \|g\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}}$$

holds for any $q \in [1, q^*]$ and any $g \in B^{-d/2-\beta}_{\infty,\infty} \cap BV$ with $\operatorname{supp} g \subseteq [0,1]^d$.

b) If $q^* > 2$, then there is a constant $C > 0$ such that for any $n \in \mathbb{N}$ we have

$$\|g\|_{L^q} \leq C(\log n)\, \|g\|_{B^{-d/2-\beta}_{\infty,\infty}}^{\frac{2}{d+2\beta+2}} \|g\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}} + C\, n^{-1}\, \|g\|_{L^\infty}^{\frac{2}{d+2\beta+2}}\, \|g\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}}$$

for any $q \in [1, q^*]$ and any $g \in L^\infty \cap BV$ with $\operatorname{supp} g \subseteq [0,1]^d$.

The reason for the two regimes in the proposition is that Theorem 3 gives a bound on the Besov $B^0_{q^*,q^*}$ norm. If $q^* \leq 2$, then $B^0_{q^*,q^*} \hookrightarrow L^{q^*}$ and the bound can be readily translated to the L^q-risk. If $q^* > 2$, however, the embedding does not hold, and we have to use additional regularity of the functions (being in L^∞) in order to upper bound the L^{q^*}-risk by the $B^0_{q^*,q^*}$-risk. The proof of Proposition 8 is given in Section 7.1.1 below. We are now ready to prove Theorem 4.

Proof of part a) of Theorem 4. We prove the claim of part a) of Theorem 4 conditionally on the event $\widetilde{\mathcal{A}}_n$ in (3.10), which by Proposition 11 in Section 7.3.2 happens with probability $\mathbb{P}(\widetilde{\mathcal{A}}_n) \geq 1 - (\#\Omega_n)^{1-\kappa^2}$.

Consider first the regime $q \le q^* := 1 + 2/(d + 2\beta)$. For $d \ge 2$, part a) of Proposition 8 gives the inequality

$$\|\hat{f}_{\Phi,T} - f\|_{L^q} \le C\, \|\hat{f}_{\Phi,T} - f\|_{B^{-d/2}_{\infty,\infty}}^{\frac{2}{d+2\beta+2}} \, \|\hat{f}_{\Phi,T} - f\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}} \tag{7.1}$$

for any $q \in [1, q^*]$, as long as $q^* \le 2$ which is guaranteed by the assumption that $d \ge 2$ and $\beta \ge 0$. Conditionally on $\widetilde{\mathcal{A}}_n$, Proposition 7 gives bounds for the terms in the right-hand side of (7.1), which altogether yield

$$\begin{aligned}
\|\hat{f}_{\Phi,T} - f\|_{L^q} &\le C\Big(\gamma_n + C\frac{\|f\|_{L^\infty} + \log n}{\sqrt{n}}\Big)^{\frac{2}{d+2\beta+2}} (\|f\|_{L^\infty} + 2|f|_{BV} + \log n)^{\frac{d+2\beta}{d+2\beta+2}} \\
&\le Cn^{-\frac{1}{d+2\beta+2}} (\sqrt{\log \#\Omega_n} + L + \log n)^{\frac{2}{d+2\beta+2}} (L + \log n)^{\frac{d+2\beta}{d+2\beta+2}} \\
&\le C\, n^{-\frac{1}{d+2\beta+2}} \log n
\end{aligned}$$

using that $f \in BV_L$. Since $\#\Omega_n$ grows polynomially in n (recall Assumption 1 in Chapter 2), the last inequality follows.

For the case when $d = 1$ and $\beta \ge 1/2$, we have $q^* \le 2$ and the argument goes through as above. Finally, the case $d = 1$ and $\beta < 1/2$ requires a special treatment, since we have $q^* > 2$ and the embedding $B^0_{q^*,q^*} \hookrightarrow L^{q^*}$ does not hold. However, part b) of Proposition 8 yields the somewhat weaker statement

$$\|g\|_{L^q} \le C\,(\log n)\, \|g\|_{B^{-1/2-\beta}_{\infty,\infty}}^{\frac{2}{d+2\beta+2}} \|g\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}} + C\, n^{-1}\, \|g\|_{L^\infty}^{\frac{2}{d+2\beta+2}} \|g\|_{BV}^{\frac{d+2\beta}{d+2\beta+2}}$$

for $g = \hat{f}_\Phi - f$ and $q \in [1, q^*]$. Proceeding as above, Proposition 7 now implies that, conditionally on $\widetilde{\mathcal{A}}_n$, we have

$$\|\hat{f}_{\Phi,T} - f\|_{L^q} \le C\, n^{-\frac{1}{d+2\beta+2}} \,(\log n)^2 + C\, n^{-1}\, \log n,$$

which yields the claim.

We have proved the claim for the L^q-risk with $q \le 1 + 2/(d + 2\beta)$. For larger q, we use Hölder's inequality between the $L^{1+2/(d+2\beta)}$ and the L^∞-risk, which gives the bound

$$\|\hat{f}_{\Phi,T} - f\|_{L^q} \le \|\hat{f}_{\Phi,T} - f\|_{L^{1+2/(d+2\beta)}}^{\frac{d+2\beta+2}{q(d+2\beta)}} \|\hat{f}_{\Phi,T} - f\|_{L^\infty}^{1-\frac{d+2\beta+2}{q(d+2\beta)}} \le C\, n^{-\frac{1}{q(d+2\beta)}}\, (\log n)^{3-\min\{d,2\}}$$

for $q \ge 1 + 2/(d + 2\beta)$. This completes the proof. □

Proof of part b) of Theorem 4. Using the convergence conditionally on $\widetilde{\mathcal{A}}_n$ proved in part a) of the theorem, we can bound the expected risk for $q \in [1, \infty)$ as

$$\begin{aligned}\mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}] &= \mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}\, 1_{\widetilde{\mathcal{A}}_n}] + \mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}\, 1_{\widetilde{\mathcal{A}}^c_n}] \\ &\leq C\, r_n\, \mathbb{P}(\widetilde{\mathcal{A}}_n) + \mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}\, 1_{\widetilde{\mathcal{A}}^c_n}], \end{aligned} \tag{7.2}$$

where $r_n = n^{-\min\{\frac{1}{d+2\beta+2}, \frac{1}{(d+2\beta)q}\}} (\log n)^{3-\min\{d,2\}}$. We show now that the second term behaves as $o(n^{-1/(d+2\beta+2)})$ for $\kappa^2 > 1 + \frac{1}{(d+2\beta+2)\Gamma}$. Indeed, since the functions f and $\hat{f}_{\Phi,T}$ are supported inside the unit hypercube, we can bound their L^q-norm by their supremum norm. Then we use that $\|f\|_{L^\infty} \leq L$ and that $\|\hat{f}_{\Phi,T}\|_{L^\infty} \leq \log n$ for $n \geq e^L$ by construction, so we have

$$\mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}\, 1_{\widetilde{\mathcal{A}}^c_n}] \leq \mathbb{E}[(L + \log n)\, 1_{\widetilde{\mathcal{A}}^c_n}] \leq (L + \log n)\mathbb{P}(\widetilde{\mathcal{A}}^c_n).$$

By Proposition 11 in Section 7.3.2 we have $\mathbb{P}(\widetilde{\mathcal{A}}^c_n) \leq (\#\Omega_n)^{1-\kappa^2}$. Inserting this back in (7.2) and using that $\#\Omega_n \geq c\, n^\Gamma$ yields

$$\mathbb{E}[\|\hat{f}_{\Phi,T} - f\|_{L^q}] \leq C\, n^{-\min\{\frac{1}{d+2\beta+2}, \frac{1}{(d+2\beta)q}\}} (\log n)^{3-\min\{d,2\}} + C\, n^{(1-\kappa^2)\Gamma} \log n.$$

Choosing $\kappa^2 > 1 + \frac{1}{(d+2\beta+2)\Gamma}$ yields the claim. □

7.1.1 Proof of Proposition 8

For simplicity, we prove the two parts of Proposition 8 separately.

Proof of part a) of Proposition 8. First, Theorem 3 with $s = -d/2 - \beta$ and $p = \infty$ gives

$$\|g\|_{B^0_{q^*,q^*}} \leq C\, \|g\|^{\frac{2}{d+2\beta+2}}_{B^{-d/2-\beta}_{\infty,\infty}} \|g\|^{\frac{d+2\beta}{d+2\beta+2}}_{BV}$$

for any smooth enough g. It remains to show that the L^q-norm, $q \in [1, q^*]$, can be upper bounded by the $B^0_{q^*,q^*}$-norm. But that is indeed the case, due to the continuous embedding

$$B^0_{r,r}(\mathbb{R}^d) \hookrightarrow L^r(\mathbb{R}^d) \tag{7.3}$$

for $r \in (1, 2]$. Indeed, continuity of the embedding follows from Proposition 2 in Section 2.3.2 in Triebel (1983). It states that, for $0 < q \leq \infty$, $0 < p < \infty$ and $s \in \mathbb{R}$, the embedding

$$B^s_{p,\min\{p,q\}}(\mathbb{R}^d) \hookrightarrow F^s_{p,q}(\mathbb{R}^d)$$

is continuous. Moreover, equation (2) in Section 2.3.5 in Triebel (1983) states that

$$F^0_{p,2}(\mathbb{R}^d) = L^p(\mathbb{R}^d)$$

for $p \in (1, \infty)$. These two facts imply that

$$B^0_{r,r}(\mathbb{R}^d) = B^0_{r,\min\{r,2\}}(\mathbb{R}^d) \hookrightarrow F^0_{r,2}(\mathbb{R}^d) = L^r(\mathbb{R}^d) \quad \forall r \in (1, 2],$$

which completes the proof of (7.3). The extension to the L^1-risk follows by compact support. □

The proof of part b) of Proposition 8 relies on the following result.

Proposition 9. Let $g \in L^\infty \cap BV$ satisfy supp $g \subseteq [0, 1]^d$, and let $q \in [2, 3]$. Then for any $J \in \mathbb{N}$ we have

$$\|g\|_{L^q} \le C\, J\, \|g\|_{B^0_{q,q}} + C\, 2^{-J/q} \|g\|_{L^\infty}^{1-1/q} \|g\|_{BV}^{1/q}$$

for a constant $C > 0$ independent of g.

The proof of Proposition 9 uses the following lemma.

Lemma 1. Let $\{\psi_{j,k,e} \,|\, (j, k, e) \in \Omega\}$ denote a basis of compactly supported wavelets in $L^2(\mathbb{R}^d)$. For any $q \in [2, 3]$ there is a constant $C_{\psi,q}$ such that

$$\int_{\mathbb{R}^d} \Bigg| \sum_{(k,e)\in P^d_j \times E_j} c_{j,k,e} \psi_{j,k,e}(x) \Bigg|^q dx \le C_{\psi,q}\, 2^{jqd(1/2-1/q)} \sum_{(k,e)\in P^d_j \times E_j} |c_{j,k,e}|^q$$

for any $j \in \mathbb{N}$ and any coefficients $\{c_{j,k,e}\}$, where

$$P^d_j := \{k \in \mathbb{Z}^d \,|\, (j, k, e) \in \Omega, \quad \text{supp}\, \psi_{j,k,e} \cap (0, 1)^d \neq \emptyset\}.$$

Proof of Lemma 1. We prove the lemma by showing the extreme cases $q = 2$ and $q = 3$, and then applying the Riesz-Thorin interpolation theorem (see e.g. Stein and Weiss (1971)) to the bounded operator

$$\begin{aligned} A_j : \ell^q(P^d_j \times E_j) &\to L^q(\mathbb{R}^d) \\ \{c_{j,k,e}\}_{(k,e)\in P^d_j \times E_j} &\mapsto \sum_{(k,e)\in P^d_j \times E_j} c_{j,k,e} \psi_{j,k,e}. \end{aligned}$$

This gives the claim for all $q \in [2, 3]$.
Notice that the claim for $q = 2$ follows by the orthonormality of the wavelet basis.

For $q = 3$, we begin with an observation. Due to the compact support of the wavelets, there is a constant c_ψ such that, for each $j \geq 0$ and $(k, e) \in P_j^d \times E_j$, at most c_ψ wavelets at scale j have support intersecting the support of $\psi_{j,k,e}$, i.e.,

$$\max_{(j,k,e)\in\mathbb{N}\times P_j^d\times E_j} \#\{(k', e') \in P_j^d \times E_j \,|\, \operatorname{supp} \psi_{j,k,e} \cap \operatorname{supp} \psi_{j,k',e'} \neq \emptyset\} \leq c_\psi. \tag{7.4}$$

For instance, for Daubechies wavelets with S continuous partial derivatives we can take $c_\psi = 2^d\,(12\,S + 1)^d$ (see Section 2.1). As a consequence, we have the following inequalities

$$\begin{aligned}
\int_{\mathbb{R}^d} \Big| \sum_{(k,e)\in P_j^d\times E_j} c_{j,k,e}\psi_{j,k,e}(x)\Big|^3 dx &= \sum_{(k,e)\in P_j^d\times E_j} \int_{\mathbb{R}^d} \big|c_{j,k,e}\psi_{j,k,e}(x)\big|^3 dx \\
&\quad + 3 \sum_{(k,e)\neq(k',e')} \int_{\mathbb{R}^d} \big|c_{j,k,e}\psi_{j,k,e}(x)\big|^2 \big|c_{j,k',e'}\psi_{j,k',e'}(x)\big| dx \\
+ 6 \sum_{(k,e)\neq(k',e')\neq(k'',e'')} &\int_{\mathbb{R}^d} \big|c_{j,k,e}\psi_{j,k,e}(x)\big|\big|c_{j,k',e'}\psi_{j,k',e'}(x)\big|\big|c_{j,k'',e''}\psi_{j,k'',e''}(x)\big| dx \\
&\leq (1 + 3c_\psi + 6c_\psi^2) \sum_{(k,e)\in P_j^d\times E_j} |c_{j,k,e}|^3 \|\psi_{j,k,e}\|_{L^3}^3 \\
&= (1 + 3c_\psi + 6c_\psi^2)\, \|\psi\|_{L^3}^3\, 2^{j3d(1/2-1/3)} \sum_{(k,e)\in P_j^d\times E_j} |c_{j,k,e}|^3
\end{aligned} \tag{7.5}$$

where we used (7.4) to bound the number of summands, and in the last equality we used that $\|\psi_{j,k,e}\|_{L^3} = 2^{jd(1/2-1/3)}\,\|\psi\|_{L^3}$. The inequality is justified as follows. By Young's inequality and the support properties of $\psi_{j,k,e}$ we have

$$\begin{aligned}
&\sum_{(k,e)\neq(k',e')} \int_{\mathbb{R}^d} \big|c_{j,k,e}\psi_{j,k,e}(x)\big|^2 \big|c_{j,k',e'}\psi_{j,k',e'}(x)\big| dx \\
&\qquad\leq \sum_{(k,e)\neq(k',e')} \int_{\mathbb{R}^d} \frac{2}{3}\big|c_{j,k,e}\psi_{j,k,e}(x)\big|^3 + \frac{1}{3}\big|c_{j,k',e'}\psi_{j,k',e'}(x)\big|^3 dx \\
&\qquad\leq \frac{2}{3}\, c_\psi \sum_{(k,e)} \int_{\mathbb{R}^d} \big|c_{j,k,e}\psi_{j,k,e}(x)\big|^3 dx + \frac{1}{3}\, c_\psi \sum_{(k',e')} \int_{\mathbb{R}^d} \big|c_{j,k',e'}\psi_{j,k',e'}(x)\big|^3 dx \\
&\qquad= c_\psi \sum_{(k,e)} \big|c_{j,k,e}\big|^3 \|\psi_{j,k,e}\|_{L^3}^3.
\end{aligned}$$

The same argument with Young's inequality gives the desired bound for the product of 3 terms in (7.5). This completes the proof. □

Proof of Proposition 9. Let $\{\psi_{j,k,e}\}$ be a basis of compactly supported wavelets. Writing g formally as its wavelet series we have for any $q \in [2,3]$

$$\|g\|_{L^q} = \Big\| \sum_{j\in\mathbb{N}} \sum_{k,e} c_{j,k,e}\psi_{j,k,e} \Big\|_{L^q} \le \Big\| \sum_{j\le J} \sum_{k,e} c_{j,k,e}\psi_{j,k,e} \Big\|_{L^q} + \Big\| \sum_{j>J} \sum_{k,e} c_{j,k,e}\psi_{j,k,e} \Big\|_{L^q} \tag{7.6}$$

for any $J \in \mathbb{N}$. Since $\operatorname{supp} g \subseteq [0,1]^d$, the sums are over $(k,e) \in P_j^d \times E_j$. Using Lemma 1, the first term can be bounded as

$$\begin{aligned}
\Big\| \sum_{j\le J} \sum_{k,e} c_{j,k,e}\psi_{j,k,e} \Big\|_{L^q} &\le \sum_{j\le J} \Big(C_{\psi,q} 2^{jqd(1/2-1/q)} \sum_{(k,e)} |c_{j,k,e}|^q \Big)^{1/q} \\
&\le C_{\psi,q}^{1/q}\, J \Big(\max_{j\le J} 2^{jqd(1/2-1/q)} \sum_{(k,e)} |c_{j,k,e}|^q \Big)^{1/q} \\
&\le C_{\psi,q}^{1/q}\, J\, \|g\|_{B^0_{q,q}},
\end{aligned}$$

which gives the first term of the claim. For the second term, we use that $g \in L^\infty$ and $g \in BV$, which means that the wavelet coefficients of g satisfy the bounds

$$\max_{(k,e)\in P_j^d\times E_j} |c_{j,k,e}| \le 2^{-jd/2}\, \|g\|_{L^\infty} \quad \text{and} \quad \sum_{(k,e)\in P_j^d\times E_j} |c_{j,k,e}| \le 2^{j(d/2-1)}\, \|g\|_{BV},$$

for any $j \in \mathbb{N}$, where the first inequality follows from the compact support of the wavelets and Hölder's inequality, and the second follows from the embedding $BV \subset B^1_{1,\infty}$. Using Lemma 1 and these bounds, the second term in (7.6) can be bounded as

$$\begin{aligned}
\Big\| \sum_{j>J} \sum_{k,e} c_{j,k,e}\psi_{j,k,e} \Big\|_{L^q} &\le \sum_{j>J} \Big(C_{\psi,q} 2^{jqd(1/2-1/q)} \sum_{(k,e)\in P_j^d\times E_j} |c_{j,k,e}|^q \Big)^{1/q} \\
&\le C_{\psi,q}^{1/q} \sum_{j>J} \Big(2^{jqd(1/2-1/q)}\, 2^{-jd(q-1)/2}\, \|g\|_{L^\infty}^{q-1} 2^{j(d/2-1)} \|g\|_{BV} \Big)^{1/q} \\
&\le C_{\psi,q}^{1/q}\, \|g\|_{L^\infty}^{1-1/q}\, \|g\|_{BV}^{1/q} \sum_{j>J} 2^{-j/q},
\end{aligned}$$

which gives the claim. □

Proof of part b) of Proposition 8. Let $q^* := 1 + 2/(d+2\beta)$ and assume that $q^* > 2$. Notice that $q^* \le 3$ for $d \in \mathbb{N}$ and $\beta \ge 0$. The claim follows from Theorem 3 with $s = -d/2 - \beta$ and $p = \infty$, which gives a bound on the $B^0_{q^*,q^*}$ norm. The L^q-norm, $q \in [1, q^*]$, can be upper bounded by the L^{q^*}-norm, which itself can be upper bounded by the $B^0_{q^*,q^*}$ norm using Proposition 9 below. Choosing $J = \lceil q^* \log n \rceil$ yields the claim. □

7.2 Proof of the minimax lower bounds

Here we prove Theorem 6. The proofs of the lower bounds in the regimes $q < p\frac{d+2s+2\beta}{d+2\beta}$ (dense case) and $q \geq p\frac{d+2s+2\beta}{d+2\beta}$ and $s > d/p$ (sparse case) are well-known, and can be found e.g. in Chapter 10 of Härdle et al. (2012) for $d = 1$ and $T = id$, so we do not reproduce them here. Indeed, the generalization from $d = 1$ to $d \geq 2$ is trivial. Concerning the generalization to inverse problems, we show below how to adapt the construction of the alternatives in the "multiscale" regime $s < d/p$, which indicates how to proceed in the other regimes (see e.g. Theorem 3 in Cavalier (2011) for a different strategy for computing the minimax risk in inverse problems for the L^2-risk).

On the other hand, the regime $q \geq p\frac{d+2s}{d}$ and $s \leq d/p$ is far less popular, so we give the complete proof of the lower bound here. The proof follows the same idea as in the other regimes: we reduce the estimation problem to a testing problem, and construct a set of alternatives that cannot be perfectly distinguished by any statistical procedure. As in the dense regime, our construction is based on Assouad's cube (Assouad, 1983).

Proof of Theorem 6. Our proof follows the proof of Theorem 10.3 in Härdle et al. (2012) closely. We structure it in several steps. For conceptual simplicity we first give the proof for $T = id$, and then consider general T.

Construction of alternatives: Let $g_0 \in B^s_{p,t} \cap L^\infty$ satisfy

$$\|g_0\|_{B^s_{p,t}} \leq L/2, \quad \text{and} \quad \|g_0\|_{L^\infty} \leq L/2.$$

Let $\psi_{j,k,e}$ be a basis of Daubechies wavelets with S continuous partial derivatives, where $S > \max\{s, d/2\}$. For $j \geq 0$ to be fixed later, let $R_j \subseteq \{0, \ldots, 2^j - 1\}^d \times E_j$ denote a subset of wavelet indices such that

$$\text{supp}\, \psi_{j,k,e} \cap \text{supp}\, \psi_{j,k',e'} = \emptyset \quad \text{for } (k,e) \neq (k',e') \in R_j.$$

Since Daubechies wavelets are compactly supported, we have $\#R_j \leq c2^{jd}$ for a constant $c > 0$. Let further $S_j = \#R_j = \lfloor 2^{j\Delta} \rfloor$ for a real number $\Delta \in [0, d]$ to be chosen later. Consider now vectors $\epsilon \in \{-1, +1\}^{S_j}$ with components indexed by $(k, e) \in R_j$. Our alternatives will have the form

$$g^\epsilon := g_0 + \gamma \sum_{(k,e)\in R_j} \epsilon_{k,e}\psi_{j,k,e}$$

for $\gamma > 0$ to be chosen later. Define the set $\mathcal{G} := \{g^\epsilon \,|\, \epsilon \in \{-1, +1\}^{S_j}\}$. Notice that all functions in this set satisfy

$$\|g^{\epsilon}\|_{B^s_{p,t}} \leq L \quad \text{and} \quad \|g^{\epsilon}\|_{L^\infty} \leq L$$

provided that

$$\gamma \leq \frac{L}{2} 2^{-j(s+d(\frac{1}{2}-\frac{1}{p})+\frac{\Delta}{p})} \quad \text{and} \quad \gamma \leq \frac{L}{2\,\|\psi\|_{L^\infty}} 2^{-jd/2}, \tag{7.7}$$

respectively. In the following we choose Δ in order to balance these two terms, i.e., $\Delta = d - ps$. Notice that in the sparse regime one chooses $\Delta = 0$, while in the dense regime it is enough to choose $\Delta = d$. The regime we consider here is a middle ground in which the difficulty of the problem is encoded by signals g^{ϵ} that have many spikes at each scale, but are not dense.

In this setting, the L^q-separation between these alternatives is

$$\delta := \inf_{\epsilon \neq \epsilon'} \|g^{\epsilon} - g^{\epsilon'}\|_{L^q} = 2\|\gamma \psi_{j,k,e}\|_{L^q} = 2\gamma\, 2^{jd(\frac{1}{2}-\frac{1}{q})} \, \|\psi\|_{L^q}, \tag{7.8}$$

where the first equality follows from the disjoint supports of the wavelets.

Lower bound: We use now Assouad's lemma for lower bounding the L^q-risk over $(B^s_{p,t} \cap L^\infty)_L$. We reproduce the claim (Lemma 10.2 in Härdle et al. (2012)) for completeness.

Lemma 2. For $\epsilon \in \{-1,+1\}^{S_j}$ and $(k,e) \in R_j$, define $\epsilon_{*k,e} := (\epsilon'_{(k_1,e_1)}, \dots, \epsilon'_{(k_{S_j}, e_{S_j})})$, where

$$\epsilon'_{(k'e')} = \begin{cases} \epsilon_{(k,e)} & \text{if } (k',e') \neq (k,e), \\ -\epsilon_{(k,e)} & \text{if } (k',e') = (k,e). \end{cases}$$

Assume there exist constants $\lambda, p_0 > 0$ such that

$$\mathbb{P}_{g^\epsilon}(LR(g^{\epsilon_{*k,e}}, g^{\epsilon}) > e^{-\lambda}) \geq p_0, \;\; \forall \epsilon, \; \forall n, \tag{7.9}$$

where $\mathbb{P}_{g^\epsilon}$ denotes the probability with respect to observations drawn from g^{ϵ} in the white noise model, and $LR(g^{\epsilon_{*k,e}}, g^{\epsilon})$ denotes the likelihood ratio between the observations associated to $g^{\epsilon_{*k,e}}$ and g^{ϵ}. Then any estimator $\hat{f}$ satisfies

$$\sup_{g^\epsilon \in \mathcal{G}} \mathbb{E}_{g^\epsilon} \|\hat{f} - g^{\epsilon}\|_{L^q} \geq \frac{e^{-\lambda}\, p_0}{2} \, \delta \, S_j^{1/q},$$

where δ is defined in (7.8).

Verification of (7.9)**:** The condition (7.9) is easily verified in our setting with Gaussian observations under the condition that $n\gamma^2 \leq c$ for n large enough (see Section 10.5 in Härdle et al. (2012)). Indeed, by Markov's inequality we have

$$\mathbb{P}_{g^\epsilon}\big(LR(g^{\epsilon_{*k,e}}, g^\epsilon) > e^{-\lambda}\big) \geq 1 - \frac{1}{\log e^\lambda} \mathbb{E}_{g^\epsilon}\Big|\log LR(g^{\epsilon_{*k,e}}, g^\epsilon)\Big|,$$

and using Proposition 6.1.7 in Giné and Nickl (2015) to bound the expectation by the Kullback-Leibler divergence we get

$$\mathbb{P}_{g^\epsilon}\big(LR(g^{\epsilon_{*k,e}}, g^\epsilon) > e^{-\lambda}\big) \geq 1 - \frac{1}{\lambda}\Big(K(dP_{g^{\epsilon_{*k,e}}}, dP_{g^\epsilon}) + \sqrt{2K(dP_{g^{\epsilon_{*k,e}}}, dP_{g^\epsilon})}\Big).$$

Using the Cameron-Martin Theorem to interpret the Gaussian probability measures (see Theorem 2.6.13 in Giné and Nickl (2015)), the Kullback-Leibler divergence between Gaussian measures is easily computed and gives

$$K(dP_{g^{\epsilon_{*k,e}}}, dP_{g^\epsilon}) = \frac{n}{2\sigma^2}\|g^{\epsilon_{*k,e}} - g^\epsilon\|_{L^2}^2 = \frac{n\gamma^2}{2\sigma^2}\|\psi_{j,k,e}\|_{L^2}^2 = \frac{n\gamma^2}{2\sigma^2}.$$

Hence, choosing $\gamma = t_0\, n^{-1/2}$ for a small enough constant $t_0 > 0$ gives (7.9).

Application of Lemma 2: The conclusion of the lemma applies, and we can lower bound the L^q-risk over the class $(B^s_{p,t} \cap L^\infty)_L$ by the risk over $\mathcal{G}$, i.e.,

$$\sup_{f \in (B^s_{p,t} \cap L^\infty)_L} \mathbb{E}_f \|\hat{f} - f\|_{L^q} \geq \sup_{g^\epsilon \in \mathcal{G}} \mathbb{E}_{g^\epsilon}\|\hat{f} - g^\epsilon\|_{L^q} \geq \frac{e^{-\lambda}\, p_0}{2}\, \delta\, 2^{j\Delta/q} \tag{7.10}$$

for any estimator $\hat{f}$. It remains to choose the scale parameter $j \geq 0$. Recall that we have chosen $\gamma = t_0\, n^{-1/2}$. Further, by (7.7) we also need $\gamma \leq c\, 2^{-j(s + d(\frac{1}{2} - \frac{1}{p}) + \frac{\Delta}{p})} = c\, 2^{-jd/2}$, for the choice $\Delta = d - sp$. We choose j such that

$$2^{-jd/2} = c\, n^{-1/2},$$

which using the definition (7.8) for δ gives the bound in (7.10)

$$\delta\, 2^{j\Delta/q} = c\, \gamma\, 2^{jd(\frac{1}{2} - \frac{1}{q})}\, 2^{j\Delta/q} = c\left(\frac{1}{n}\right)^{\frac{1}{2} - (\frac{1}{2} - \frac{1}{q}) - \frac{\Delta}{dq}} = c\, n^{-\frac{ps}{dq}}.$$

This completes the proof for $T = id$.

Modification for general T: For general T we construct the alternatives as above, using a wavelet basis for which (3.11) is satisfied. The first difference occurs for the application of the lower bound from Assouad's lemma: here we need to ensure condition (7.9) for the transformed alternatives $Tg^{\epsilon_{*k,e}}$ and Tg^ϵ, since our observations arise from those functions. Proceeding as above, we reduce the problem to bounding the Kullback-Leibler divergence between the associated Gaussian measures. The final condition that we need to ensure is that

$$K(dP_{Tg^{\epsilon_{*k,e}}}, dP_{Tg^{\epsilon}}) = \frac{n\gamma^2}{2\sigma^2}\|T\psi_{j,k,e}\|_{L^2}^2 \leq c_0$$

for some small constant $c_0 > 0$. Since our operator T satisfies (3.11), this condition holds if

$$\gamma^2 \asymp 2^{-jd} \asymp n^{-\frac{d}{d+2\beta}}.$$

Plugging this in gives the claim for $\beta \geq 0$. □

7.3 Proofs of auxiliary results

7.3.1 Proof of existence of a minimizer

In this section we prove Propositions 1 and 6, which guaranty the existence of the multiscale TV estimators in the regression and inverse problems settings. In fact, we prove a slightly stronger result. Let X denote a finite set, and let $K : L^\infty \to \ell^\infty(X)$ be a linear, bounded operator. Let further $Y \in \ell^\infty(X)$, $\gamma > 0$ and $L > 0$ be given. Consider the optimization problem

$$\operatorname*{argmin}_{g \in BV} |g|_{BV} \ \text{ such that } \ \|Kg - Y\|_{\ell^\infty} \leq \gamma, \ \ \|g\|_{L^\infty} \leq L, \ \ \operatorname{supp} g \subseteq [0,1]^d. \tag{7.11}$$

Proposition 10 below shows that (7.11) admits a minimizer in $BV \cap L^\infty$. In order to prove Proposition 1, choose $X = \Omega_n$ and let K denote the operator that maps a function $g \in L^\infty$ to its coefficients $\langle \phi_\omega, g\rangle$ for $\omega \in \Omega_n$, which is linear and bounded. Since the observations Y_ω, $\omega \in \Omega_n$ are almost surely finite, we conclude that there exists almost always a minimizer $\hat{f}_\Phi \in BV \cap L^\infty$ of (2.5), so Proposition 1 is proven.

Proposition 6 is proven analogously: the only difference being that we choose $(Kg)_{j,\theta} = \langle u_{j,\theta}, Tg\rangle$ for the vaguelette system $u_{j,\theta}$. Such K is a bounded operator from L^∞ to ℓ^∞, so by the same argument we conclude that the minimizer $\hat{f}_{\Phi,T}$ exists almost always.

It remains to prove the existence of minimizers of (7.11).

Proposition 10. Assume that $Y \in \ell^\infty(X)$, $\gamma > 0$ and $L > 0$. Then there exists a minimizer $\hat{f} \in BV \cap L^\infty$ of (7.11).

Proof. Note that we are minimizing a convex functional bounded from below subject to convex constraints. Let $\{g_k\}_{k\in\mathbb{N}}$ be a minimizing sequence of $|\cdot|_{BV}$ satisfying

$$\|g_k\|_{L^\infty} \leq L, \quad \|Kg_k - Y\|_{\ell^\infty(X)} \leq \gamma, \quad \operatorname{supp} g_k \subseteq [0,1]^d, \quad \text{for all } k \in \mathbb{N}.$$

Then $\{g_k\}_{k\in\mathbb{N}}$ is bounded in L^∞, which means that we can take a subsequence (still denoted by g_k) that converges weakly, i.e. $g_k \rightharpoonup \hat{f} \in L^\infty$. Moreover, this sequence is bounded in BV, so again we can conclude that, taking a subsequence, $\nabla g_k \rightharpoonup \nabla\hat{f}$, where here we mean weak convergence of Radon measures. Finally, we can take another subsequence such that the bounded sequence Kg_k converges weakly in $\ell^\infty(X)$, in which case it converges to $K\hat{f}$ (by linearity and boundedness of the mapping K).
The lower-semicontinuity of $|\cdot|_{BV}$, $||\cdot||_{L^\infty}$ and $||\cdot||_{\ell^\infty}$, together with the weak convergence stated above, implies that

$$\begin{aligned}|\hat{f}|_{BV} &\leq \liminf_{k\to\infty} |g_k|_{BV},\\ ||\hat{f}||_{L^\infty} &\leq \liminf_{k\to\infty} ||g_k||_{L^\infty} \leq L,\\ ||K\hat{f}-Y||_{\ell^\infty(X)} &\leq \liminf_{k\to\infty} ||Kg_k - Y||_{\ell^\infty(X)} \leq \gamma.\end{aligned}$$

Finally, it is clear that $\hat{f}$ satisfies supp $\hat{f} \subseteq [0,1]^d$, since it is the limit of functions supported there. This implies that $\hat{f} \in BV \cap L^\infty$ is a minimizer of (7.11). □

7.3.2 Tail bound for the noise

We prove an elementary tail bound for the dictionary coefficients of white noise.

Proposition 11 (Tail bounds on the coefficients of white noise)**.** Let $\Phi = \{\phi_\omega \,|\, \omega\in\Omega\} \subset L^2(\mathbb{M})$ be a family of functions defined on an open domain $\mathbb{M} \subseteq \mathbb{R}^d$ and satisfying $\sup_{\omega\in\Omega} ||\phi_\omega||_{L^2} \leq c$ for a constant $c > 0$. Then for any $n \in \mathbb{N}$ we have

$$\mathbb{P}\Big(\max_{\omega\in\Omega_n}\Big|\int_{\mathbb{M}} \phi_\omega(x)\,dW(x)\Big| \geq c\,t\Big) \leq \#\Omega_n\, e^{-t^2/2} \quad \text{for any } t \geq 0.$$

Proof. By the union bound we have

$$\mathbb{P}(\max_{\omega\in\Omega_n} |\epsilon_\omega| \geq t) \leq \sum_{\omega\in\Omega_n} \mathbb{P}(|\epsilon_\omega| \geq t)$$

for any $t \geq 0$. The random variables $\epsilon_\omega := c^{-1}\int_{\mathbb{M}} \phi_\omega(x)\,dW(x)$ are normal with variance smaller than 1, since $||\phi_\omega||_{L^2(\mathbb{M})} \leq c$. They are therefore stochastically dominated by standard normal random variables, so the probabilities in the right-hand side can be bounded as

$$\begin{aligned}\mathbb{P}(|\epsilon_\omega| \geq t) \leq 2\int_t^\infty e^{-x^2/2}\,\frac{dx}{\sqrt{2\pi}} &= 2\,e^{-t^2}\int_t^\infty e^{-x^2/2+t^2}\,\frac{dx}{\sqrt{2\pi}}\\ &\leq 2\,e^{-t^2}\int_t^\infty e^{-x^2/2+xt}\,\frac{dx}{\sqrt{2\pi}} = e^{-t^2/2}.\end{aligned}$$ □

7.3.3 Proofs for Section 2

Proof of Proposition 2

Proof of Proposition 2. We begin with the inequality in Assumption 1. Recall from Section 2.1 that the Besov norm of a function can be represented in terms of its wavelet coefficients with respect to a smooth enough wavelet basis. In particular, for a function g with supp $g \subseteq [0,1]^d$ we have

$$\|g\|_{B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)} \asymp \sup_{j\geq 0}\max_{k\in\mathbb{Z}^d}\max_{e\in E_j}|\langle\psi_{j,k,e},g\rangle| \leq \max_{0\leq j<J}\max_{k\in\mathbb{Z}^d}\max_{e\in E_j}|\langle\psi_{j,k,e},g\rangle| + \sup_{j\geq J}\max_{k\in\mathbb{Z}^d}\max_{e\in E_j}|\langle\psi_{j,k,e},g\rangle|.$$

Note that the first term is precisely $\max_{(j,k,e)\in\Omega_n}|\langle\psi_{j,k,e},g\rangle|$ for $J = \lfloor\frac{1}{d}\log_2 n\rfloor$ and Ω_n as in equation (2.3). Indeed, since g is supported in the unit cube, only the coefficients with (j,k,e) such that supp $\psi_{j,k,e}\cap(0,1)^d \neq \emptyset$ are nonzero.

It remains to show that the second term is dominated by $C\|g\|_{L^\infty(\mathbb{R}^d)}\, n^{-1/2}$. For that, Hölder's inequality yields

$$\sup_{j\geq J}\max_{k\in\mathbb{Z}^d}\max_{e\in E_j}|\langle\psi_{j,k,e},g\rangle| \leq \sup_{j\geq J}\max_{k\in\mathbb{Z}^d}\max_{e\in E_j}\|\psi_{j,k,e}\|_{L^1(\mathbb{R}^d)}\|g\|_{L^\infty(\mathbb{R}^d)} \leq C\,2^{-Jd/2}\|g\|_{L^\infty(\mathbb{R}^d)}, \quad (7.12)$$

where we used that the wavelets are of the form $\psi_{j,k,e}(x) = 2^{jd/2}\psi_e(2^j x - k)$, compactly supported and normed in L^2. Using now that $2^{-Jd/2} \leq 2^{d/2}\, n^{-1/2}$, the inequality follows. Moreover, since the index sets Ω_n satisfy $2^{-d}n \leq \#\Omega_n/(12\,S+1)^d \leq n$, we can choose $Q(x) = c\,x$ and $\Gamma = 1$ in Assumption 1. This completes the proof. □

Proof of Proposition 3

In order to prove Proposition 3 we rely on the characterization of Besov spaces in terms of local means.

Proposition 12 (Norm equivalence)**.** Let $\psi \in C^\infty(\mathbb{R}^d)$ satisfy supp $\psi \subseteq [0,1]^d$ and $|\mathcal{F}[\psi](\xi)| > 0$ for $|\xi| \leq 2$. Then the norm equivalence

$$\|g\|_{B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)} \asymp \sup_{j\geq 0} 2^{jd/2}\sup_{x\in\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\psi(2^j(x-y))\,g(y)\,dy\right|$$

holds for any function $g \in B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)$.

Proposition 12 is a consequence of Theorem 3 in Triebel (1988). We refer to Section A.3 of the Appendix for the proof.

Proof of Proposition 3. Note that by part b) of Remark 5, we have $n^{\max\{1,d/2\}} \leq \#\Omega_n \leq n^{\max\{1,d/2\}+1}$ for all $n \in \mathbb{N}$, so we have $\Gamma = \max\{1, d/2\}$ in Assumption 1.
For the inequality in Assumption 1, we have to show that there is a constant $C > 0$ such that for any $n \in \mathbb{N}$ we have

$$\|g\|_{B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)} \leq \frac{C}{\sqrt{n}}\|g\|_{L^\infty(\mathbb{R}^d)} + C \max_{(j,k)\in\Omega_n} \left| \int_{[0,1]^d} \psi_{j,k}(z) g(z)\, dz \right|$$

for any $g \in L^\infty(\mathbb{R}^d)$ with $\operatorname{supp} g \subseteq [0,1]^d$. By Proposition 12 we have the bound

$$\begin{aligned}\|g\|_{B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)} \leq{}& C \sup_{j\geq J} 2^{jd/2} \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \psi(2^j(x-y))\, g(y)\, dy \right| \\ &+ C \sup_{j<J} 2^{jd/2} \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \psi(2^j(x-y))\, g(y)\, dy \right| \qquad (7.13)\end{aligned}$$

for any $J > 0$, to be fixed later. The first term in the right-hand side can be bounded as

$$\begin{aligned}\sup_{j\geq J} 2^{jd/2} \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \psi(2^j(x-y))\, g(y)\, dy \right| &= \sup_{j\geq J} 2^{-jd/2} \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \psi(z)\, g(x-2^{-j}z)\, dz \right| \\ &\leq \sup_{j\geq J} 2^{-jd/2} \sup_{x\in\mathbb{R}^d} \|\psi\|_{L^1(\mathbb{R}^d)} \|g\|_{L^\infty(\mathbb{R}^d)} \\ &\leq 2^{-Jd/2}\|\psi\|_{L^1}\|g\|_{L^\infty},\end{aligned}$$

and the right-hand side is bounded, since ψ is a compactly supported smooth function, so it is integrable in particular. It remains to bound the second term in the right-hand side of (7.13). For $j \in \mathbb{N}$ and $x \in \mathbb{R}^d$, define the integral

$$\mathcal{I}_{j,x} = \int_{\mathbb{R}^d} \psi(2^j(x-y))\, g(y)\, dy = \int_{[0,1]^d} \psi(2^j(x-y))\, g(y)\, dy,$$

which can be restricted to integration over $[0,1]^d$, since we assume that g is supported in the unit cube. Moreover notice that, since $\operatorname{supp}\psi \subseteq [0,1]^d$, if x is such that $|x-y|_\infty > 2^{-j}$ for all $y \in [0,1]^d$, then $\mathcal{I}_{j,x} = 0$. Here, $|z|_\infty := \max_{i=1,\dots,d} |z_i|$ denotes the $\ell\infty$ norm of $z \in \mathbb{R}^d$. With this observations, we can write the second term in the right-hand side of (7.13) as

$$\sup_{j<J} 2^{jd/2} \sup_{x\in[-2^{-j},1+2^{-j}]^d} \left| \int_{[0,1]^d} \psi(2^j(x-y))\, g(y)\, dy \right|.$$

We take the supremum over $x \in [-2^{-j}, 1+2^{-j}]^d$ because the integral vanishes for x outside of that cube. We now approximate this expression by proving the following:

$$\begin{aligned} &\forall j = 0,\dots,J-1,\ \ \forall x \in [-2^{-j}, 1+2^{-j}]^d,\ \ \exists (j,k) \in \Omega_n \\ &\qquad \text{such that } \ |\mathcal{I}_{j,x} - \mathcal{I}_{j,k}| \le C\,\|g\|_{L^\infty}\, 2^{-R-jd+j}, \end{aligned} \tag{7.14}$$

for a certain $R > 0$, i.e., for each j, the integrals $\mathcal{I}_{j,x}$ can be approximated uniformly in x by the integrals $\{\mathcal{I}_{j,k} \,|\, (j,k) \in \Omega_n\}$. Before we prove (7.14), let us see what it implies. With it, we can bound the second term in the right-hand side of (7.13) as

$$\begin{aligned} \max_{j<J} 2^{jd/2} \sup_{x\in[-2^{-j},1+2^{-j}]^d} |\mathcal{I}_{j,x}| &\le \max_{j<J} 2^{jd/2} \sup_{x\in[-2^{-j},1+2^{-j}]^d} \min_{k\ s.t.\ (j,k)\in\Omega_n} |\mathcal{I}_{j,x} - \mathcal{I}_{j,k}| + |\mathcal{I}_{j,k}| \\ &\le \max_{j<J} 2^{jd/2} \max_{k\ s.t.\ (j,k)\in\Omega_n} C\,\|g\|_{L^\infty}\, 2^{-R-jd+j} + |\mathcal{I}_{j,k}| \\ &\le C\,\|g\|_{L^\infty}\, 2^{-R} \max_{j<J} 2^{j-jd/2} + \max_{j\le J} 2^{jd/2} \max_{k\ s.t.\ (j,k)\in\Omega_n} |\mathcal{I}_{j,k}|. \end{aligned}$$

This equation with $R = J\max\{1, d/2\}$ and $J = \lceil \frac{1}{d}\log_2 n \rceil$ yields the claim. It just remains to prove (7.14).

Proof of (7.14)**:**
Recall from Assumption 2 that for each $n \in \mathbb{N}$ we have

$$\begin{aligned} &\Omega_n = \{(j,k) \,|\, j = 0,\dots,J-1,\ k \in \mathcal{D}_j\}, \\ &\mathcal{D}_j = \{k = (k_1,\cdots,k_d) \,|\, k_i = -2^{-j} + l_i\, 2^{-R}(1+2^{1-j}),\ \ l_i = 0,\dots,2^R-1,\ i = 1,\dots,d\}, \end{aligned}$$

where $J = \lceil \frac{1}{d}\log_2 n \rceil$ and $R = J\max\{1, d/2\}$. Consequently, for each $x \in [-2^{-j}, 1+2^{-j}]^d$ we can find $k \in \mathcal{D}_R$ such that $|x-k| \le \sqrt{d}\,2^{-R}\,(1+2^{1-j})$. With this in mind, we can bound

$$\begin{aligned} \mathcal{I}_{j,x} - \mathcal{I}_{j,k} &= \int_{[0,1]^d} g(z)\Big(\psi(2^j(x-z)) - \psi(2^j(k-z))\Big)\,dz \\ &\le \|g\|_{L^\infty} \int_{[0,1]^d} \Big|\psi(2^j(x-z)) - \psi(2^j(k-z))\Big|\,dz \\ &= \|g\|_{L^\infty}\, 2^{-jd} \int_{2^jx-[0,2^j]^d} \Big|\psi(y) - \psi(y + 2^j(k-x))\Big|\,dy \\ &\le \|g\|_{L^\infty}\, 2^{-jd} \int_{2^jx-[0,2^j]^d} |\nabla\psi(y)|\big|2^j(k-x)\big|\,dy \\ &\le \|g\|_{L^\infty}\, 2^{-jd}\, \|\nabla\psi\|_{L^1}\, \big|2^j(k-x)\big| \\ &\le \sqrt{d}\,(1+2^{1-j})\, \|g\|_{L^\infty}\, \|\nabla\psi\|_{L^1}\, 2^{-jd+j-R}. \end{aligned}$$

This proves (7.14) and finishes the proof. □

Proof of Proposition 4

Proof of Proposition 4. The inequality in Assumption 1 follows in both cases (curvelet and shearlet) from the inequality (7.12) for the wavelet basis (see the proof of Proposition 2 above). Indeed, denoting the elements of Φ by

$$\phi_\omega = \begin{cases} \psi_{j,k,e} & \text{if } \omega = (j,k,e) \in \Theta^W \quad \text{(wavelets)}, \\ \varphi_{j,\tilde{\theta}} & \text{if } \omega = (j,\tilde{\theta}) \in \Theta \quad \text{(curvelets or shearlets)}, \end{cases}$$

we have

$$\begin{aligned} \|g\|_{B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)} &\le C \max_{(j,k,e)\in\Theta_n^W} |\langle g, \psi_{j,k,e}\rangle| + C\frac{\|g\|_{L^\infty(\mathbb{R}^d)}}{\sqrt{n}} \\ &\le C \max_{\omega\in\Theta_n^W\cup\Theta_n} |\langle g, \phi_\omega\rangle| + C\frac{\|g\|_{L^\infty(\mathbb{R}^d)}}{\sqrt{n}}, \end{aligned}$$

for any function g supported on the unit cube, where we just enlarge the right-hand side by taking the maximum over a larger index set. Concerning the cardinality of $\Omega_n \cup \Theta_n$, by Assumption 3 we have

$$\#(\Omega_n \cup \Theta_n) = 2^{d\lfloor \frac{1}{d}\log_2 n\rfloor} + 2^{d\lfloor \frac{1}{d}\log_2 n\rfloor},$$

and hence we have Assumption 1 with $Q(x) = 2x$ and $\Gamma = 1$. □

7.3.4 Error bound in the discretized model

In this section we prove Proposition 5 from Section 2.5.

Proof of Proposition 5 . Plug in the definition of $\phi^n_\omega = n^{-1/2}\,\phi_\omega(x_i)$ into δ_n. Using that one of the functions ϕ_ω is the indicator function of the unit cube, this gives

$$\delta_n \ge \left| \sum_{x_i\in\Gamma_n} \int_{x_i+[0,n^{-1/d})^d} (h(x_i) - h(y))\, dy \right|. \tag{7.15}$$

Fix an irrational number $\alpha \in (0,1)$, and consider the function

$$h(x) = \begin{cases} 0 & \text{if } x^{(1)} \le \alpha \\ 1 & \text{else.} \end{cases}$$

Here, $x^{(1)}$ denotes the first coordinate of the vector $x = (x^{(1)}, \dots, x^{(d)})$. Due to the definition of

h, the summands in the lower bound for δ_n satisfy the following:

$$\int_{x_i+[0,n^{-1/d}]^d} (h(x_i) - h(y))\, dy = 0 \quad \text{if } \alpha \leq x_i^{(1)} \quad \text{or} \quad x_i^{(1)} + n^{-1/d} \leq \alpha.$$

For $x_i^{(1)} < \alpha < x_i^{(1)} + n^{-1/d}$ we get

$$\begin{aligned}\int_{x_i+[0,n^{-1/d}]^d} (h(x_i) - h(y))\, dy &= \int_{x_i^{(1)}}^{x_i^{(1)}+n^{-1/d}} \int_{x_i^{(2)}}^{x_i^{(2)}+n^{-1/d}} \cdots \int_{x_i^{(d)}}^{x_i^{(d)}+n^{-1/d}} (-h(y))\, dy^{(1)} \cdots dy^{(d)} \\ &= \int_{\alpha}^{x_i^{(1)}+n^{-1/d}} \int_{x_i^{(2)}}^{x_i^{(2)}+n^{-1/d}} \cdots \int_{x_i^{(d)}}^{x_i^{(d)}+n^{-1/d}} (-1)\, dy^{(1)} \cdots dy^{(d)} \\ &= (-1)(x_i^{(1)} + n^{-1/d} - \alpha)(n^{-1/d})^{d-1}.\end{aligned}$$

Furthermore, note that there are $n^{1-1/d}$ nonzero summands in the lower bound (7.15) for δ_n. Indeed, there are $n^{1-1/d}$ elements $x_i \in \Gamma_n$ with $x_i^{(1)} \in [\alpha - n^{-1/d}, \alpha)$. Consequently, the discretization error δ_n is lower bounded by

$$\delta_n \geq x_i^{(1)} + n^{-1/d} - \alpha. \tag{7.16}$$

Now, since $x_i \in \Gamma_n$, we have $x_i^{(1)} = k\, n^{-1/d} = k/m$ for some $k \in \mathbb{N}$ and $m = n^{1/d}$. In order to show that the lower bound is larger that $m^{-1}/2$ for infinitely many $m \in \mathbb{N}$, we use a classical result in irrational approximation: the sequence $z_m = m\,\alpha$ is uniformly distributed modulo 1 for α irrational (see e.g. Theorem 3.3 in Chapter 1 of Kuipers and Niederreiter (1974)). Now, the lower bound (7.16) is exactly of the form $m^{-1}(k + 1 - m\,\alpha)$, and by the above result and the definition of k, the sequence $k + 1 - m\,\alpha$ is uniformly distributed in $[0, 1]$ as m varies. Consequently, there are infinitely many $m \in \mathbb{N}$ such that $k + 1 - m\,\alpha > 1/2$, so we conclude that

$$\delta_n \geq \frac{1}{2} m^{-1}$$

for infinitely many $m \in \mathbb{N}$, which is what we wanted to prove. □

7.3.5 Proofs for Section 3

Proposition 13. In the setting of Section 3.3 we have

$$\begin{aligned} &T^* u_{j,k,e} = \kappa_j \psi_{j,k,e} \quad \text{where } \kappa_j = 2^{-j\beta}, \\ &c_1 \leq \|u_{j,k,e}\|_{L^2} \leq c_2 \quad \forall (j,k,e) \in \Omega, \end{aligned}$$

where we can choose $c_1 = \min_{e\in\{0,1\}^d} \|(-\Delta)^{\beta/2}\psi_{0,0,e}\|_{L^2}$ and $c_2 = \max_{e\in\{0,1\}^d} \|\psi_{0,0,e}\|_{H^\beta}$.

Proof. Notice that the Fourier transform of the elements $u_{j,k,e}$ is given by

$$\mathcal{F}[u_{j,k,e}](\xi) = 2^{-jd/2-j\beta} e^{-i\xi\cdot k2^{-j}} \frac{\mathcal{F}[\psi_{0,0,e}](2^{-j}\xi)}{\mathcal{F}[K](-\xi)}. \tag{7.17}$$

The first claim of the proposition follows trivially by construction of the $u_{j,k,e}$: we essentially use that T^* acts by convolution with $K(-\cdot)$, which in Fourier domain is the product with $\mathcal{F}[K](-\cdot)$. For the bounds in the L^2 norm, we use Plancherel's theorem, i.e.

$$\begin{aligned}
\|u_{j,k,e}\|_{L^2}^2 = \|\mathcal{F}[u_{j,k,e}]\|_{L^2}^2 &= 2^{-jd-2j\beta} \int_{\mathbb{R}^d} \left| \frac{\mathcal{F}[\psi_{0,0,e}](2^{-j}\xi)}{\mathcal{F}[K](-\xi)} \right|^2 \frac{d\xi}{(2\pi)^d} \\
&\asymp 2^{-jd-2j\beta} \int_{\mathbb{R}^d} (1+|\xi|^2)^\beta \left|\mathcal{F}[\psi_{0,0,e}](2^{-j}\xi)\right|^2 d\xi \\
&= 2^{-2j\beta} \int_{\mathbb{R}^d} (1+|2^j\xi|^2)^\beta \left|\mathcal{F}[\psi_{0,0,e}](\xi)\right|^2 d\xi, \qquad (7.18)
\end{aligned}$$

where in the second line we used the bounds (3.14) on the Fourier transform of the kernel K. The expression in the right-hand side can now be easily bounded from below as

$$\begin{aligned}
2^{-2j\beta} \int_{\mathbb{R}^d} (1+|2^j\xi|^2)^\beta \left|\mathcal{F}[\psi_{0,0,e}](\xi)\right|^2 d\xi &\geq 2^{-2j\beta} \int_{\mathbb{R}^d} |2^j\xi|^{2\beta} \left|\mathcal{F}[\psi_{0,0,e}](\xi)\right|^2 d\xi \\
&= \left\| |\xi|^\beta \mathcal{F}[\psi_{0,0,e}] \right\|_{L^2}^2 = \|(-\Delta)^{\beta/2}\psi_{0,0,e}\|_{L^2}^2,
\end{aligned}$$

again by Plancherel's theorem. On the other hand, the right-hand side of (7.18) can be upper-bounded as

$$\begin{aligned}
2^{-2j\beta} \int_{\mathbb{R}^d} (1+|2^j\xi|^2)^\beta \left|\mathcal{F}[\psi_{0,0,e}](\xi)\right|^2 d\xi &\leq 2^{-2j\beta} \int_{\mathbb{R}^d} (2^j+|2^j\xi|^2)^\beta \left|\mathcal{F}[\psi_{0,0,e}](\xi)\right|^2 d\xi \\
&= \left\| (1+|\xi|^2)^{\beta/2} \mathcal{F}[\psi_{0,0,e}] \right\|_{L^2}^2 = \|\psi_{0,0,e}\|_{H^\beta}^2.
\end{aligned}$$

This yields the claim. □

APPENDIX A

Harmonic Analysis

A.1 S-regularity of wavelet bases

We give here the definition of S-regularity for a wavelet basis of $L^2(\mathbb{R})$ as stated in Definition 4.2.14 of Giné and Nickl (2015). This property is extended to wavelet bases of $L^2(\mathbb{R}^d)$ by tensorization of one-dimensional bases.

Definition 3 (Definition 4.2.14 in Giné and Nickl (2015))**.** Consider a multiresolution basis of $L^2(\mathbb{R})$ of the form

$$\Phi = \{\varphi_k = \varphi(\cdot - k),\ \ \psi_{j,k} = 2^{j/2}\,\psi(2^j\cdot - k)\,\big|\, k \in \mathbb{Z},\ j \in \mathbb{N}_0\},$$

and define $K(x, y) = \sum_k \varphi(x - k)\varphi(y - k)$. For $S \in \mathbb{N}$, the basis Φ is said to be S-regular if the following conditions hold:

(1) $\int_{\mathbb{R}} \varphi(x)\,dx = 1, \quad \int_{\mathbb{R}} x^l\,\psi(x)\,dx = 0$ for $l = 0, \dots, S - 1$, and

$$\int_{\mathbb{R}} K(y, y + x)\,dx = 1, \qquad \int_{\mathbb{R}} x^l\,K(y, y + x)\,dx = 0 \quad \text{for } l = 1, \dots, S - 1, \forall y \in \mathbb{R};$$

(2) $\sum_{k\in\mathbb{Z}} |\varphi(\cdot - k)| \in L^\infty(\mathbb{R})$ and $\sum_{k\in\mathbb{Z}} |\psi(\cdot - k)| \in L^\infty(\mathbb{R})$;

(3) for a kernel $\kappa(x, y)$ equal to either $K(x, y)$ or $\sum_{k\in\mathbb{Z}} \psi(x - k)\psi(y - k)$, there are constants $c_1, c_2 > 0$ such that

$$\sup_{y\in\mathbb{R}} |\kappa(y, y - x)| \leq c_1\,G(c_2|x|) \ \ \forall x \in \mathbb{R},$$

where G is a real-valued, bounded and integrable function satisfying $\int_{\mathbb{R}} |x|^S\,G(|x|)\,dx < \infty$. ♣

A.2 Schwartz space and temperate distributions

We denote by $\mathcal{S}(\mathbb{R}^d)$ the set of all functions $\varphi \in C^\infty(\mathbb{R}^d)$ such that

$$\sup_{x\in\mathbb{R}^d} \left|x^\alpha \partial^\beta \varphi(x)\right| \le C_{\alpha,\beta} < \infty$$

for all multi-indices $\alpha, \beta \in \mathbb{N}_0^d$. For each α and β, the left-hand side defines a seminorm, and the topology induced by all these seminorms turns $\mathcal{S}(\mathbb{R}^d)$ into a Fréchet space (see Chapter VII in Hörmander (1990)). We refer to $\mathcal{S}(\mathbb{R}^d)$ as *Schwartz space*, and to its elements as *Schwartz functions*.

Continuous linear functionals on Schwartz functions are called *temperate distributions*. The set of all such functionals is denoted by $\mathcal{S}^*(\mathbb{R}^d)$.

A.3 Characterization of Besov spaces by local means

In this section we give the proof of Proposition 12 in Section 7.3.3, which gives a characterization of a Besov norm in terms of local means. The proof relies on Theorem 3 of (Triebel, 1988), which we recall here for completeness. Let $h, H \in \mathcal{S}(\mathbb{R}^d)$ be Schwartz functions satisfying

$$\begin{aligned} h(x) = 1 \quad \text{in} \quad |x| \le 1, \quad &\text{and} \quad \operatorname{supp} h \subseteq \{|x| \le 2\}, \\ H(x) = 1 \quad \text{in} \quad 1/2 \le |x| \le 2, \quad &\text{and} \quad \operatorname{supp} H \subseteq \{1/4 \le |x| \le 4\}. \end{aligned}$$

Let further $\varphi \in C^\infty(\mathbb{R}^d)$ satisfy $|\varphi(\xi)| > 0$ in $|\xi| < 2$, as well as the bounds

$$\begin{aligned} &\int_{\mathbb{R}^d} |\mathcal{F}^{-1}[\varphi(\xi)h(\xi)](y)|\, dy < \infty, \\ \sup_{j\ge1} 2^{-js_0} &\int_{\mathbb{R}^d} |\mathcal{F}^{-1}[\varphi(2^j\xi)H(\xi)](y)|\, dy < \infty, \end{aligned} \tag{A.1}$$

for some number $s_0 < -d/2$.

Theorem 7 (Particular case of Theorem 3 in Triebel (1988))**.** For φ as above, define $\varphi_j(\cdot) := \varphi(2^{-j}\cdot)$ for $j \in \mathbb{N}$. Then the expression

$$\sup_{j\ge0} 2^{-jd/2} \sup_{x\in\mathbb{R}^d} |\mathcal{F}^{-1}[\varphi_j(\xi)\mathcal{F}[g](\xi)](x)|$$

for $g \in B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)$ is equivalent to the norm of $B^{-d/2}_{\infty,\infty}(\mathbb{R}^d)$.

The proof of Theorem 7 consists essentially in relating the functions ψ_j to the functions used in the Paley-Littlewood decomposition. Notice that the inverse Fourier transform in the theorem can be rewritten as

$$\begin{aligned}
\mathcal{F}^{-1}[\varphi(2^{-j}\cdot)\mathcal{F}[g]](x) &= \int_{\mathbb{R}^d} e^{i\xi x}\varphi(2^{-j}\xi)\mathcal{F}[g](\xi)\,d\xi/(2\pi)^d \\
&= \int_{\mathbb{R}^d} e^{i\xi x}\int_{\mathbb{R}^d} e^{-i2^{-j}\xi z}\mathcal{F}[\varphi](z)\,dz \int_{\mathbb{R}^d} e^{-i\xi y} g(y)\,dy\,d\xi/(2\pi)^d \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathcal{F}[\varphi](z)\,g(y)\int_{\mathbb{R}^d} \exp\{i\xi(x-y-2^{-j}z)\}\,d\xi/(2\pi)^d\,dy\,dz \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathcal{F}[\varphi](z)\,g(y)\,\delta(x-y-2^{-j}z)\,dy\,dz \\
&= \int_{\mathbb{R}^d} \mathcal{F}[\varphi](z)\,g(x-2^{-j}z)\,dz.
\end{aligned}$$

Notice that inserting this in the expression of the theorem, we recover the claim of Proposition 12 with $\psi(z) = \mathcal{F}[\varphi](z)$. Hence, in order to prove Proposition 12, we need to show that for a function ψ satisfying Assumption 2, its inverse Fourier transform $\mathcal{F}^{-1}[\psi]$ satisfies the assumptions of Theorem 7. Recall that Assumption 2 implies that

$$\begin{gathered}
\psi \in C^\infty(\mathbb{R}^d), \quad \operatorname{supp}\psi \subseteq [0,1]^d, \quad |\mathcal{F}[\psi](\xi)| > 0 \ \text{ in } |\xi| < 2, \\
\text{and } \|\psi\|_{L^2(\mathbb{R}^d)} = 1, \ \|\psi\|_{L^\infty(\mathbb{R}^d)} \le 2.
\end{gathered}$$

Since the Fourier transform of ψ does not vanish near zero, neither does its inverse Fourier transform. This means that we just have to verify the bounds (A.1) for $\varphi = \mathcal{F}^{-1}[\psi]$. In Proposition 14 we show that the second bound in (A.1) holds. The same strategy can be used to show that the first bound in (A.1) holds.

Proposition 14. Let $s_0 \in \mathbb{R}$ satisfy $s_0 < -d/2$. Under the assumptions above, the inequality

$$\sup_{j\in\mathbb{N}} 2^{-js_0}\int_{\mathbb{R}^d} |\mathcal{F}^{-1}[\mathcal{F}^{-1}[\psi](2^j\cdot)H(\cdot)](y)|\,dy < \infty \tag{A.2}$$

holds.

Proof. For simplicity of the notation, we denote the inverse Fourier transform of ψ by $\Psi := \mathcal{F}^{-1}[\psi]$. We prove (A.2) by taking advantage of the support properties of ψ and H. Since ψ is a smooth function of compact support, its inverse Fourier transform Ψ is a Schwartz function, and in particular it satisfies

$$\sup_{x\in\mathbb{R}^d} |x|^\alpha|\partial_x^\beta\Psi(x)| \le C_{\alpha,\beta} < \infty$$

for any $\alpha \in \mathbb{N}_0, \beta \in \mathbb{N}_0^d$. Consequently we have

$$|\partial_x^\beta \Psi(2^j x)| = 2^{j|\beta|} |(\partial_x^\beta \Psi)(2^j x)| \leq C_{|\beta|+M,\beta}\, 2^{-jM}\, |x|^{-M-|\beta|}$$

uniformly in $|x| \geq 1$ for any multi-index $\beta \in \mathbb{N}_0^d$ and any $M \in \mathbb{N}_0$. Hence we have

$$\int_{\mathbb{R}^d} |\partial_y^\beta \Psi(2^j y) H(y)|^2\, dy \leq 2^{-2jM}\, C^2_{|\beta|+M,\beta} \int_{\mathbb{R}^d} |y|^{-2M-2|\beta|}\, |\max_{\gamma\leq\beta}\{\partial_y^\gamma H(y)\}|^2\, dy \leq C\, C^2_{|\beta|+M,\beta}\, 2^{-2jM}$$

for any $M \in \mathbb{N}_0$, where the integral is finite due to the support properties of H. Taking $|\beta| = 2\sigma \in \mathbb{N}_0$, this equation implies that $\Psi(2^j\cdot)H(\cdot)$ belongs to the Sobolev space $H^{2\sigma}(\mathbb{R}^d)$ with norm $\|\Psi(2^j\cdot)H(\cdot)\|_{H^{2\sigma}(\mathbb{R}^d)} \leq C_{M,\sigma}\, 2^{-jM}$. In order to prove (A.2) we show that

$$\int_{\mathbb{R}^d} |\mathcal{F}^{-1}[\Psi(2^j\cdot)H](y)|\, dy \leq C\, \|\Psi(2^j\cdot)H(\cdot)\|_{H^{2\sigma}(\mathbb{R}^d)} \tag{A.3}$$

holds for $\sigma > d/2$, and (A.2) follows then taking $M > -s_0$ (recall that $s_0 < s < 0$). It remains to show (A.3), which follows from the bounds

$$\begin{aligned}\int_{\mathbb{R}^d} |\mathcal{F}^{-1}[\Psi(2^j\cdot)H](y)|\, dy &= \int_{\mathbb{R}^d} (1+|y|^2)^{-\sigma}(1+|y|^2)^{\sigma} |\mathcal{F}^{-1}[\Psi(2^j\cdot)H](y)|\, dy \\ &\leq C \int_{\mathbb{R}^d} (1+|y|^2)^{-\sigma} \Big| \sum_{|\beta|\leq 2\sigma} \int_{\mathbb{R}^d} e^{-2\pi i\xi y} \partial_\xi^\beta \Psi(2^j\xi) H(\xi)\, d\xi \Big|\, dy \\ &\leq C \int_{\mathbb{R}^d} (1+|y|^2)^{-\sigma}\, dy \sum_{|\beta|\leq 2\sigma} \int_{\mathbb{R}^d} |\partial_\xi^\beta \Psi(2^j\xi) H(\xi)|\, d\xi.\end{aligned}$$

Now since $\sigma > d/2$, the first integral in the right-hand side is finite. Furthermore the compact support of H implies that the integrand in the second integral also has compact support. Since the L^1 norm of a compactly supported function can be upper bounded by its L^2 norm times a constant, and we conclude that

$$\int_{\mathbb{R}^d} |\mathcal{F}^{-1}[\Psi(2^j\cdot)H](y)|\, dy \leq C \sum_{|\beta|\leq 2\sigma} \|\partial_y^\beta \Psi(2^j\cdot)H\|_{L^2(\mathbb{R}^d)} \leq C\|\Psi(2^j\cdot)H\|_{H^{2\sigma}(\mathbb{R}^d)},$$

which is what we wanted to prove. □

Bibliography

Abramovich, F. U. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1):115–129.

Assouad, P. (1983). Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024.

Bauer, F., Hohage, T., and Munk, A. (2009). Iteratively regularized Gauss–Newton method for nonlinear inverse problems with random noise. *SIAM Journal on Numerical Analysis*, 47(3):1827–1846.

Bauschke, H. H., Combettes, P. L., and Luke, D. R. (2006). A strongly convergent reflection method for finding the projection onto the intersection of two closed convex sets in a Hilbert space. *Journal of Approximation Theory*, 141(1):63 – 69.

Bertero, M., Boccacci, P., Desiderà, G., and Vicidomini, G. (2009). Image deblurring with Poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006.

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636.

Blanchard, G. and Mathé, P. (2012). Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse Problems*, 28(11):115011.

Borup, L. and Nielsen, M. (2007). Frame decomposition of decomposition spaces. *Journal of Fourier Analysis and Applications*, 13(1):39–70.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398.

Cai, T. T. and Low, M. G. (2005). Nonparametric estimation over shrinking neighborhoods: superefficiency and adaptation. *The Annals of Statistics*, 33(1):184–213.

Candès, E. J. and Donoho, D. L. (2000). Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Technical report, Stanford University, California, Department of Statistics.

Candès, E. J. and Donoho, D. L. (2002). Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *The Annals of Statistics*, 30:784–842.

Candès, E. J. and Guo, F. (2002). New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction. *Signal Processing*, 82(11):1519–1543.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Cavalier, L. (2011). Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, pages 3–96. Springer.

Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, 123(3):323–354.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1):89–97.

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

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145.

Christensen, O. (2003). *An introduction to frames and Riesz bases*, volume 7. Springer.

Clason, C., Jin, B., and Kunisch, K. (2010). A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration. *SIAM Journal on Imaging Sciences*, 3(2):199–231.

Clason, C., Kruse, F., and Kunisch, K. (2018). Total variation regularization of multi-material topology optimization. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(1):275–303.

Cohen, A. (2003). *Numerical analysis of wavelet methods*, volume 32. Elsevier.

Cohen, A., Dahmen, W., Daubechies, I., and DeVore, R. (2003). Harmonic analysis of the space BV. *Revista Matematica Iberoamericana*, 19(1):235–263.

Cohen, A., Hoffmann, M., and Reiss, M. (2004). Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM Journal on Numerical Analysis*, 42(4):1479–1501.

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

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

Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61. Society for Industrial and Applied Mathematics, Philadelphia.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29:1–48.

del Álamo, M., Li, H., and Munk, A. (2018). Frame-constrained total variation regularization for white noise regression. *arXiv preprint arXiv:1807.02038*.

del Álamo, M. and Munk, A. (2019). Total variation multiscale estimators for linear inverse problems. *arXiv preprint arXiv:1905.08515*.

Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Applied and Computational Harmonic Analysis*, 3(3):215–228.

Dong, Y., Hintermüller, M., and Rincon-Camacho, M. M. (2011). Automated regularization parameter selection in multi-scale total variation models for image restoration. *Journal of Mathematical Imaging and Vision*, 40(1):82–104.

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):100–115.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Applied and Computational Harmonic Analysis*, 2(2):101–126.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1997). Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 183–218. Springer, New York.

Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli*, 9(3):423–449.

Dümbgen, L. and Kovac, A. (2009). Extensions of smoothing via taut strings. *Electronic Journal of Statistics*, 3:41–75.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, 29(1):124–152.

Durand, S. and Froment, J. (2001). Artifact free signal denoising with wavelets. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3685–3688.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842.

Evans, L. C. and Gariepy, R. F. (2015). *Measure theory and fine properties of functions*. CRC Press.

Frick, K., Marnitz, P., and Munk, A. (2012). Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electronic Journal of Statistics*, 6:231–268.

Frick, K., Marnitz, P., and Munk, A. (2013). Statistical multiresolution estimation for variational imaging: With an application in Poisson-biophotonics. *Journal of Mathematical Imaging and Vision*, 46(3):370–387.

Garnett, J. B., Le, T. M., Meyer, Y., and Vese, L. A. (2007). Image decompositions using bounded variation and generalized homogeneous Besov spaces. *Applied and Computational Harmonic Analysis*, 23(1):25–56.

Giné, E. and Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press.

Gobet, E., Hoffmann, M., and Reiß, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, 32(5):2223–2253.

Goldenshluger, A. and Lepskii, O. (2014). On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, 159(3-4):479–543.

Grasmair, M., Li, H., and Munk, A. (2018). Variational multiscale nonparametric regression: smooth functions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54(2):1058–1097.

Grohs, P., Keiper, S., Kutyniok, G., and Schaefer, M. (2013). Alpha molecules: curvelets, shearlets, ridgelets, and beyond. In *Wavelets and Sparsity XV*. International Society for Optics and Photonics.

Guntuboyina, A., Lieu, D., Chatterjee, S., and Sen, B. (2017). Spatial adaptation in trend filtering. *To appear in The Annals of Statistics*.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594.

Guo, K., Kutyniok, G., and Labate, D. (2006). Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines (Athens, GA, 2005)*. Nashboro Press, Nashville, TN.

Haddad, A. and Meyer, Y. (2007). An improvement of Rudin–Osher–Fatemi model. *Applied and Computational Harmonic Analysis*, 22(3):319–334.

Haltmeier, M. (2013). Inversion of circular means and the wave equation on convex planar domains. *Computers & Mathematics with Applications*, 65(7):1025–1036.

Haltmeier, M. and Munk, A. (2014). Extreme value analysis of empirical frame coefficients and implications for denoising by soft-thresholding. *Applied and Computational Harmonic Analysis*, 36(3):434–460.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2017). Isotonic regression in general dimensions. *To appear in The Annals of Statistics*.

Han, Q. and Wellner, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *To appear in The Annals of Statistics*.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012). *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media.

Hintermüller, M. (2010). Semismooth Newton methods and applications. Technical report, Department of Mathematics, Humboldt-University of Berlin.

Hintermüller, M. and Rasch, J. (2015). Several path-following methods for a class of gradient constrained variational inequalities. *Computers & Mathematics with Applications*, 69(10):1045 – 1067.

Hoffmann, M. and Reiss, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, 36(1):310–336.

Hohage, T. (2000). Regularization of exponentially ill-posed problems. *Numerical Functional Analysis and Optimization*, 21(3-4):439–464.

Hohage, T. and Miller, P. (2019). Optimal convergence rates for sparsity promoting wavelet-regularization in Besov spaces. *Inverse Problems*.

Hörmander, L. (1990). *The analysis of linear partial differential operators*. Grundlehren der mathematischen Wissenschaften. Springer.

Hütter, J.-C. and Rigollet, P. (2016). Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146.

Kabluchko, Z. (2011). Extremes of the standardized Gaussian noise. *Stochastic Processes and their Applications*, 121(3):515 – 533.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657.

Kuipers, L. and Niederreiter, H. (1974). *Uniform distribution of sequences*. Wiley-Interscience, John Wiley & Sons, New York.

Kutyniok, G., Lemvig, J., and Lim, W.-Q. (2012). Compactly supported shearlets. In *Approximation Theory XIII: San Antonio 2010*, pages 163–186. Springer, New York.

Labate, D., Lim, W.-Q., Kutyniok, G., and Weiss, G. (2005). Sparse multidimensional representation using shearlets. In *Wavelets XI*, volume 5914. International Society for Optics and Photonics.

Labate, D., Mantovani, L., and Negi, P. (2013). Shearlet smoothness spaces. *Journal of Fourier Analysis and Applications*, 19(3):577–611.

Lassas, M., Saksman, E., and Siltanen, S. (2009). Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122.

Lassas, M. and Siltanen, S. (2004). Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537–1563.

Ledoux, M. (2003). On improved Sobolev embedding theorems. *Mathematical Research Letters*, 10(5/6):659–670.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lepskii, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466.

Lepskii, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *The Annals of Statistics*, 43(3):1178–1242.

Lepskii, O. V., Mammen, E., and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25(3):929–947.

Li, H. (2016). *Variational estimators in statistical multiscale analysis*. PhD thesis, Georg-August-Universität Göttingen.

Li, H., Guo, Q., and Munk, A. (2017). Multiscale change-point segmentation: Beyond step functions. *arXiv preprint arXiv:1708.03942*.

Luke, D. R. and Malitsky, Y. (2018). Block-coordinate primal-dual method for nonsmooth minimization over linear constraints. In *Large-Scale and Distributed Optimization*, pages 121–147. Springer.

Malgouyres, F. (2001). A unified framework for image restoration. Technical report, University of California, Los Angeles.

Malgouyres, F. (2002). Mathematical analysis of a model which combines total variation and wavelet for image restoration. *Journal of Information Processes*, 2(1):1–10.

Malitsky, Y. and Pock, T. (2018). A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432.

Mallat, S. (2008). *A wavelet tour of signal processing: the sparse way*. Academic press.

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

Mathé, P. and Pereverzev, S. V. (2003). Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789.

Meyer, Y. (1990). *Ondelettes et operateurs* I: Ondelettes. Hermann, (English translation: Wavelets and Operators. Cambridge, UK: Cambridge University Press, 1993.).

Meyer, Y. (2001). *Oscillating patterns in image processing and nonlinear evolution equations: the fifteenth Dean Jacqueline B. Lewis memorial lectures*, volume 22. American Mathematical Society.

Morozov, V. A. (1966). Regularization of incorrectly posed problems and the choice of regularization parameter. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 6(1):170–175.

Munk, A., Bissantz, N., Wagner, T., and Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):19–41.

Natterer, F. (1986). *The mathematics of computerized tomography*, volume 32. Siam.

Natterer, F. and Wübbeling, F. (2001). *Mathematical methods in image reconstruction*, volume 5. Siam.

Nemirovski, A. (1985). Nonparametric estimation of smooth regression functions. *Izvestiya Akademii Nauk SSR Tekhnieheskaya Kibernetika*, 3:50–60.

Nemirovski, A. (2000). Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85.

Nesterov, Y. and Nemirovsky, A. (1994). *Interior-point polynomial methods in convex programming*, volume 13. Studies in Applied Mathematics.

Nirenberg, L. (1959). On elliptic partial differential equations. In *Il principio di minimo e sue applicazioni alle equazioni funzionali*, pages 1–48. Springer.

Osher, S., Solé, A., and Vese, L. (2003). Image Decomposition and Restoration using Total Variation Minimization and the H^{-1} norm. *Multiscale Modeling & Simulation*, 1(3):349–370.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239.

Petsa, A. and Sapatinas, T. (2009). Minimax convergence rates under the L^p-risk in the functional deconvolution model. *Statistics & Probability Letters*, 79(13):1568–1576.

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

Picard, D. and Kerkyacharian, G. (2006). Estimation in inverse problems and second-generation wavelets. In *Proceedings of International Congress of Mathematicians: Madrid*, volume 3, pages 713–740.

Proksch, K., Werner, F., and Munk, A. (2018). Multiscale scanning in inverse problems. *The Annals of Statistics*, 46(6B):3569–3602.

Reed, M. and Simon, B. (1972). *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, San Diego.

Reiß, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *The Annals of Statistics*, 36(4):1957–1982.

Rockafellar, R. T. (2015). *Convex analysis*. Princeton University Press.

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

Sadhanala, V., Wang, Y.-X., and Tibshirani, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, pages 3513–3521.

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. (2009). *Variational Methods in Imaging*. Springer Science & Business Media.

Schmeisser, H.-J. and Triebel, H. (1987). *Topics in Fourier analysis and function spaces*. John Wiley & Sons.

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features. *The Annals of Statistics*, 41(3):1299–1328.

Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82(1):111–133.

Starck, J.-L., Candès, E. J., and Donoho, D. L. (2002). The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 11(6):670–684.

Starck, J.-L., Donoho, D. L., and Candes, E. J. (2001). Very high quality image restoration by combining wavelets and curvelets. In *Wavelets: Applications in Signal and Image Processing IX*, volume 4478, pages 9–20. International Society for Optics and Photonics.

Stein, E. M. and Weiss, G. (1971). *Introduction to Fourier analysis on Euclidean spaces*. Princeton University Press.

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559.

Tahmasebi, P., Javadpour, F., and Sahimi, M. (2016). Stochastic shale permeability matching: Three-dimensional characterization and modeling. *International Journal of Coal Geology*, 165:231–242.

Triebel, H. (1983). *Theory of Function Spaces*. Monographs in Mathematics, Volume 78, Birkhäuser-Verlag, Basel.

Triebel, H. (1988). Characterizations of Besov-Hardy-Sobolev spaces: a unified approach. *Journal of Approximation Theory*, 52(2):162–203.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667.

Wang, Y.-X., Sharpnack, J., Smola, A., and Tibshirani, R. J. (2016). Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. *arXiv preprint arXiv:1902.01778*.

Wunsch, C. (1996). *The ocean circulation inverse problem*. Cambridge University Press.