Total variation multiscale estimators for linear inverse problems

Miguel del Álamo$^1$ and Axel Munk$^{1,2}$

$^1$Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

$^2$Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

May 22, 2019

Abstract

Even though the statistical theory of linear inverse problems is a well-studied topic, certain relevant cases remain open. Among these is the estimation of functions of bounded variation (BV), meaning $L^1$ functions on a $d$-dimensional domain whose weak first derivatives are finite Radon measures. The estimation of BV functions is relevant in many applications, since it involves minimal smoothness assumptions and gives simplified, interpretable cartoonized reconstructions. In this paper we propose a novel technique for estimating BV functions in an inverse problem setting, and provide theoretical guarantees by showing that the proposed estimator is minimax optimal up to logarithms with respect to the $L^q$-risk, for any $q \in [1, \infty)$. This is to the best of our knowledge the first convergence result for BV functions in inverse problems in dimension $d \geq 2$, and it extends the results of Donoho (1995) in $d = 1$. Furthermore, our analysis unravels a novel regime for large $q$ in which the minimax rate is slower than $n^{-1/(d+2\beta+2)}$, where $\beta$ is the degree of ill-posedness: our analysis shows that this slower rate arises from the low smoothness of BV functions. The proposed estimator combines variational regularization techniques with the wavelet-vaguelette decomposition of operators.

Keywords Inverse problems Minimax estimation Total variation Interpolation inequalities Wavelet-vaguelette

Mathematics Subject Classification (2010) 62G05 65J22 62G20

Contents

1 Introduction 2
  1.1 Multiscale total variation estimation 3
  1.2 Main result 4
  1.3 Related work 6

2 Results 7
  2.1 Notation 7
  2.2 Main results 8
  2.3 Proofs of the main theorems 11
    2.3.1 Proof of Theorem 13 11
2.3.2 Proof of Theorem \[\text{[1]}\] .......................... 13
2.4 Examples ................................................... 15
2.4.1 Radon transform ........................................ 15
2.4.2 Convolution .............................................. 16
2.5 Nonparametric inverse regression model ................. 16

3 Auxiliary analytical results 18

1 Introduction

We consider the problem of estimating a real-valued function \(f\) from observations of \(Tf\) in a white noise regression model (see e.g. Tsybakov (2008))

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

Here \(\mathbb{M}\) denotes an open subset of \(\mathbb{R}^d\), \(T: L^2(\mathbb{R}^d) \to L^2(\mathbb{M})\) (see Section 2.1.2 in Giné and Nickl (2015)). The domain \(\mathbb{M}\) on which the data \(Y\) is defined is given by the inverse problem under consideration. In case of regression or deconvolution, we may have \(\mathbb{M} = \mathbb{R}^d\), while for certain types of tomography we have \(\mathbb{M} = \mathbb{R} \times S^{d-1}\) (Natterer, 1986), where \(S^{d-1}\) denotes the \(d\)-dimensional unit sphere. 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 \(\sigma^2\) (see e.g. Brown and Low (1996), Reiss (2008) or Section 1.10 in Tsybakov (2008)). Specifically, the white noise model does not take into account discretization effects, thus simplifying the theoretical analysis (see however Section 2.5 for a discussion of this). In the following we will often refer to \(n\) as the sample size, keeping in mind that this is only an analogy.

In this setting, our goal is to reconstruct the function \(f\) from observations \(dY\) and quantify the error made as \(n\) grows. In order to do so, we assume that \(f\) is supported inside the unit hypercube \([0,1]^d\). This restriction is somewhat arbitrary; we merely need the support of \(f\) to be contained in a compact set. Additionally, we make the structural assumption that \(f\) is a function of bounded variation, written \(f \in BV\).

**Definition 1** (Functions of bounded variation). The space of functions of bounded variation \(BV\) consists of functions \(g \in L^1\) whose weak distributional gradient \(\nabla g = (\partial_x g_1, \cdots, \partial_x g_d)\) is a \(\mathbb{R}^d\)-valued finite Radon measure on \(\mathbb{R}^d\). The finiteness implies that the bounded variation seminorm of \(g\), defined by

\[|g|_{BV} := \sup \left\{ \int_{\mathbb{R}^d} g(x) \, \text{div}(h)(x) \, dx \mid h \in C^1(\mathbb{R}^d; \mathbb{R}^d), \|h\|_{L^\infty} \leq 1 \right\}, \]

is finite, where \(\text{div}(h) = \sum_{i=1}^d \partial_x h_i(x)\) 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 \(\mathbb{R}^d\).

Functions of bounded variation have been used manifold in imaging applications since their introduction in the seminal work by Rudin et al. (1992). The reason for their success is that
they produce cartoonized reconstructions with sharp edges, which eases interpretability and makes them suitable for applications as diverse as medical imaging, microscopy, astronomy and geology, to mention just a few (see Scherzer et al. (2009) and references therein). However, in spite of their widespread use, a statistical theory for the estimation of BV functions in inverse problems is still lacking. To the best of our knowledge, the only available result for minimax optimal reconstructions of BV functions in inverse problems is Donoho (1995). He introduced the wavelet-vaguelette decomposition (WVD) associated with an operator, and showed that thresholding the WVD yields minimax optimal reconstruction over a range of Besov spaces. His results cover the case of BV functions for $d = 1$ and $\beta$-smoothing operators with $\beta \in [0, 1/2)$, meaning operators whose singular values behave like $\kappa_j = O(2^{-j\beta})$ as $j \to \infty$. This includes convolution operators with smooth enough convolution kernels, among others. In contrast, there is no statistical guaranty for estimating BV functions in dimension $d \geq 2$, which covers the very relevant imaging applications.

In this paper we propose an estimator that combines variational regularization with the WVD and multiscale dictionaries. We show 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 Radon inversion and deconvolution.

1.1 Multiscale total variation estimation

We consider the variational estimator

$$
\hat{f}_n \in \arg \min_{g \in \mathcal{F}_n} |g|_{BV} \text{ subject to } \max_{\omega \in \Omega_n} |\langle u_\omega, Tg \rangle - \langle u_\omega, dY \rangle| \leq \gamma_n,
$$

where $\gamma_n$ is a threshold to be chosen, and we minimize over a set of functions $\mathcal{F}_n$ to be specified later. $\Omega_n$ is a finite set of indices, and $\{u_\omega\}$ is a vaguelette system associated to the operator $T$, meaning that $T^* u_\omega = \kappa_\omega \psi_\omega \forall \omega \in \Omega$ for a wavelet basis $\{\psi_\omega \mid \omega \in \Omega\}$ and generalized singular values $\kappa_\omega$ (see Assumption 1 for the details). The set $\Omega_n$ depends monotonically on the parameter $n$ in (1.1), which plays the role of the sample size: the larger $n$, the larger the set $\Omega_n$. The reason is that, if the observations $dY$ are very noisy ($n$ small), we do not want to include too many terms in (1.2), since $\hat{f}_n$ would then be dominated by the noise. Conversely, the smaller the noise level, the more observations we want to include in the data-fidelity in (1.2), which is then able to extract more information about $f$.

Notice that, by the definition of the vaguelettes, the data-fidelity in (1.2) 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. Hence, the data-fidelity in (1.2) amounts to denoising of wavelet coefficients, while the regularization term $|g|_{BV}$ ensures that $\hat{f}_n$ is well-behaved in the BV seminorm.

We deliberately pose the optimization problem (1.2) in constrained form, but emphasize its equivalence to the penalized form

$$
\hat{f} \in \arg \min_{g \in \mathcal{F}_n} |g|_{BV} + \lambda \max_{\omega \in \Omega_n} |\langle u_\omega, Tg \rangle - \langle u_\omega, dY \rangle|.
$$

Indeed, both forms are equivalent for suitable parameters $\gamma_n$ and $\lambda$, but these will depend on the data and cannot be transformed easily from one problem to the other. For the penalized formulation (1.3), the optimal $\lambda$ could then be chosen in a data-driven way (e.g. by cross validation (Wahba, 1977) or by a version of Lepskii’s balancing principle (Lepskii, 1991), see
e.g. Mathé and Pereverzev (2003) in the context of inverse problems). In the constrained formulation (1.2), the optimal $\gamma_n$ in (1.2) can be chosen in a universal, non data-dependent manner, see equation (2.7).

To see that, notice that the role of $\gamma_n$ is to decide which functions are allowed for the minimization problem (1.2): a smaller $\gamma_n$ would yield very few admissible functions, and conversely for larger $\gamma_n$. Since the best reconstruction we can hope for is the true regression function $f$, the optimal $\gamma_n$ would be the one that is large enough to let $f$ be a feasible function, but not larger. In this sense, note that $f$ satisfies the constraint in (1.2) precisely when

$$\max_{\omega\in\Omega_n} |\langle u_\omega, Tf \rangle - \langle u_\omega, dY \rangle| = \max_{\omega\in\Omega_n} \sigma \frac{\langle u_\omega, dW \rangle}{\sqrt{n}} \leq \gamma_n.$$ (1.4)

Assume for a moment that $u_\omega \in L^2$ with $\|u_\omega\|_{L^2} = 1$ for all $\omega$. Then the left-hand side behaves like the maximum of the absolute value of $\#\Omega_n$ standard normal random variables times $\sigma n^{-1/2}$. Consequently, we see that (1.4) holds asymptotically with probability one if we choose $\gamma_n \sim \sigma n^{-1/2} \sqrt{2 \log \#\Omega_n}$. This argument can be adapted to the case that the $u_\omega$ do not have norm one, as long as their norms remain bounded above and below by positive constants. We remark that this canonical choice of $\gamma_n$ makes the estimator in constrained form (1.2) more convenient from a practical point of view than the one in penalized form (1.3).

At this point we can argue why the choice of the data-fidelity term in (1.2) is in a sense optimal: if we had chosen it to be the maximum of weighted coefficients, these weights would appear in (1.3), which would then be the maximum of normals with different variances. The maximum would hence be dominated by the terms with larger variances, which would lead to overfitting (if small scales dominate) or oversmoothing (if the large scales dominate).

Finally, we argue that the multiscale data-fidelity in (1.2) is in a sense preferable over the $L^2$ data-fidelity, which acts globally on the residuals. Indeed, consider an estimator like (1.2) with an $L^2$ constraint, which would take the form

$$\sum_{\omega\in\Omega_n} |\langle u_\omega, Tg \rangle - \langle u_\omega, dY \rangle|^2 \leq \gamma_n$$

for some $\gamma_n$, where we used the fact that $\{u_\omega\}$ is a frame for $L^2$ (Donoho, 1995) to express the $L^2$ norm in terms of the vaguelette coefficients. Arguing as above, the optimal $\gamma_n$ is the one for which the true function $f$ satisfies the constraint. Plugging in $g = f$, the left-hand side is a $\chi^2$-distributed random variable, so $\gamma_n$ should be chosen as $\gamma_n \sim \sigma^2 \#\Omega_n/n$. The difference between the multiscale and $L^2$ constraints is now apparent:

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

where both constraints are on the vaguelette domain. If we assume that the number of constraints $\#\Omega_n$ grows polynomially in $n$ (see Assumption 1), then 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.

Before we turn to the discussion of the convergence properties of $\hat{f}_n$, let us discuss two potential limitations of our approach. First, not every operator $T$ has an associated vaguelette system $\{u_\omega\}$, as we use in (1.2). In fact, only reasonably homogeneous operators admit such a system (see Donoho (1995)). However, for our theory we do not need the whole generality of the WVD (see Assumption 1), and many relevant operators such as the Radon transform, convolution or integration satisfy our assumptions (see Examples 2 below).
The second limitation concerns the numerical solution of the optimization problem in (1.2), which in general is a non-smooth, high-dimensional optimization problem (since $n$ and $\#\Omega_n$ might be large). While classical techniques such as interior point methods (Nesterov and Nemirovsky, 1994) find their limitations here, the computation of (1.1) is meanwhile feasible due to recent progress in convex optimization, e.g. in primal-dual methods (Chambolle and Pock, 2011) and accelerations thereof (Malitsky and Pock, 2018), or semismooth Newton methods with the path-following technique (Clason et al., 2010). We will not elaborate on this issue further and postpone this to future work.

1.2 Main result

The main result of this paper states that the estimator (1.2) is minimax optimal (up to logarithmic factors) for estimating $BV$ functions in any dimension for certain inverse problems. In order to formulate our result we need to introduce some notation. For $L > 0$ define the intersection of a $BV$-ball of radius $L$ with an $L^\infty$-ball as

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

(1.5)

where $D(T) \subset L^2$ denotes the domain of the operator $T$. The reason for the support condition in (1.5) is the following: 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.

For given $d, \beta \geq 0$ and $q \in [1,\infty]$, define the number

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

(1.6)

Our main result (Theorems 3 and 4) can be stated informally as follows.

**Main Theorem (Informal).** For $d \in \mathbb{N}$ and $\beta \geq 0$, let $T$ have a WVD with singular values behaving as $\kappa_j = 2^{-j\beta}$ (see Assumption 1 in Section 2). Let the threshold $\gamma_n$ be as in (2.7) 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,

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

(1.7)

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 $L, \sigma, d$ and $T$.

The convergence rate in (1.7) is indeed minimax optimal over the class $BV_L$ up to the logarithmic factor, as it is the optimal rate over the smaller class of bounded Besov $B^1_{1,1}$ functions, see Theorem 4 and Section 2.1 for the definition of Besov spaces. The minimax rate $n^{-1/(d+2\beta+2)}$ is well-known for inverse problems when $q \leq 1 + 2/(d+2\beta)$ (see e.g. Cavalier (2011)). In contrast, the "slow" regime with rate $n^{-1/((d+2\beta) + 2)}$ for $q > 1 + 2/(d+2\beta)$ has been observed for the specific case $\beta = 0$ in density estimation (Goldenshluger and Lepskii, 2014) and nonparametric regression (Lepskii, 2013 and del Álamo et al., 2018) when estimating over anisotropic Nikol’skii classes $N^p_s$ and Besov classes $B^s_{p,t}$ with $s < d/p$. Moreover, the slow regime explains the recently observed phase transition in the $L^2$ minimax risk for estimating discretized $TV$ functions in the particular case $\beta = 0$, see Sadhanala et al. (2016). Our result extends these findings to linear inverse problems.
The proof of the minimax optimality of that rate is based on the construction of a set of alternatives in the smaller space \(B_{1,1}^{1} \subset BV\). Interestingly, the set of alternatives that attains the minimax rate is neither sparse nor dense: it presents blocks of signals at different locations. We conjecture that only estimators that incorporate a form of spatial adaptation can be minimax optimal in this regime, as the ones proposed in [Lepski (2015)], [del Alamo et al. (2018)] and in the present paper.

The proof of the Main Theorem is based on an upper bound on the \(L^q\)-risk with an interpolation inequality in terms of the \(BV\) norm and a Besov norm of negative smoothness,

\[
\|\hat{f}_n - f\|_{L^q} \leq C \|\hat{f}_n - f\|_{B_{\infty,\infty}^{d/2-\beta}} \|\hat{f}_n - f\|_{BV}^{d/2+2\beta} \quad \forall g \in B_{\infty,\infty}^{-d/2-\beta} \cap BV
\]

for any \(q \in [1, \frac{d+2\beta+2}{d+2\beta}], \ d \geq 2\). See Section 2.1 for the definition of Besov spaces. This inequality follows from a result by [Cohen et al. (2003)], proved by an analysis of the wavelet coefficients of \(BV\) functions. Since we have \(\hat{f}_n \in BV\) by construction, the \(BV\) norm in (1.8) is easily bounded by a constant. On the other hand, the Besov norm can be related to the constraint in the right-hand side of (1.2), and some analysis yields the bound

\[
\|\hat{f}_n - f\|_{B_{\infty,\infty}^{-d/2-\beta}} \leq C n^{-1/2} \log n
\]

with high probability. Plugging this expression in (1.8), we get the desired bound for the \(L^q\)-risk. The bound is extended to \(q > 1 + \frac{2}{d+2\beta}\) using Hölder’s inequality. For \(d = 1\) we proceed analogously with some modifications. See Section 2.3 for a complete proof.

### 1.3 Related work

Notwithstanding the success of \(BV\) functions in imaging applications (see [Rudin et al. (1992)] for the first reference), there are very few works that analyze the estimation of \(BV\) functions in a statistical setting. In nonparametric regression \((T = id)\), classical results ([Mammen and van de Geer (1997)] and [Donoho and Johnstone (1998)]) established minimax optimality results for estimation in dimension \(d = 1\), and recently a class of multiscale variational estimators was shown to perform optimally in any dimension ([del Alamo et al., 2018]), whose approach we generalize here to \(T \neq id\). 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 \(B_{s,t}^{p}\) and for a class of \(\beta\)-smoothing inverse problems. In the case relevant for \(BV\) \((s = p = 1)\), the minimax optimality holds for the range \(\beta < 1 - d/2\), i.e. for \(\beta\) 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 \(\beta\) nor on the dimension \(d\). On the other hand, we get a suboptimal logarithmic factor in (1.7), while [Donoho (1995)] achieves the exact optimal rate.

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 to 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 with multiscale dictionaries. We refer exemplarily to [Candès and Guo (2002)], [Dong et al. (2011)], [Frick et al. (2012)] and [Frick et al. (2013)] for an empirical analysis of such methods in simulations. Moreover, the
proof of our main result is based on the above mentioned interpolation technique: an interpolation inequality of the form (1.3) 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 by del Alamo et al. (2018) for the estimation of BV functions, employing a generalization thereof (Meyer, 2001). Cohen et al. (2003)). In that sense, the present work combines the tools developed in del Alamo et al. (2018) with the WVD from Donoho (1995), and it generalizes both results.

Organization of the paper

The rest of the paper is organized as follows. In Section 2 we state our assumptions and main theorems, and give their proofs. We also discuss the particular inverse problems of deconvolution and Radon inversion. The proofs of auxiliary results are given in Section 3.

2 Results

2.1 Notation

Basic notation. We denote the Euclidean norm of a vector \( v = (v_1, \ldots, v_d) \in \mathbb{R}^d \) by \(|v| := (v_1^2 + \cdots + v_d^2)^{1/2} \). For a real number \( x \), define \(|x| := \max\{m \in \mathbb{Z} | m \leq x\}\) and \([x] := \min\{m \in \mathbb{Z} | m > x\}\). The cardinality of a finite set \( X \) is denoted by \#X. We say that two sequences \( a_n \) and \( b_n \), \( n \in \mathbb{N} \), grow at the same rate, written \( a_n \asymp b_n \), if there are (potentially zero) constants \( c_1, c_2 \geq 0 \) such that \( c_1a_n \leq b_n \leq c_2a_n \) for all \( n \in \mathbb{N} \). Finally, we denote by \( C \) a generic positive constant that may change from line to line.

Wavelet bases. Let \( \{\psi_{j,k,e} \mid (j,k,e) \in \Lambda\} \) denotes a wavelet basis of \( L^2(\mathbb{R}^d) \) formed by tensorization of Daubechies wavelets (Daubechies, 1992) with \( D \) continuous partial derivatives and whose mother wavelet has \( R \) vanishing moments. Here \( j \geq 0 \) is a scale index, \( k \in \mathbb{Z}^d \) is a position index, and \( e = (e_1, \ldots, e_d) \in \{0,1\}^d \) denotes whether \( \psi_{j,k,e} \) is a mother or a father wavelet along each coordinate. We recall that one-dimensional Daubechies wavelets with \( R \) vanishing moments have support of size \( 2R-1 \) (with respect to the Lebesgue measure) and are \( 0.18 \cdot (R-1) \) times continuously differentiable (see Theorem 4.2.10 in Giné and Nickl (2015)). A D-smooth wavelet basis formed by tensorization of one-dimensional Daubechies wavelets needs to satisfy \( R = 1 + 6D \) in order to have \( |0.18 \cdot 6 \cdot D| > D \) continuous derivatives. Consequently, the mother and father wavelets have support of size \( (12D + 1)^d \).

In this work we will mainly deal with functions \( g \) supported inside the unit cube, \( \text{supp} \ g \subseteq [0,1]^d \). We will use their wavelet expansion intensively, so let us introduce the set of wavelets with nonzero overlap with the unit cube

\[ \Omega = \{(j,k,e) \in \Lambda \mid \text{supp} \ \psi_{j,k,e} \cap (0,1)^d \neq \emptyset\}. \tag{2.1} \]

For each \( n \in \mathbb{N}, n \geq 2 \), let

\[ \Omega_n := \{(j,k,e) \in \Omega \mid j \leq \lceil d^{-1} \log n \rceil \} \]

denote the set of indices of wavelets at scales rougher than \( \lceil d^{-1} \log n \rceil \). Since the wavelets at scale \( j = 0 \) have support of size \( (12D + 1)^d \), it follows that there are \( O(2^{(j+1)d}) \) indices \((j,k,e) \in \Omega \) at level \( j \), and hence the cardinality of \( \Omega_n \) is of the order \( \#\Omega_n \asymp 2^{d(d-1) \log n} \asymp n \).
**Besov spaces.** Let \( \{ \psi_{j,k,e} \} \) be a wavelet basis with \( D \) continuous partial derivatives and whose mother wavelet has \( R \) vanishing moments. For \( p, q \in [1, \infty] \) and \( s \in \mathbb{R} \) with \( \min\{R, D\} > |s| \), the Besov space \( B^s_{p,q}(\mathbb{R}^d) \) consists of all functions (or distributions) \( g \) with finite Besov norm

\[
\|g\|_{B^s_{p,q}} := \left( \sum_{j \geq 0} 2^{js/d} \left( \sum_{k \in \mathbb{Z}^d} \sum_{e \in \{0,1\}^d} |\langle \psi_{j,k,e}, g \rangle|^p \right)^{q/p} \right)^{1/q}.
\]  

(2.2)

We refer to Section 4.3 in Giné and Nickl (2015) for more details.

Finally, we define the Fourier transform of a function \( g \in L^1(\mathbb{R}^d) \) by

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

The Fourier transform can be extended as an operator to \( L^2 \) and, by duality, to distributions \( \mathcal{D}'(\mathbb{R}^d) \) (see e.g. Section 4.1.1 in Giné and Nickl (2015)).

### 2.2 Main results

We make the following assumptions on the operator \( T \).

**Assumption 1.** Let \( T : L^2(\mathbb{R}^d) \to L^2(\mathcal{M}) \) denote a bounded, linear operator. For \( \beta \geq 0 \), assume that the following hold:

- there is a wavelet basis \( \{ \psi_{j,k,e} \mid (j, k, e) \in \Lambda \} \) of \( L^2(\mathbb{R}^d) \) (see Section 2.1) with \( D \) continuous partial derivatives and whose mother wavelet has \( R \) vanishing moments, such that \( \min\{R, D\} > \max\{1, d/2 + \beta\} \);
- there is a set of functions \( \{ u_{j,k,e} \mid (j, k, e) \in \Lambda \} \subset L^2(\mathcal{M}) \), which we call vaguelette system, s.t.

\[
T^* u_{j,k,e} = \kappa_j \psi_{j,k,e} \quad \forall (j,k,e) \in \Lambda,
\]

(2.3)

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

\[
c_1 \leq \|u_\omega\|_{L^2} \leq c_2 \quad \forall \omega \in \Lambda
\]

for some real constants \( 0 < c_1 < c_2 \).

We remark that a vaguelette system as constructed in Donoho (1995) is a frame. However, we will not need that property in the following.

**Remark 1.**

a) Assumption 1 is slightly weaker than assuming that the operator \( T \) has a wavelet-vaguelette decomposition (WVD) (Donoho, 1995). In the following we nevertheless call \( \{ u_{j,k,e} \} \) a vaguelette system for simplicity.

b) As remarked in Section 2.1 we will only need the wavelets with nonzero overlap with the unit cube, which we index by the set \( \Omega \) in (2.1). In the following we index the vaguelettes accordingly.
c) 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 basis \( \{\psi_{j,k,e}\} \) (see Section 2.1 or Section 4.3 in Giné and Nickl (2015)).

d) Let \( \{\psi_{j,k,e}\} \) be a smooth enough wavelet basis. Then condition (2.3) implies that the inverse problem (1.1) is mildly ill-posed with degree of ill-posedness \( \beta \).

**Examples 2.** We list here some examples of operators satisfying Assumption 1.

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 wavelet basis, and the critical values are \( \kappa_j = 2^{-j} \). Fractional integration, iterated integration and higher dimensional integrals also define operators satisfying Assumption 1. We refer to Donoho (1995) for more details.

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

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

where the integral is taken over the hyperplane defined by vectors \( x \) satisfying \( x \cdot \theta = r \). See Section 2.4.1 for more details on how our estimator (2.5) works for the Radon transform.

c) The convolution operator

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

for a regular enough kernel \( K \in L^1(\mathbb{R}^d) \) satisfies Assumption 1. See Section 2.4.2 for the details.

d) The identity operator, in which case we are in the white noise regression model. We can take \( \{\psi_{j,k,e}\} \) to be a smooth enough wavelet basis, and the estimator (2.5) reduces (with minor modifications) to the multiscale total variation estimator analyzed in del Alamo et al. (2018). Besides some differences in the setting (here we estimate compactly supported functions, there periodic ones), the convergence rate that we prove here coincides for \( \beta = 0 \) with the result in del Alamo et al. (2018).

More generally, operators satisfying a certain homogeneity condition with respect to dilations have a WVD (see Donoho (1993) for a general result). Conversely, Assumption 1 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 1 with a vaguelette system \( \{u_{j,k,e}\} \). We denote
\[
\hat{f}_n \in \text{argmin}_{g \in F_n} |g|_{BV} \text{ subject to } \max_{\omega \in \Omega_n} \langle u_\omega, Tg \rangle - \langle u_\omega, dY \rangle \leq \gamma_n,
\]
as the \textit{multiscale total variation estimator} for the operator \( T \). In (2.5) we minimize over the set
\[
F_n = \{ g \in BV \cap L^\infty \mid |g|_{L^\infty} \leq \log n, \ \text{supp} \ g \subseteq [0, 1]^d \}.
\]
We use the convention that, whenever the feasible set of the problem (2.5) is empty (which happens with vanishing probability as \( n \) grows), see Remark 2, the estimator \( \hat{f}_n \) is set to zero.

The reason for requiring the support to be inside the closed unit cube in (2.6) is to make the set \( F_n \) closed. This is important for ensuring existence of a minimizer in (2.5) as the limit of a minimizing sequence.

Concerning the choice of the threshold \( \gamma_n \), let \( \sigma > 0 \) be as in (1.1), and let \( c_2 \) be the upper bound in Assumption 1. For \( \kappa > 0 \), we choose
\[
\gamma_n = \kappa c_2 \sigma \sqrt{\frac{2 \log \# \Omega_n}{n}}.
\]
Notice that the upper bound \( c_2 \) can be computed from the dictionary, as we do in the examples in Section 2.4.

Remark 2. Let us discuss the feasible set of the problem (2.5), which consists of the constraints
\[
\max_{\omega \in \Omega_n} \langle u_\omega, Tg \rangle - \langle u_\omega, dY \rangle \leq \gamma_n, \ |g|_{L^\infty} \leq \log n, \ \text{supp} \ g \subseteq [0, 1]^d.
\]
Here we assume that the observations \( dY \) arise from a function \( f \in BV_L \), as defined in (1.5). By Proposition 2 below and the choice (2.7) for \( \gamma_n \), the probability that the true regression function \( f \) satisfies the first constraint in (2.8) is not smaller than \( 1 - O(n^{1-\kappa^2}) \). As long as \( f \) satisfies the first constraint in (2.8), it also satisfies the others for \( n \) large enough (\( n \geq e^L \)), since we assume that \( f \in BV_L \). As a consequence, the feasible set of (2.5) is nonempty with probability of the order \( 1 - O(n^{1-\kappa^2}) \). 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 \( \hat{f}_n \).

Theorem 3. For \( d \in \mathbb{N} \), let \( T \) satisfy Assumption 1 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 (1.6).

a) Let \( \gamma_n \) be as in (2.7) with \( \kappa > 1 \). Then for any \( n \in \mathbb{N} \) with \( n \geq e^L \), the estimator \( \hat{f}_n \) in (2.6) with parameter \( \gamma_n \) satisfies
\[
\sup_{f \in BV_L} \| \hat{f}_n - f \|_{L^q} \leq C n^{-\vartheta_{q,\beta}} (\log n)^{3-\min\{d,2\}}
\]
for any \( q \in [1, \infty) \) with probability at least \( 1 - (\# \Omega_n)^{1-\kappa^2} \), for a constant \( C > 0 \) independent of \( n \), but depending on \( L, \sigma \) and \( d \).

b) Under the assumptions of part a), if \( \kappa^2 > 1 + 1/(d + 2\beta + 2) \), then
\[
\sup_{f \in BV_L} \mathbb{E}[\| \hat{f}_n - f \|_{L^q}] \leq C n^{-\vartheta_{q,\beta}} (\log n)^{3-\min\{d,2\}}
\]
holds for any \( q \in [1, \infty) \), \( n \) large enough and a constant \( C > 0 \) independent of \( n \).
Theorem 3 gives an upper bound for the expected error of $\hat{f}_n$. We now prove a matching lower bound. For that, we assume that $T$ satisfies

$$\|T\psi_{j,k,e}\|_{L^2} \leq c2^{-j\beta}\ |\forall (j,k,e) \in \Omega \tag{2.11}$$

for a constant $c > 0$, where $\{\psi_{j,k,e}\}$ is a wavelet basis of compactly supported wavelets. We remark that (2.11) is satisfied by any operator with a WVD (see Donoho (1995)).

**Theorem 4.** Consider the setting of Theorem 3 and assume that the operator $T$ admits a WVD. Then the minimax $L^q$-risk over $BV_L$ given observations (1.1) is lower bounded by $cn^{-d/2-\beta}$. In particular, the estimator (2.5) 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 $q \in [1, \infty)$.

### 2.3 Proofs of the main theorems

#### 2.3.1 Proof of Theorem 3

The proof of Theorem 3 relies on a variant of an interpolation inequality prove by Cohen et al. (2003).

**Proposition 1.** 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
  $$\|g\|_{L^q} \leq C \|g\|_{B^{d/2-\beta}_{\infty,\infty}}^{\beta} \|g\|_{BV}^{d+2\beta}$$
  holds for any $q \in [1, q^*]$ and any $g \in B^{d/2-\beta}_{\infty,\infty} \cap BV$ with 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
  $$\|g\|_{L^q} \leq C(\log n) \|g\|_{B^{d/2-\beta}_{\infty,\infty}}^{\beta} \|g\|_{BV}^{d+2\beta} + Cn^{-1} \|g\|_{L^\infty}^{\beta} \|g\|_{BV}^{d+2\beta}$$
  for any $q \in [1, q^*]$ and any $g \in L^\infty \cap BV$ with supp $g \subseteq [0, 1]^d$.

The proof of Proposition 1 is given in Section 3 below. Define the event

$$\mathcal{A}_n := \left\{ \max_{(j,k,e) \in \Omega_n} \left| \int_M u_{j,k,e}(x) dW(x) \right| \leq \frac{\sqrt{n}}{\sigma} \gamma_n \right\}, \tag{2.12}$$

where $\{u_{j,k,e}\}$ is the vaguelette system from Assumption 1.

**Proposition 2.** Let $\{u_{j,k,e}\}$ be a vaguelette system as described in Assumption 1. For any $n \in \mathbb{N}$ we have

$$\mathbb{P} \left( \max_{(j,k,e) \in \Omega_n} \left| \int_M u_{j,k,e}(x) dW(x) \right| \geq c_2 t \right) \leq \#\Omega_n e^{-t^2/2}$$

for any $t \geq 0$, where $c_2$ is the upper bound in Assumption 1.

**Proof.** The random variables $\epsilon_{j,k,e} := c_2^{-1} \int_M u_{j,k,e}(x) dW(x)$ are normal with variance smaller than one, since $\|u_{j,k,e}\|_{L^2} \leq c_2$ by the inequality in Assumption 1. By the union bound we have

$$\mathbb{P} \left( \max_{(j,k,e) \in \Omega_n} |\epsilon_{j,k,e}| \geq t \right) \leq \sum_{(j,k,e) \in \Omega_n} \mathbb{P}(|\epsilon_{j,k,e}| \geq t)$$
for any $t \geq 0$, and the probability in the right-hand side can be bounded as
\[
\mathbb{P}(|\epsilon_{j,k,e}| \geq t) \leq 2 \int_t^\infty e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} \leq 2 e^{-t^2} \int_t^\infty e^{-x^2/2+xt} \frac{dx}{\sqrt{2\pi}} = e^{-t^2/2}.
\]
In the first inequality, we bounded the probability that $|\epsilon_{j,k,e}| \geq t$ by the probability that a standard normal random variable is larger than $t$ in absolute value. This is justified by the fact that $\epsilon_{j,k,e}$ has variance smaller than one for all indices.

We begin with an auxiliary result for the proof of Theorem 3 which is essentially a regularity result for $\hat{f}_n$ conditionally on the event $A_n$ in (2.12). In the following proofs, $C > 0$ denotes a generic constant that may change from line to line.

**Proposition 3.** Let $\{\psi_{j,k,e}\}$ and $\{u_{j,k,e}\}$ denote the wavelet and vaguelet systems from Assumption 1. For $n \geq e^L$, let $\hat{f}_n$ denote the estimator (2.5) with parameter $\gamma_n$ given by (2.7). Then conditionally on the event $A_n$ in (2.12) we have
\[
\text{(i) } \|\hat{f}_n - f\|_{B_{\infty,\infty}^{-d/2-\beta}} \leq C \gamma_n + C \frac{\|f\|_{L^\infty} + \log n}{\sqrt{n}},
\]
\[
\text{(ii) } \|\hat{f}_n - f\|_{BV} \leq \|f\|_{L^\infty} + 2|f|_{BV} + \log n,
\]
for any $f \in BV \cap L^\infty(\mathbb{R}^d)$ with supp $f \subseteq [0,1]^d$, and a constant $C > 0$ independent of $n$, $f$ and $\hat{f}_n$.

**Proof.** For part (i), the definition of the Besov $B_{\infty,\infty}^{-d/2-\beta}$ norm in terms of wavelet coefficients (see Section 2.1) yields
\[
\|\hat{f}_n - f\|_{B_{\infty,\infty}^{-d/2-\beta}} = \max_{(j,k,e) \in \Omega} 2^{-j\beta} |\langle \psi_{j,k,e}, \hat{f}_n - f \rangle|
\leq \max_{(j,k,e) \in \Omega} 2^{-j\beta} |\langle \psi_{j,k,e}, \hat{f}_n \rangle| + \max_{(j,k,e) \in \Omega} 2^{-j\beta} |\langle \psi_{j,k,e}, f \rangle|
\leq \max_{(j,k,e) \in \Omega} 2^{-j\beta} |\kappa_j^{-1} T^* u_{j,k,e} \cdot \hat{f}_n - f \rangle| + C \max_{(j,k,e) \in \Omega} 2^{-j\beta} \|u_{j,k,e}\|_{L^1} \|\hat{f}_n - f\|_{L^\infty}
\leq C \frac{\|\hat{f}_n - f\|_{L^\infty}}{\sqrt{n}},
\]
where we used that $\kappa_j = 2^{-j\beta}$ and that $\|u_{j,k,e}\|_{L^1} \leq C 2^{-jd/2} \|u_{j,k,e}\|_{L^2}$ for Daubechies wavelets (which are supported on a compact set). The numerator in the second term can be bounded by $\|f\|_{L^\infty} + \log n$ by construction of $\hat{f}_n$, while the first term can be bounded as
\[
\max_{(j,k,e) \in \Omega} |\langle u_{j,k,e}, T \hat{f}_n - f \rangle|
\leq \max_{(j,k,e) \in \Omega} |\langle u_{j,k,e}, T \hat{f}_n \rangle| + \max_{(j,k,e) \in \Omega} |\langle u_{j,k,e}, T f \rangle| - \max_{(j,k,e) \in \Omega} \sigma \frac{\sigma}{\sqrt{n}} \int_M u_{j,k,e}(x) dW(x) \leq 2 \gamma_n
\]
conditionally on $\mathcal{A}_n$, where in the second inequality we used the definition of $\hat{f}_n$. This completes the proof of (i). The proof of (ii) is analogous to the proof of Proposition 4 in [del Álamo et al. (2018)], so we do not reproduce it here. \qed
Proof of part a) of Theorem 3. We prove the claim of part a) of Theorem 3 conditionally on the event $\mathcal{A}_n$ in (2.12), which by Proposition 2 happens with probability $P(\mathcal{A}_n) \geq 1 - (\#\Omega_n)^{1-\kappa^2}$. Consider first the case $d \geq 2$, which gives $q^* := 1 + 2/(d + 2\beta) \leq 2$. In this case, Proposition 1 gives the interpolation inequality

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

for $q \leq 1 + 2/(d + 2\beta)$. Conditionally on $\mathcal{A}_n$ and for $n \geq e^L$, Proposition 3 gives bounds for the terms in the right-hand side of (2.13), and putting the last three equations together then yields

$$
\|\hat{f}_n - f\|_{L^q} \leq C\left(\gamma_n + C \frac{\|f\|_{L^\infty} + \log n}{\sqrt{n}}\right)^{\frac{2}{d+2\beta+2}} \left(\|f\|_{L^\infty} + 2\|f\|_{BV} + \log n\right)^{\frac{d+2\beta}{d+2\beta+2}} \\
\leq Cn^{-\frac{1}{d+2\beta+2}} \left(\sqrt{\log \#\Omega_n} + \log n\right)^{\frac{2}{d+2\beta+2}} (L + \log n)^{\frac{d+2\beta}{d+2\beta+2}} \\
\leq Cn^{-\frac{1}{d+2\beta+2}} \log n
$$

using that $f \in BV_L$. Since $\#\Omega_n$ grows linearly in $n$ (recall Section 2.1), the claim follows.

For the case when $d = 1$ and $\beta \geq 1/2$, we have $q^* \leq 2$ and the argument goes through as above.

Finally, the case $d = 1$ and $\beta < 1/2$ requires a special treatment, since then $q^* > 2$. We use part b) of Proposition 1 which gives

$$
\|\hat{f}_n - f\|_{L^q} \leq C(\log n) \|\hat{f}_n - f\|_{B^{\beta/2-\beta}_{\infty,\infty}} \|\hat{f}_n - f\|_{BV} + Cn^{-1} \|\hat{f}_n - f\|_{L^\infty} \|\hat{f}_n - f\|_{BV} \tag{2.14}
$$

for a constant $C > 0$ and any $q \leq q^*$. Conditionally on $\mathcal{A}_n$, we bound the terms in the right-hand side by Proposition 3 which for $n \geq e^L$ yields

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

which gives the claim.

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

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

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

Proof of part b) of Theorem 3. It follows from the convergence conditionally on $\mathcal{A}_n$ proved in part a) of the theorem. We omit the proof, as it is analogous to the proof of part b) of Theorem 1 in [del Alamo et al. (2018)].

2.3.2 Proof of Theorem 4

Here we prove Theorem 4 by showing that the minimax rate over the smaller set

$$(B_{1,1}^1 \cap L^\infty)_L := \{g \in B_{1,1}^1 \mid \|g\|_{L^\infty} \leq L, \|g\|_{B_{1,1}^1} \leq L, \text{ sup} \ g \subseteq [0,1]^d \} \subset BV_L$$

with respect to the $L^q$-risk, $q \in [1,\infty)$, is not faster than $n^{-\varrho_{q,\beta}}$. The proof of this is well-known in the dense case $q < 1 + 2/(d + 2\beta)$, where $\varrho_{q,\beta} := \frac{1}{d+2\beta+2}$; it can be found e.g. in Chapter
10 of Härdle et al. (2012) for \( d = 1 \) and \( T = id \), so we do not reproduce it here. Indeed, the generalization from \( d = 1 \) to \( d \geq 2 \) is trivial. Concerning the difference between \( T = id \) and general \( T \), we show below how to adapt the construction of the alternatives in the case \( \gamma \geq 1 + 2/(d + 2\beta) \), which indicates how to proceed in the dense regime (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, we have not found a lower bound in the literature for the regime \( \gamma \geq 1 + 2/(d + 2\beta) \): only the construction in del Álamo et al. (2018) for the particular case \( \beta = 0 \) deals with that regime. Here we modify that proof and give a lower bound for general \( \beta \geq 0 \).

**Proof of Theorem 4.** The proof follows the proof of Theorem 2 in del Álamo et al. (2018) closely. **Construction of alternatives:** In the proof of Theorem 2 in del Álamo et al. (2018), a set of alternatives \( \mathcal{G} := \{ g^\epsilon | \epsilon \in \{-1, +1\}^{S_j} \} \) is constructed such that

\[
g^\epsilon := \gamma \sum_{(k, e) \in R_j} \varepsilon_{k, e} \psi_{j, k, e},
\]

where \( \gamma \asymp 2^{-dj/2} \) is the signal strength, \( \psi_{j, k, e} \) are orthonormal Daubechies wavelets, and \( (k, e) \in R_j \subseteq \{0, \ldots, 2^j - 1\}^2 \times E_j, \ E_j = \{0, \ldots, 2^j - 1\} \setminus \{0\} \), are indices such that \( S_j = \# R_j = 2^{j(d - 1)} \). These functions are chosen to satisfy \( \|g^\epsilon\|_{B^1_{1, 1}} \leq L, \|g^\epsilon\|_{L^\infty} \leq L \) and

\[
\delta := \inf_{\epsilon \notin \epsilon'} \|g^\epsilon - g^{\epsilon'}\|_{L^q} = 2\|\gamma \psi_{j, k, e}\|_{L^q} = 2\gamma 2^{j(d - 1)/d} \|\psi_j\|_{L^q} \asymp 2^{-jd/q}. \tag{2.15}
\]

**Lower bound:** We use now Assouad’s lemma for lower bounding the \( L^q \)-risk over \( (B^1_{1, 1} \cap L^\infty)_L \). We reproduce the claim (Lemma 10.2 in Härdle et al. (2012)) for completeness.

**Lemma 1.** For \( \epsilon \in \{-1, +1\}^{S_j} \) and \( (k, e) \in R_j \), define \( \varepsilon_{k, e} := (\varepsilon'_{(k, e_1)}), \ldots, \varepsilon'_{(k_{S_j}, e_{S_j})} \), where

\[
\varepsilon'_{(k', e')} = \begin{cases} 
\varepsilon_{(k, e)} & \text{if } (k', e') \neq (k, e), \\
-\varepsilon_{(k, e)} & \text{if } (k', e') = (k, e).
\end{cases}
\]

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

\[
\mathbb{P}_{T^g}(LR(Tg^{\varepsilon_{k, e}}, Tg^\epsilon) > e^{-\lambda}) \geq p_0, \quad \forall \epsilon, \forall n, \tag{2.16}
\]

where \( \mathbb{P}_{T^g} \) denotes the probability with respect to observations drawn from \( Tg^\epsilon \) in the white noise model \( (1.1) \), and \( LR(Tg^{\varepsilon_{k, e}}, Tg^\epsilon) \) denotes the likelihood ratio between the observations associated to \( Tg^{\varepsilon_{k, e}} \) and \( Tg^\epsilon \). Then any estimator \( \hat{f} \) based on observations \( (1.1) \) satisfies

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

where \( \delta \) is defined in \( (2.15) \).

**Verification of (2.16):** With the same argument as the proof of Theorem 2 in del Álamo et al. (2018), condition \( (2.16) \) holds provided that the Kullback-Leibler divergence between observations from two alternatives satisfies \( K(dP_{Tg^{\varepsilon_{k, e}}}, dP_{Tg^\epsilon}) \leq c \) for a small enough constant \( c > 0 \). A standard computation gives

\[
K(dP_{Tg^{\varepsilon_{k, e}}}, dP_{Tg^\epsilon}) = \frac{n}{2\sigma^2} \|Tg^{\varepsilon_{k, e}} - Tg^\epsilon\|_{L^2}^2 = \frac{n\gamma^2}{2\sigma^2} \|T\psi_{j, k, e}\|_{L^2}^2 \leq \frac{n\gamma^2 2^{-2j\beta}}{2\sigma^2}
\]
using (2.11), so choosing $\gamma^2 \approx 2^{-jd} \approx n^{-d/(d+\beta)}$ gives (2.16).

**Application of Lemma 1:** The conclusion of the lemma applies, and we can lower bound the $L^q$-risk over the class $(B_{1,1}^1 \cap L^\infty)_L$ by the risk over $G$, i.e.,

$$\sup_{f \in (B_{1,1}^1 \cap L^\infty)_L} \mathbb{E}_T f \| \hat{f} - f \|_{L^q} \geq \sup_{g^* \in G} \mathbb{E}_T g^* \| \hat{f} - g^* \|_{L^q} \geq \frac{e^{-\lambda} p_0}{2} \delta 2^{j\Delta/q}$$

for any estimator $\hat{f}$. Choosing as above $2^j \approx n^{1/(d+2\beta)}$, the definition (2.15) for $\delta$ gives the bound

$$\sup_{f \in (B_{1,1}^1 \cap L^\infty)_L} \mathbb{E}_T f \| \hat{f} - f \|_{L^q} \geq c \delta 2^{j\Delta/q} \approx 2^{-j(d-\Delta)/q} \approx n^{-d/(d+\beta)},$$

which completes the proof.

### 2.4 Examples

#### 2.4.1 Radon transform

Due to its application in nondestructive imaging, in particular in medical 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 (2.4).

Functions in the range of $T$ are supported on cylindrical sets of the form $M = \mathbb{R} \times [0, 2\pi)$. Moreover, the domain of $T$ consists of functions $g \in L^2(\mathbb{R}^d)$ whose Fourier transform satisfies $|\xi|^{-1/2} \mathcal{F}[g](\xi) \in L^2$, see Donoho (1995). 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 $(j,k,e) \in \Omega$, define the vaguelettes by

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

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 generalized critical 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 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).

Altogether, the observations above imply that the Radon transform satisfies Assumption 1 with $\beta = 1/2$ in dimension $d = 2$. By Theorem 4, the multiscale total variation estimator (2.5) 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 \( \text{[Natterer, 1986]} \), as well as for other tomography operators, such as photoacoustic and thermoacoustic tomography (see e.g. \textit{Haltmeier} (2013)).

2.4.2 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 \( M = \mathbb{R}^d \), and by Young’s inequality \( T \) is a bounded operator from \( D(T) = L^2(\mathbb{R}^d) \) 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. \textit{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 (2.5) 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

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

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 \( \min\{R,D\} > \max\{1, d/2 + \beta\} \), define the system of functions

\[
u_{j,k,e}(x) := 2^{j(d/2-\beta)} \left[\frac{F[\psi_{0,0,e}](\cdot)}{F[K](\cdot)}\right] (2^j x - k)
\]

indexed by the set \( \Omega \) in (2.1). These functions satisfy the following relations

\[
T^* u_{j,k,e} = \kappa_j \psi_{j,k,e} \quad \text{where} \quad \kappa_j = 2^{-j\beta},
\]

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

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 3 for the proof). These results show that the convolution operator \( T \) under the assumptions above satisfies Assumption 1. By Theorem 4 we conclude that the multiscale TV-estimator is minimax optimal for estimating functions \( f \in BV_L \), up to logarithmic factors.

2.5 Nonparametric inverse regression model

So far we have discussed the estimator \( \hat{f}_n \) based on observations from the white noise model (1.1). In practice, however, one naturally has access to discretely sampled data, which makes it more realistic to model the observations with the nonparametric regression model

\[
Y_i = Tf(x_i) + \sigma \epsilon_i, \quad x_i \in \Gamma_n, \quad i = 1, \ldots, n.
\]

Here we assume that \( n = m^d \) for some \( m \in \mathbb{N} \), and that the design points belong to an equidistant grid

\[
\Gamma_n := \left\{ \left( \frac{k_1}{m}, \ldots, \frac{k_d}{m} \right) \mid k_i \in \{1, \ldots, m\}, \ i = 1, \ldots, d \right\}.
\]
Of course, different grids may be used, depending on the operator $T$ and the domain $\mathcal{M}$ under consideration. For simplicity of the analysis, we assume in this section that $\mathcal{M} = (0, 1)^d$. This is the case when $T$ is the identity operator, a suitable convolution operator, or integration, to mention just a few examples. In (2.21), $\epsilon_i$ are independent standard normal random variables, and $\sigma > 0$ plays the role of the standard deviation of the noise.

Given observations (2.21), our goal is to estimate the function $f$. We do so by discretizing our construction of the multiscale TV-estimator from Definition [2]. Let $\{u_n^\omega | \omega \in \Omega_n\}$ be a dictionary of discretized vaguelettes, i.e., each $u_n^\omega$ is a vector of $n$ values

$$(u_n^\omega)_i = n^{-1/2} u_\omega(x_i) \quad \text{for } i = 1, \ldots, n, \quad x_i \in \Gamma_n,$$

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

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

for any $\omega \in \Omega_n$, i.e., so that the vectors $u_n^\omega$ have asymptotically unit norm in an $L^2$ sense.

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

$$\hat{f}_D \in \operatorname{argmin}_{g \in \mathcal{F}_n} \|g\|_{BV} \quad \text{subject to} \quad \max_{\omega \in \Omega_n} \left| \sum_{x_i \in \Gamma_n} (u_n^\omega)_i (Tg(x_i) - Y_i) \right| \leq \kappa c_2 \sigma \sqrt{2 \log \#\Omega_n}, \quad (2.22)$$

where $c_2 > 0$ is the upper frame constant for the continuous vaguelettes in Assumption [1].

We can now analyze the estimator (2.22) following the same strategy as we did in the white noise model. The only difference will be that, above, we related the constraint on the vaguelette coefficients to the Besov $B_{\infty, \infty}^{-d/2 - \beta}$ norm. Since here we only have access to the discretized vaguelette coefficients, there is an additional discretization error caused by the approximation of the vaguelette coefficients by their discretized counterparts. That error is given by

$$\delta_n := \sup_{g \in \mathcal{F}_n} \left| \int_{[0, 1]^d} u_\omega(x) Tg(x) \, dx - n^{-1} \sum_{i=1}^n u_\omega(x_i) Tg(x_i) \right|. \quad (2.23)$$

Proceeding as in the proof of Proposition [3] we see that $\hat{f}_D$ satisfies the error bounds

$$\begin{align*}
(i) \quad & \|\hat{f}_D - f\|_{B_{\infty, \infty}^{-d/2 - \beta}} \leq C \gamma_n + C \frac{\|f\|_{L^\infty} + \max\{L, \log n\}}{\sqrt{n}} + 2\delta_n, \\
(ii) \quad & \|\hat{f}_D - f\|_{BV} \leq \|f\|_{L^\infty} + 2\|f\|_{BV} + \max\{L, \log n\},
\end{align*}$$

conditionally on the event $\mathcal{A}_n$ in (2.12). Following the proof of Theorem [3] we get the result

$$\sup_{f \in BV_L} \mathbb{E} [\|\hat{f}_D - f\|_{L^q}] \leq C \max\{n^{-1/2}, \delta_n\}^{2\delta_n, \beta} (\log n)^{3-\min(d, 2)}$$

for $q \in [1, \infty)$ and $n$ large enough. Here we have the following trade-off: if $\delta_n$ is of smaller order that $n^{-1/2}$, then $\hat{f}_D$ attains the same rate as the multiscale TV-estimator based on observations from the white noise model. On the other hand, if $\delta_n$ is of bigger order than $n^{-1/2}$, the discretization error dominates and $\hat{f}_D$ performs worse than $\hat{f}_n$. The different performance of the multiscale TV-estimator in the white noise and in the nonparametric regression models hence boils down to a purely approximation theoretic question.

It remains now to bound the discretization error $\delta_n$. For that, notice that it is entirely determined by the smoothness of $u_\omega Tg$. Recall that $g \in BV \cap L^\infty$ and that $T$ is a smoothing operator. Consider the following examples.
1) Let \( T \) be the identity operator. Then \( u_\omega = \psi_\omega \) is a smooth wavelet basis, and \( Tg \in BV \cap L^\infty \). Consequently, the product \( u_\omega Tg \) is at most a function of bounded variation, for which we have \( \delta_n = O(n^{-1/d}) \) (see e.g. Chapter 5 in [Evans and Gariepy (2015)]). In this case, the discretization error is of lower order for \( d = 1, 2 \), while it dominates for \( d \geq 3 \).

2) In particular cases, for \( T = id \), the error in \( d \geq 3 \) can be improved. For instance, if \( g \) is a piecewise constant function and if \( u_\omega \) is smooth enough and has vanishing moments. In that case, the discretization error can be of smaller order due to the vanishing moments of \( \psi_\omega \). We do not pursue this idea further.

3) If \( T \) is a convolution operator as in Section 2.4.2, then by Fourier inversion we can show that \( u_\omega \) is continuous. Moreover, if the kernel decays fast enough, \( Tg \) will be a continuous function as well, and so will be \( u_\omega Tg \). Hence we have the same discretization error \( \delta_n = O(n^{-1/d}) \) as above. There is nevertheless an important caveat here: as opposed to wavelets, vaguelettes do not have in general vanishing moments. Consequently, this error cannot be improved by assuming that \( Tg \) is e.g. piecewise constant.

We have argued that the difference between the multiscale TV-estimator in the white noise and the nonparametric inverse regression models arises from a discretization error. In particular, the error appears in the convergence rate in the nonparametric regression model, eventually making it slower. Importantly, for \( d = 2 \) the error behaves as \( \delta_n = O(n^{-1/2}) \), and so the multiscale estimator attains the optimal convergence rate \( n^{-\vartheta_{q,\beta}} \) for imaging problems in the discretized model (2.21).

More generally, the difference between the white noise and the nonparametric inverse problem models can be measured with the theory of asymptotic equivalence. While that theory is well understood when \( T = id \) ([Brown and Low (1996), Reiss (2008)]), there are considerably fewer results for general operators \( T \) (see [Grama and Nussbaum (1998) and Meister (2011)]). In particular, [Meister (2011)] proves asymptotic equivalence in a functional linear regression model provided that the unknown function is suitably smooth, which is reminiscent of our analysis above to control \( \delta_n \) based on the smoothness of \( Tg \).

3 Auxiliary analytical results

For simplicity, we prove the two parts of Proposition 1 separately. They rely on an interpolation inequality proved by [Cohen et al. (2003)] which we reproduce here.

**Theorem 5** (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

\[
\|g\|_{B_{q,t} \cap \mathcal{L}^\infty} \leq C \|g\|^{1-\theta}_{B_{p,t}^s} \|g\|^\theta_{BV}
\]

for any function \( g \in BV \cap \mathcal{L}^\infty \) and a constant \( C > 0 \) depending on \( p, q, s \) and \( d \) only.

**Proof of part a) of Proposition 1** First, Theorem 5 with \( s = -d/2 - \beta \) and \( p = \infty \) gives

\[
\|g\|_{B_{q,t} \cap \mathcal{L}^\infty} \leq C \|g\|^{2/(2d+4\beta)}_{B_{d/2+\beta} \cap \mathcal{L}^\infty} \|g\|^{d+2\beta/(2d+4\beta)}_{BV}
\]


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

\[
B^0_{r,r}(\mathbb{R}^d) \hookrightarrow L^r(\mathbb{R}^d)
\]  

(3.2)

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

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

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 (3.2). The extension to the \( L^1 \)-risk follows by compact support. \( \square \)

The proof of part b) of Proposition 1 relies on the following result.

**Proposition 4.** Let \( g \in L^\infty \cap BV \) satisfy \( \text{supp} \ g \subseteq [0, 1]^d \), and let \( q \in [2, 3] \). Then for any \( J \in \mathbb{N} \) we have

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

for a constant \( C > 0 \) independent of \( g \).

The proof of Proposition 4 uses the following lemma.

**Lemma 2.** Let \( \{\psi_{j,k,e} \mid (j, k, e) \in \Omega\} \) 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

\[
\int_{\mathbb{R}^d} \left| \sum_{(k,e) \in P^d_j \times E_j} c_{j,k,e} \psi_{j,k,e}(x) \right|^q dx \leq C_{\psi,q} 2^{jd(1/2-1/q)} \sum_{(k,e) \in P^d_j \times E_j} |c_{j,k,e}|^q
\]

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

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

**Proof of Lemma 2.** 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

\[
A_j : l^q(P^d_j \times E_j) \to L^q(\mathbb{R}^d)
\]

\[
\{c_{j,k,e}\}_{(k,e) \in P^d_j \times E_j} \mapsto \sum_{(k,e) \in P^d_j \times E_j} c_{j,k,e} \psi_{j,k,e}
\]

which gives the claim for all \( q \in [2, 3] \). The claim for \( q = 2 \) follows by the orthonormality of the wavelet basis. For \( q = 3 \), the claim follows with the same argument as Lemma 2 in [del Álamo et al. (2018)]: the only difference is that the functions there are defined on the torus \( \mathbb{T}^d \), and here on the cube \([0, 1]^d\). This completes the proof. \( \square \)
Proof of Proposition 4: Let \( \{ \psi_{j,k,e} \} \) be a basis of compactly supported wavelets. Writing \( g \) formally as its wavelet series we have for any \( q \in [2, 3] \)

\[
\|g\|_{L^q} = \left\| \sum_{j \in \mathbb{N}} \sum_{k,e} c_{j,k,e} \psi_{j,k,e} \right\|_{L^q} \leq \left\| \sum_{j \leq J} \sum_{k,e} c_{j,k,e} \psi_{j,k,e} \right\|_{L^q} + \left\| \sum_{j > J} \sum_{k,e} c_{j,k,e} \psi_{j,k,e} \right\|_{L^q} \tag{3.3}
\]

for any \( J \in \mathbb{N} \). Since \( \text{supp } g \subseteq [0, 1]^d \), the sums are over \((k,e) \in P_J \times E_j\). Using Lemma 2 the first term can be bounded as

\[
\left\| \sum_{j \leq J} \sum_{k,e} c_{j,k,e} \psi_{j,k,e} \right\|_{L^q} \leq \sum_{j \leq J} \left( C_{\psi,q} 2^{j/(d-1)} \sum_{(k,e)} |c_{j,k,e}|^q \right)^{1/q} \leq C_{\psi,q}^{1/q} \left( \max_{j \leq J} 2^{j/(d-1)} \sum_{(k,e)} |c_{j,k,e}|^q \right)^{1/q} \leq C_{\psi,q}^{1/q} J \|g\|_{B^0_{q,q}},
\]

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

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

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 2 and these bounds, the second term in (3.3) can be bounded as

\[
\left\| \sum_{j > J} \sum_{k,e} c_{j,k,e} \psi_{j,k,e} \right\|_{L^q} \leq \sum_{j > J} \left( C_{\psi,q} 2^{j/(d-1)} \sum_{(k,e) \in P_J \times E_j} |c_{j,k,e}|^q \right)^{1/q} \leq C_{\psi,q}^{1/q} \sum_{j > J} \left( 2^{j/(d-1)} 2^{-jd(q-1)}/2 \|g\|_{L^\infty}^{-1} 2^{j(d-1)/2-1} \|g\|_{BV} \right)^{1/q} \leq C_{\psi,q}^{1/q} \|g\|_{L^\infty}^{-1/2} \|g\|_{BV}^{1/2} \sum_{j > J} 2^{-j/q},
\]

which gives the claim.

Proof of part b) of Proposition 5: Let \( q^* := 1 + 2/(d + 2\beta) \) and assume that \( q^* > 2 \). Notice that \( q^* \leq 3 \) for \( d \in \mathbb{N} \) and \( \beta \geq 0 \). The claim follows from Theorem 2 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 4 below. Choosing \( J = \lfloor q^* \log n \rfloor \) yields the claim.

Proposition 5: In the setting of Section 2.3.2 we have

\[
T^* u_{j,k,e} = \kappa_j \psi_{j,k,e} \quad \text{where} \quad \kappa_j = 2^{-j\beta},
\]

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

where we can choose \( c_1 = \min_{e \in \{0, 1\}^d} \|(-\Delta)^{d/2} \psi_{0,0,e}\|_{L^2} \) 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

$$
\mathcal{F}[u_{j,k,e}](\xi) = 2^{-jd/2-j\beta} e^{-i\xi_j k - j} \frac{\mathcal{F}[\psi_{0,0,e}](2^{-j}\xi)}{\mathcal{F}[K](-\xi)}.
$$

(3.4)

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.

$$
\|u_{j,k,e}\|_{L^2}^2 = \|\mathcal{F}[u_{j,k,e}]\|_{L^2}^2 = 2^{-jd-2j\beta} \int_{\mathbb{R}^d} \left| \frac{\mathcal{F}[\psi_{0,0,e}](2^{-j}\xi)}{\mathcal{F}[K](-\xi)} \right|^2 \frac{d\xi}{(2\pi)^d}
\leq 2^{-jd-2j\beta} \int_{\mathbb{R}^d} (1 + |\xi|^2)^{\beta} \left| \mathcal{F}[\psi_{0,0,e}](2^{-j}\xi) \right|^2 d\xi,
$$

(3.5)

where in the second line we used the bounds (2.19) on the Fourier transform of the kernel $K$. The expression in the right-hand side can now be easily bounded from below as

$$
2^{-2j\beta} \int_{\mathbb{R}^d} (1 + |2^j \xi|^2)^{\beta} \left| \mathcal{F}[\psi_{0,0,e}](\xi) \right|^2 d\xi \geq 2^{-2j\beta} \int_{\mathbb{R}^d} |2^j \xi|^2 \left| \mathcal{F}[\psi_{0,0,e}](\xi) \right|^2 d\xi = \|\xi^{\beta} \mathcal{F}[\psi_{0,0,e}]\|_{L^2}^2 = \|(-\Delta)^{\beta/2} \psi_{0,0,e}\|_{L^2}^2,
$$

again by Plancherel’s theorem. On the other hand, the right-hand side of (3.5) can be upper-bounded as

$$
2^{-2j\beta} \int_{\mathbb{R}^d} (1 + |2^j \xi|^2)^{\beta} \left| \mathcal{F}[\psi_{0,0,e}](\xi) \right|^2 d\xi \leq 2^{-2j\beta} \int_{\mathbb{R}^d} (2^{2j} + |2^j \xi|^2)^{\beta} \left| \mathcal{F}[\psi_{0,0,e}](\xi) \right|^2 d\xi
= \|(1 + |\xi|^2)^{\beta/2} \mathcal{F}[\psi_{0,0,e}]\|^2_{L^2} = \|\psi_{0,0,e}\|_{H^\beta}^2.
$$

This yields the claim. \qed

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [RTG 2088-B2 to M.A., CRC 755-A4 to A.M.].

Acknowledgment

The authors thank Dr. Housen Li and Dr. Frank Werner for helpful discussions.

References

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. C. R. Math. Acad. Sci. Paris, 296(23):1021–1024.

Bertero, M., Boccacci, P., Desiderà, G., and Vicidomini, G. (2009). Image deblurring with Poisson data: from cells to galaxies. Inverse Problems, 25(12):123006.
Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398.

Candès, E. J. and Donoho, D. L. (2002). Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Ann. Statist.*, 30(3):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.

Cavalier, L. (2011). Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, pages 3–96. Springer Berlin Heidelberg.

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145.

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 J. Imaging Sci.*, 3(2):199–231.

Cohen, A., Dahmen, W., Daubechies, I., and DeVore, R. (2003). Harmonic analysis of the space BV. *Rev. Mat. Iberoam.*, 19(1):235–263.

Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61. Society for Industrial and Applied Mathematics, Philadelphia.

del Álamo, M., Li, H., and Munk, A. (2018). Frame-constrained total variation regularization for white noise regression. *arXiv preprint arXiv:1807.02038*.

Dong, Y., Hintermüller, M., and Rincon-Camacho, M. M. (2011). Automated regularization parameter selection in multi-scale total variation models for image restoration. *J. Math. Imaging Vision*, 40(1):82–104.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921.

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. *Electron. J. Stat.*, 6:231–268.

Frick, K., Marnitz, P., and Munk, A. (2013). Statistical multiresolution estimation for variational imaging: With an application in Poisson-biophotonics. *J. Math. Imaging Vision*, 46(3):370–387.

Giné, E. and Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press.

Goldenshluger, A. and Lepskii, O. (2014). On adaptive minimax density estimation on $\mathbb{R}^d$. *Probab. Theory Related Fields*, 159(3-4):479–543.
Grama, I. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields*, 111(2):167–214.

Grasmair, M., Li, H., and Munk, A. (2018). Variational multiscale nonparametric regression: smooth functions. *Ann. Inst. Henri Poincaré Probab. Stat.*, 54(2):1058–1097.

Haltmeier, M. (2013). Inversion of circular means and the wave equation on convex planar domains. *Comput. Math. Appl.*, 65(7):1025–1036.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012). *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media.

Lepski, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, 35(3):454–466.

Lepski, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.*, 43(3):1178–1242.

Malitsky, Y. and Pock, T. (2018). A first-order primal-dual algorithm with linesearch. *SIAM J. Optim.*, 28(1):411–432.

Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *Ann. Statist.*, 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.

Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.*, 39(3):1471–1495.

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.

Munk, A., Bissantz, N., Wagner, T., and Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 67(1):19–41.

Natterer, F. (1986). *The mathematics of computerized tomography*, volume 32. Siam.

Nemirovski, A. (1985). Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet.*, 3:50–60.

Nesterov, Y. and Nemirovsky, A. (1994). Interior-point polynomial methods in convex programming. *Stud. Appl. Math.*, 13.

Nirenberg, L. (1959). On elliptic partial differential equations. In *Il principio di minimo e sue applicazioni alle equazioni funzionali*, pages 1–48. Springer.

Proksch, K., Werner, F., and Munk, A. (2018). Multiscale scanning in inverse problems. *Ann. Statist.*, 46(6B):3569–3602.

Reiss, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982.
Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D*, 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.

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features. *Ann. Statist.*, 41(3):1299–1328.

Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivariate Anal.*, 82(1):111–133.

Stein, E. M. and Weiss, G. (1971). *Introduction to Fourier analysis on Euclidean spaces*, volume 32. Princeton University Press.

Triebel, H. (1983). *Theory of Function Spaces*, volume 78. Monographs in Mathematics, Birkhäuser-Verlag, Basel.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4):651–667.