The Annals of Statistics
2021, Vol. 49, No. 3, 1318–1346
https://doi.org/10.1214/20-AOS2001

FRAME-CONSTRAINED TOTAL VARIATION REGULARIZATION FOR WHITE NOISE REGRESSION

BY MIGUEL DEL ÁLAMO*, HOUSEN LI† AND AXEL MUNK‡

*Institute for Mathematical Stochastics, University of Göttingen, *miguel.del-alamo@mathematik.uni-goettingen.de; †housen.li@mathematik.uni-goettingen.de; ‡munk@math.uni-goettingen.de*

Despite the popularity and practical success of total variation (TV) regularization for function estimation, surprisingly little is known about its theoretical performance in a statistical setting. While TV regularization has been known for quite some time to be minimax optimal for denoising one-dimensional signals, for higher dimensions this remains elusive until today. In this paper, we consider frame-constrained TV estimators including many well-known (overcomplete) frames in a white noise regression model, and prove their minimax optimality w.r.t. L^q-risk ($1 \le q < \infty$) up to a logarithmic factor in any dimension $d \ge 1$. Overcomplete frames are an established tool in mathematical imaging and signal recovery, and their combination with TV regularization has been shown to give excellent results in practice, which our theory now confirms. Our results rely on a novel connection between frame-constraints and certain Besov norms, and on an interpolation inequality to relate them to the risk functional. Additionally, our results explain a phase transition in the minimax risk for BV functions.

1. Introduction. We consider the problem of estimating a real-valued function f from observations in the commonly used Gaussian white noise regression model (see, e.g., [3, 69] and [76])

$$dY(x) = f(x)\,dx + \frac{\sigma}{\sqrt{n}}\,dW(x), \quad x \in [0,1)^d. \tag{1}$$

Here, dW denotes the standard Gaussian white noise process in $L^2(\mathbb{T}^d)$, and we identify the d-torus $\mathbb{T}^d \sim \mathbb{R}^d/\mathbb{Z}^d$ with the set $[0,1)^d$, that is, to simplify the presentation we assume f to be a 1-periodic function. In Section 3, we present the extension to nonperiodic functions. To ease notation, we will henceforth drop the symbol $\mathbb{T}^d$, and write for instance, L^2 instead of $L^2(\mathbb{T}^d)$, and so on. The function f is assumed to be of bounded variation (BV), written $f \in BV$, meaning that $f \in L^1$ and its weak partial derivatives of first order are finite Radon measures on $\mathbb{T}^d$ (see Section 2.1 or Chapter 5 in [29]). Note that, for (1) to be well defined, we need to assume additionally that $f \in L^2$ if $d \ge 3$, since only in $d = 1, 2$ we have $f \in BV \subset L^2$. In the following, we assume that σ is known, otherwise it can be estimated $\sqrt{n}$-efficiently (see, e.g., [62] or [74]), which will not affect our results. In the following, we use the terms bounded variation (BV) and total variation (TV) indistinctly. The former is commonly used in analysis, while the latter appears in imaging.

Functions of bounded variation can have discontinuities, and are thus ideal to model objects with edges and abrupt changes. This is a desirable property for instance, in medical imaging applications, where sharp transitions between tissues occur, and smoother functions would represent them inadequately (see, e.g., [57] for a TV-based optical flow method in real

Received May 2019; revised January 2020.
MSC2020 subject classifications. Primary 62G05; secondary 62G20, 62M40.
Key words and phrases. Nonparametric regression, minimax estimation, total variation, interpolation inequalities, wavelets, overcomplete dictionaries.

time magnetic resonance imaging or [45] for its use in photoacoustic tomography). Consequently, BV functions have been studied extensively in the applied and computational analysis literature; see, for example, [9, 61, 70, 72] 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. With the exception of the one-dimensional case $d = 1$, where total variation (TV) penalized least squares [60] and wavelet thresholding [24] applied to BV functions are known to attain the minimax optimal convergence rate $O(n^{-1/3})$, 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. More generally, BV does not admit an unconditional basis (see Sections 17 and 18 in [61]).

Our goal in this paper is to fill that gap. We consider the continuous model (1) and present estimators for $f \in BV$ that are minimax optimal up to logarithmic factors in any dimension, that is, they attain the polynomial rate $n^{-1/(d+2)}$ for the L^q-risk, $q \in [1, 1 + 2/d]$, and the rate $n^{-1/dq}$ for $q \in [1 + 2/d, \infty)$. While the first regime is well known (e.g., for $d = 1$ and $q = 2$; see again [60] and [24]), much less attention has been paid to the second regime. We mention [36] and [52] for estimation over anisotropic Nikolskii classes, which in the isotropic case coincide with Besov spaces $B^s_{p,\infty}$, and [71] for the case of discrete total variation when $q = 2$ (see "Related work" later in this section for a comprehensive discussion). These risk regimes explain the recently observed phase transitions in discrete TV-regularization [71] and componentwise isotone estimation [21, 42] (see Figure 1 and the remarks after the main theorem in the Introduction for more details). As a remarkable statistical consequence,, we also show that there is no L^∞-consistent estimator of BV functions.

The estimators that achieve these rates are not a straightforward extension of those for $d = 1$ [60]. There it is sufficient to penalize a *global* least-squares data-fidelity term by the TV function,

$$\hat{f}_{\lambda_n} \in \arg\min_g \|g - Y\|_2^2 + \lambda_n |g|_{BV} \tag{2}$$

for a suitable sequence of Lagrange multipliers λ_n, where $|g|_{BV}$ denotes the BV-seminorm of g (Section 2.1). Instead, we consider estimators that combine the strengths of TV and *multiscale* data-fidelity constraints. Multiscale data-fidelity terms and the associated reconstructions by the corresponding dictionary are widely used since the introduction of wavelets (see, e.g., [17] and [23]), and specially for imaging tasks overcomplete frames such as curvelets [5], shearlets ([39, 46]) and other multiresolution systems (see [41] for a survey) have been

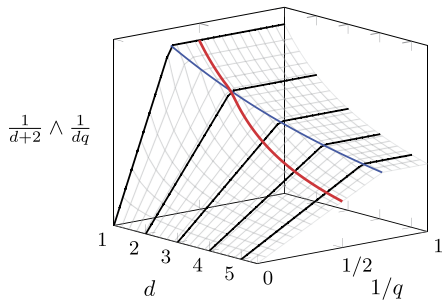

FIG. 1. *Exponent of the minimax rate over BV_L, $\min\{\frac{1}{d+2}, \frac{1}{dq}\}$, plotted as a function of $d \in \mathbb{N}$ and $1/q \in [0, 1]$. The line $1/q = d/(d + 2)$ is marked in blue, and the red line corresponds to the L^2-risk, $q = 2$. The phase transition observed in [71] for the L^2-minimax risk corresponds to the change of behavior of the red curve.*

shown to perform well in theory and numerical applications. In contrast, for the multiscale TV-estimators a theoretical understanding in a statistical setup when $d \geq 2$ is lacking, although its good empirical performance has been reported for specific choices of dictionaries in several places [7, 22, 30, 31]; see also Figure 2. Further, these methods were rarely used in routine applications, as they need large scale nonsmooth convex optimization methods for their computation. However, in the meantime such methods have become computationally feasible due to recent progress in optimization, the development of primal-dual algorithms [10] or semismooth Newton methods [11]. See Section 5, where we discuss the practical implementation of these estimators. Hence, we do see practical potential for such multiscale TV-methods, for which we give a theoretical justification in this paper in large generality.

Multiscale total variation estimators. Let $\Phi = \{\phi_\omega \mid \omega \in \Omega\} \subset L^2$ be a dictionary of functions indexed by a countable set Ω and satisfying $\|\phi_\omega\|_{L^2} = 1$, $\omega \in \Omega$. Consider the projection of the white noise model (1) onto Φ,

$$Y_\omega := \langle \phi_\omega, f \rangle + \frac{\sigma}{\sqrt{n}} \int_{\mathbb{T}^d} \phi_\omega(x)\, dW(x), \quad \omega \in \Omega, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in L^2. For each $n \in \mathbb{N}$, Φ and given the observations Y_ω, our estimator $\hat{f}_\Phi$ for f is defined as any solution to the constrained minimization problem

$$\hat{f}_\Phi \in \operatorname*{arg\,min}_{g \in X_n} |g|_{BV} \quad \text{subject to } \max_{\omega \in \Omega_n} \big|\langle \phi_\omega, g \rangle - Y_\omega\big| \leq \gamma_n. \tag{4}$$

Here, $X_n \subset BV$ is a suitable closed, convex set which may depend on n (see (15) for the definition). Hence, the existence of a minimizer is guaranteed by the convexity and lower-semicontinuity of the objective function and the constraint. The *finite* subsets $\Omega_n \subset \Omega$ indexing a proper sequence of subsets of the dictionary Φ will be specified later (see Assumption 1 and (6) below). For instance, if Φ is a wavelet basis, Ω_n corresponds to the wavelet coefficients at all scales j such that $2^{jd} \leq n$.

The constraint in (4) can be interpreted statistically as testing whether the data Y_ω is compatible with the coefficients $\langle \phi_\omega, \hat{f}_\Phi \rangle$, *simultaneously* for all $\omega \in \Omega_n$, an approach that dates back to [64]. This testing interpretation suggests how to choose the parameter γ_n in (4): the coefficients $\langle \phi_\omega, f \rangle$ of the truth should satisfy the constraint with high probability. This can be achieved by the *universal threshold*

$$\gamma_n(\kappa) = \kappa \sigma \sqrt{\frac{2 \log \#\Omega_n}{n}} \quad \text{for } \kappa > \kappa^* \tag{5}$$

with $\kappa^* > 0$ depending on d and the dictionary Φ in an explicit way (see Theorem 1). This universal choice of the parameter γ_n appears to us as a great conceptual and practical advantage of the estimator (4), in contrast to its penalized formulation, requiring more complex parameter-choice methods (e.g. [54] or [77]). In particular, γ_n in (5) can be precomputed using known or simulated quantities only.

The main conceptual contribution of this paper is to link the multiscale constraint in (4) and the Besov $B^{-d/2}_{\infty,\infty}$ norm. In fact, several dictionaries Φ used in practice have the following property: for each $n \in \mathbb{N}$ there is a finite subset $\Omega_n \subset \Omega$ such that

$$\|g\|_{B^{-d/2}_{\infty,\infty}} \leq C \max_{\omega \in \Omega_n} \big|\langle \phi_\omega, g \rangle\big| + C \frac{\|g\|_{L^\infty}}{\sqrt{n}} \tag{6}$$

holds for any function $g \in L^\infty$. This is a Jackson-type inequality [13], representing how well a function can be approximated in the Besov $B^{-d/2}_{\infty,\infty}$ norm by its coefficients with respect to Φ. It is well known that smooth enough wavelet bases satisfy this condition [13]. In

Section 2.3 we will show that (6) holds for more general multiscale systems, for example, systems of indicator functions of dyadic cubes, and mixed frames of wavelets and curvelets and of wavelets and shearlets.

For fixed $L > 0$, define the $BV \cap L^\infty$-ball of radius L,

$$BV_L := \{g \in BV \cap L^\infty \mid \|g\|_{L^\infty} \leq L, |g|_{BV} \leq L\}. \tag{7}$$

The main contribution of this paper (Theorems 1 and 2 in Section 2.2) can be informally stated as follows.

MAIN THEOREM (Informal). *Fix the dimension $d \in \mathbb{N}$, and let Φ satisfy an inequality of the form (6) (see Assumption 1 in Section 2.2). Let the threshold γ_n in (4) be as in (5). Then the estimator $\hat{f}_\Phi$ in (4) attains the minimax optimal rate of convergence over BV_L possibly up to a logarithmic factor ($(\log n)^2$ in $d = 1$ and $\log n$ else)*

$$\sup_{f \in BV_L} \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q}\big] \leq C_L \operatorname{polylog}(n) n^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}} \tag{8}$$

for n large enough, for any $q \in [1, \infty)$, any $L > 0$ and a constant $C_L > 0$ independent of n and q, but dependent on L, σ, d and Φ.

We remark that this reproduces the results by [71] for estimating BV functions in a discrete model for $q = 2$ (quadratic risk). Indeed, [71] shows that the minimax rate with respect to the empirical ℓ^2-risk scales as $n^{-\min\{\frac{1}{d+2}, \frac{1}{2d}\}}$. Our theorem explains this "phase transition" in the risk between $d \leq 2$ and $d > 2$ as arising from the low smoothness of BV functions and from the L^q-risk employed (see Figure 1 for an illustration of this).

Notably, the minimax rate in the Main Theorem for $q = 2$ also matches the minimax rate for estimating componentwise isotonic functions in the discrete regression model with respect to the empirical ℓ^q risk (see [42] for $q = 2$ and [21] for the general case $q \in [1, \infty)$). Remarkably, this means that the statistical complexity of estimating BV functions equals the complexity of estimating componentwise isotone functions. This result is well known in dimension $d = 1$, as a function of bounded variation can be written as the difference of two monotone functions, but we are not aware of any such result in $d \geq 2$. Moreover, this complements the recent finding that entirely monotone functions have the same statistical complexity as functions of bounded variation in the sense of Hardy–Krause [51]. We remark, however, that bounded variation in the sense of Hardy–Krause is a much stronger assumption than bounded variation in the sense that we use here (see "Related work" for a discussion).

The proof of (8) relies on the compatibility between the frame constraint and the $B^{-d/2}_{\infty,\infty}$ norm, as expressed in (6). Indeed, (6) allows us to relate the statistical multiscale constraint in (4) to an analytic object (the Besov norm). We can thus use techniques from harmonic analysis to analyze $\hat{f}_\Phi$, such as the interpolation inequality between $B^{-d/2}_{\infty,\infty}$ and BV [14],

$$\|g\|_{L^q} \leq C \|g\|^{\frac{2}{d+2}}_{B^{-d/2}_{\infty,\infty}} \|g\|^{\frac{d}{d+2}}_{BV} \quad \forall g \in B^{-d/2}_{\infty,\infty} \cap BV \tag{9}$$

for any $q \in [1, \frac{d+2}{d}]$, $d \geq 2$. This interpolation inequality relates the risk functional on the left-hand side with the data-fidelity and the regularization functionals on the right-hand side. It can be proven by a delicate analysis of the wavelet coefficients of functions of bounded variation (the original proof is in [14], and here we use an extension of (9) to periodic functions). The inequality (9) is the first step towards bounding the L^q-risk of $\hat{f}_\Phi$: inserting $g = \hat{f}_\Phi - f$ we can bound it in terms of the $B^{-d/2}_{\infty,\infty}$ and the BV-risks. It can be shown that the BV-risk

is bounded by a constant with high probability, while the $B^{-d/2}_{\infty,\infty}$-risk can be handled using inequality (6) as follows:

$$\begin{aligned}\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}} \le C \max_{\omega\in\Omega_n} |\langle \phi_\omega, \hat{f}_\Phi\rangle - Y_\omega| + C\frac{\sigma}{\sqrt{n}} \max_{\omega\in\Omega_n} \left|\int \phi_\omega(x)\,dW(x)\right| \\ + C\frac{\|\hat{f}_\Phi - f\|_{L^\infty}}{\sqrt{n}}.\end{aligned} \tag{10}$$

The first term is bounded by $\gamma_n = O(n^{-1/2}\sqrt{\log \#\Omega_n})$ as in (5) 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 of the estimator. The third term arises from the compatibility between Φ and the Besov space $B^{-d/2}_{\infty,\infty}$ stated in (6). Inserting the result in (9) (which requires $d \ge 2$) yields the conclusion that $\|\hat{f}_\Phi - f\|_{L^q} \le Cn^{-\frac{1}{d+2}} \log n$ with high probability. The bounds for $q \ge 1 + 2/d$ follow from Hölder's inequality between $L^{1+2/d}$ and L^∞. The proof for $d = 1$ follows the same lines, but it is slightly different. See Section 6 for the full proof.

The inequality (9) is sharp, in the sense that the norms in the right-hand side cannot *both* be replaced by weaker norms. In this sense, it is important that our estimator (4) combines a bound on the frame coefficients (related to the $B^{-d/2}_{\infty,\infty}$-norm) with control on the BV-seminorm. Finally, notice that the argument above does not rely on Gaussianity of the process dW: it holds whenever the random variables $\int \phi_\omega(x)\,dW(x)$ have sub-Gaussian tails.

EXAMPLE 1. In order to illustrate the performance of the estimator $\hat{f}_\Phi$, consider the situation where $d = 2$ and the dictionary Φ consists of normalized mollified indicator functions of dyadic squares [63],

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

where $|B|$ denotes the Lebesgue measure of the set B, $\widetilde{1}_B = \varphi_\varepsilon * 1_B$, $\varphi_\varepsilon = \varepsilon^{-d}\varphi(x/\varepsilon)$ denotes the standard mollifier with $\varphi(x) = \tilde{c}\exp\{-\frac{1}{1-|x|^2}\}1_{\{|x|<1\}}$, and $\varepsilon = cn^{-1/d}$ for $c < 1$. Now, the estimator $\hat{f}_\Phi$ in (4) has to be computed under the constraint

$$\max_{\text{dyadic } |B|\ge \frac{1}{n}} \frac{1}{\sqrt{|B|}} \left| \int \widetilde{1}_B(x)\big(g(x) - f(x)\big)\,dx - \frac{\sigma}{\sqrt{n}} \int \widetilde{1}_B(x)\,dW(x) \right| \le \gamma_n,$$

that is, Ω_n consists of all squares $B \subseteq [0,1]^2$ of area $|B| \ge 1/n$ with vertices at dyadic positions (see Section 2.3.2 for the details). In particular, the frame Φ satisfies Assumption 2 in Section 2.3.2, and hence also Assumption 1. The main peculiarity of $\hat{f}_\Phi$ is the data-fidelity term, which encourages proximity of $\hat{f}_\Phi$ to the truth f *simultaneously* at all 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 2 (see Section 5 for computational details): the estimator $\hat{f}_\Phi$ succeeds to reconstruct the image well at both the large (sky and building) and small scales (stairway). For comparison, we also show the classical TV-regularization estimator, a.k.a. Rudin–Osher–Fatemi (ROF) estimator [70], defined in (2), which employs a global L^2 data-fidelity term. The parameter λ_n in (2) is chosen in an oracle way so as to minimize the distance to the truth, which serves as a benchmark for any data-driven parameter choice. Here we measure the "distance" by the symmetrized Bregman divergence of the BV seminorm (see Section 3 of [30] for a motivation for this and other distances). The ROF estimator successfully denoises the image in the large scales at the cost of losing details in the small scales. The reason is simple: the use of the L^2 norm as a

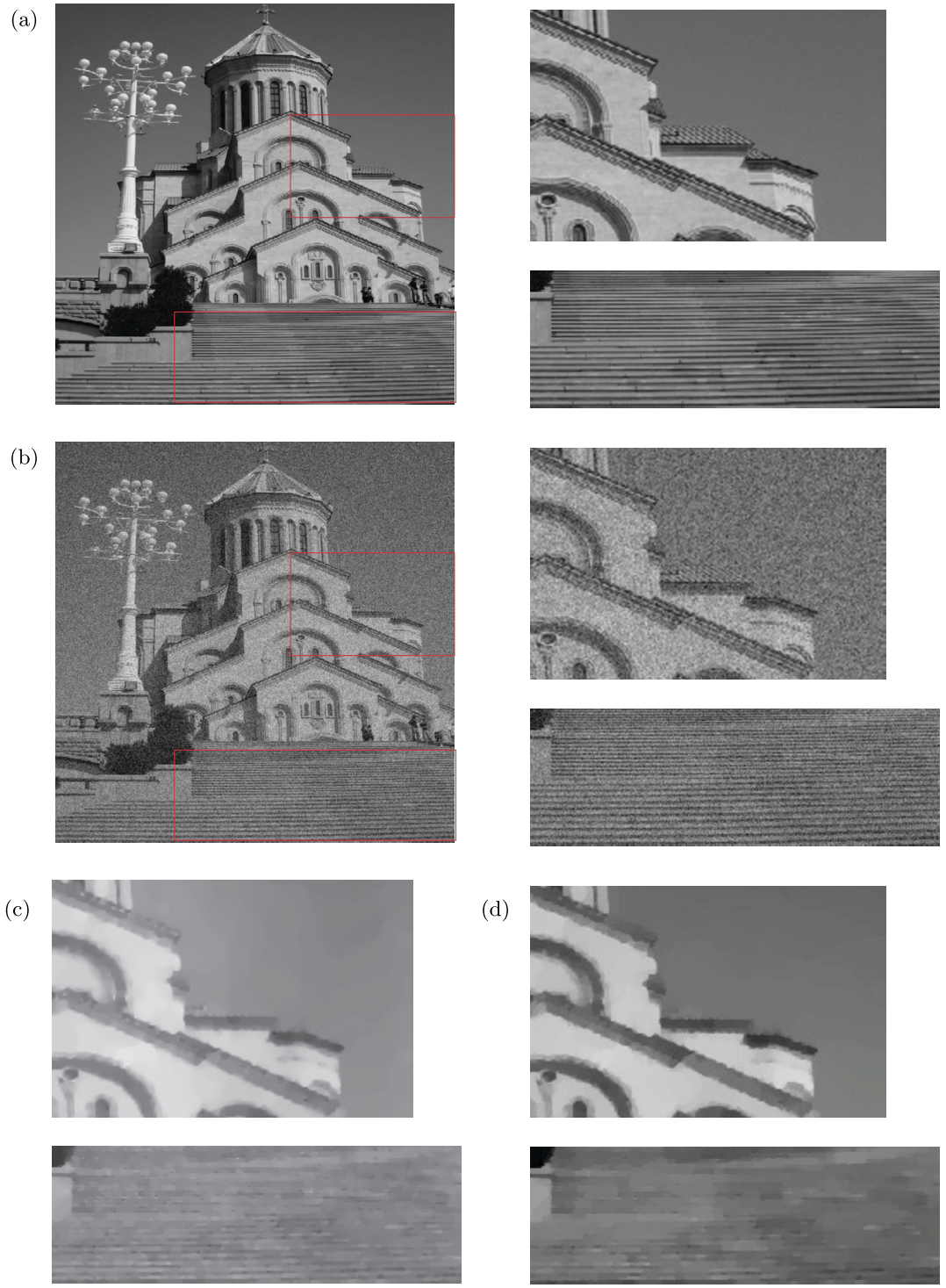

FIG. 2. (*a*) *Original image*, (*b*) *noisy version with signal-to-noise ratio* SNR $= 5$, (*c*) *zoom in of the multiscale TV estimator in Example* 1 *with* $\kappa = 1/2$ *in* (5), *and* (*d*) *zoom in of the estimator* $\hat{f}_{\lambda_n}$ *from* (2) *with oracle* $\lambda_n^* = \arg\min \mathbb{E}[D_{BV}(\hat{f}_{\lambda_n}, f)]$, *where* $D_{BV}(\cdot,\cdot)$ *denotes the symmetrized Bregman divergence of the BV seminorm*.

data-fidelity, which only measures the proximity to the data *globally*. As a consequence, the optimal parameter λ_n 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, for example, in the stairway in Figure 2(d).

OTHER EXAMPLES. Other estimators that minimize the BV seminorm and fall into our framework (4), covered by our theory, result from dictionaries Φ consisting of a wavelet basis ([23, 43]), a curvelet frame [5] or a shearlet frame [46]. Such estimators have been proposed in the literature [7, 30, 59] and have been shown to perform very well in numerical examples, outperforming wavelet and curvelet thresholding, and TV-regularization with global L^2 data-fidelity, as illustrated in Figure 2.

Related work. This paper is related to a number of results at the cutting edge of statistics, mathematical imaging and applied harmonic analysis. As the literature is vast, we only mention some selective references. Starting with the seminal paper [70] that proposed the TV-penalized least squares estimator (2) for image denoising (the ROF estimator), the subsequently developed theory of TV-based estimators depends greatly on the spatial dimension. In dimension $d = 1$, [60] showed that the ROF-estimator attains the optimal rate of convergence in the discretized nonparametric regression model, and [24] proved that wavelet thresholding for estimation over BV attains the minimax rates with the exact logarithmic factors. We also refer to [18] and [26] for a combination of TV-regularization with related multiscale data-fidelity terms in $d = 1$, and to [32] and [56] for the combination of a multiscale constraint with a jump penalty for segmentation of one-dimensional functions

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 that are related to our work. First, [61] 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 $B^{-1}_{\infty,\infty}$ [33], the G-norm [40] and the Sobolev norm H^{-1} in $d = 2$ [67] were proposed, but the statistical performance of these estimators has not been analyzed. A second variant (see [28, 58] and [59]) involved estimators of the form (4) with a wavelet basis Φ. Following this approach and the development of curvelets (see, e.g., [5] for an early reference), [7] and [25] proposed the estimator (4) with Φ being a curvelet frame and a mixed curvelet and wavelet family, respectively, which showed good numerical behavior. The third line of development that leads to the estimator (4) is based on Nemirovski's work [64], who credits S. V. Shil'man for the original idea (see also [63]), and on Donoho's work on soft-thresholding [23]. Nemirovski 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 (LR) tests: the *multiresolution norm*. In statistical inverse problems, [22] proposed an estimator using TV-regularization constrained by the *sum* of local averages of residuals, instead of the maximum we employ in (4), which was proposed by [30]. Finally, during revision of this work we became aware of the work by [51], who consider estimation of functions of bounded variation in the sense of Hardy–Krause. This class of functions has higher regularity than BV, and hence is much smaller: it corresponds roughly to Sobolev $W^{d,1}$ functions, that is, with d partial derivatives in L^1, which explains the faster minimax rate $n^{-1/3}$ in any dimension.

The other approach to TV-regularization in higher dimensions is to discretize the observational model (1), thereby reducing the problem of estimating a function $f \in BV$ to that of estimating a vector of function values $(f(x_1), \dots, f(x_n)) \in \mathbb{R}^n$, where $\{x_i\}$ are design points

in $[0,1]^d$. In particular, the risk is measured by the *Euclidean norm* ℓ^2 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 for the ℓ^2 risk. We mention [16] and [44], who proved convergence rates in any dimension d, which were shown to be minimax optimal in that model [71]. Its generalization to trend-filtering, where higher order derivatives are assumed to belong to BV, is a current research topic [38, 78]. However, this discretized model is substantially different from the continuous model that we consider. In fact, the works just mentioned deal with a finite dimensional parameter space of discretized signals and regularize with the ℓ^1-norm of the discrete gradient, which in the limit of finer discretization converges to the Sobolev $W^{1,1}$ seminorm. Hence, BV functions are indistinguishable from Sobolev $W^{1,1}$ functions in the discretized model for any dimension $d \in \mathbb{N}$. However, the difference between $W^{1,1}$ and BV functions is significant: while the gradients of the former are finite Lebesgue continuous measures, the gradients of the latter can be any finite Radon measure, that is, Lebesgue singular measures are allowed. Consequently, BV functions can have jump singularities, which makes their estimation significantly more challenging than estimating a Sobolev function. Therefore, in contrast to the analysis of discrete TV-regularization, the continuous setting is more subtle and genuinely analytical tools are needed, such as the interpolation inequality (9). Moreover, a limitation of discretized models is that they typically discretize the functions and the TV functional with respect to the *same* grid. The discretization of the signals is usually determined by the application, while different discretizations of the TV functional can have different effects (see, e.g., [15]). It is hence useful to study the estimation of BV functions in the continuous setting, since it gives insight into the problem, independently of the discretization of signals or functionals.

Regarding the tools and techniques we use, we mention in particular the concept of an interpolation inequality that relates the risk functional, the regularization functional and the data-fidelity term (see [64] and [37]). While the inequality in those papers is essentially the Gagliardo–Nirenberg inequality for Sobolev norms (see Lecture II in [66]), we extend and make use of interpolation inequalities for the BV norm, for example, equation (9), see [14] and [50]. Finally, as opposed to [37], we formulate our results in the white noise model. This eases the incorporation of results from harmonic analysis (e.g., the interpolation inequalities between BV and $B^{-d/2}_{\infty,\infty}$ and the characterization of Besov spaces by local means) into our statistical analysis, as discretization effects (due to sampling) do not occur. See, however, Section 7 for a discussion of our results in the latter case, where we show that the discretization leads to a qualitative difference between estimation in dimensions $d = 1, 2$ than in $d \geq 3$.

Organization of the paper. In Section 2, we state general assumptions on the family Φ under which the estimator $\hat{f}_\Phi$ is shown to be nearly minimax optimal over the set BV_L. We give a complete statement of the main theorem. Then we present examples of the estimator (4) where Φ is a wavelet basis, a multiresolution system, and a curvelet or shearlet frame combined with wavelets. In Sections 3 and 4, we present the extension of our results to the nonperiodic setting and to the discrete nonparametric regression model, respectively. In Section 5, we discuss the efficient numerical implementation of the frame-constrained TV estimator. The proof of the main theorem is given in Section 6, while some analytical results and proofs are relegated to the Supplementary Material [19]. In Section 7, we briefly discuss possible extensions.

Notation. We denote the Euclidean norm of a vector $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ by $|v| := (v_1^2 + \cdots + v_d^2)^{1/2}$. For a real number x, define $\lfloor x \rfloor := \max\{m \in \mathbb{Z} \mid m \leq x\}$ and $\lceil x \rceil := \min\{m \in \mathbb{Z} \mid m > x\}$. 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 $c_1\|v\|_\alpha \leq \|v\|_\beta \leq c_2\|v\|_\alpha$ for all $v \in V$. Finally, we denote by C a generic positive constant that may change from line to line.

2. Results.

2.1. *Basic definitions.* For $k \in \mathbb{N}$, let C^k denote the space of k-times continuously differentiable periodic functions on $[0,1)^d$, which we identify with the d-torus $\mathbb{T}^d$. The space of 1-periodic functions of bounded variation BV consists of functions $g \in L^1$ whose weak distributional gradient $\nabla g = (\partial_{x_1} g, \dots, \partial_{x_d} g)$ is a periodic, $\mathbb{R}^d$-valued finite Radon measure on $[0,1)^d$ [29]. The finiteness implies that the bounded variation seminorm of g, defined by

$$|g|_{BV} := \sup\left\{ \int_{\mathbb{T}^d} g(x) \operatorname{div}\big(h(x)\big)\, dx \;\middle|\; h \in C^1\big(\mathbb{T}^d; \mathbb{R}^d\big), \big\||h|\big\|_{L^\infty} \le 1 \right\}, \tag{11}$$

is finite, where $\operatorname{div}(h)$ is the divergence of the vector field $h = (h_1, \dots, h_d)$, and $\||h|\|_{L^\infty} := \sup_{x \in \mathbb{T}^d} (\sum_{i=1}^d |h_i(x)|^2)^{1/2}$ denotes the supremum of its magnitude. BV is a Banach space with the norm $\|g\|_{BV} = \|g\|_{L^1} + |g|_{BV}$, see [29]. For $S \in \mathbb{N}$, let $\Phi = \{\psi_{j,k,e} \mid (j,k,e) \in \Omega\}$ be an S-regular wavelet basis for L^2 whose elements are S times continuously differentiable with absolutely integrable Sth derivative, indexed by the set

$$\begin{aligned} \Omega &:= \big\{(j,k,e) \mid j \ge 0, k \in P_j^d, e \in E_j\big\}, \quad \text{with} \\ P_j^d &:= \big\{k = (k_1, \dots, k_d) \mid k_i = 0, \dots, 2^j - 1, i = 1, \dots, d\big\}, \\ E_j &:= \begin{cases} \{0,1\}^d & \text{if } j = 0, \\ \{0,1\}^d \backslash (0, \dots, 0) & \text{else.} \end{cases} \end{aligned} \tag{12}$$

In particular, we consider wavelets of the form

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

where $\psi_e(z_1, \dots, z_d) = \prod_{i=1}^d \psi_{e_i}(z_i)$ is a tensor product of periodized 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 or the father wavelet of a one-dimensional wavelet basis of L^2. The index $(0, \dots, 0) \in E_0$ refers here to (shifts of) the father wavelet $\psi_{0,k,0} = \varphi(\cdot - k)$. See, for example, Section 4.3.6 in [34] for the construction of such a basis. Then for $p, q \in [1, \infty]$ and $s \in \mathbb{R}$ with $S > |s|$, the Besov norm of a (generalized) function is defined by

$$\|g\|_{B^s_{p,q}} := \left(\sum_{j \in \mathbb{N}_0} 2^{jq(s + d(\frac{1}{2} - \frac{1}{p}))} \left(\sum_{k \in P_j^d} \sum_{e \in E_j} |\langle \psi_{j,k,e}, g \rangle|^p \right)^{q/p} \right)^{1/q}, \tag{13}$$

with the usual modifications if $p = \infty$ or $q = \infty$. If $s > 0$ and $p \in [1, \infty)$, the Besov space $B^s_{p,q}$ consists of L^p functions with finite Besov norm, while if $s > 0$ and $p = \infty$, then $B^s_{p,q}$ consists of continuous functions with finite Besov norm. In these cases, $\langle \cdot, \cdot \rangle$ denotes the standard inner product in L^2. If $s \le 0$, $B^s_{p,q}$ consists of periodic distributions $\mathcal{D}^*(\mathbb{T}^d)$ with finite Besov norm. Here, $\mathcal{D}^*(\mathbb{T}^d)$ denotes the space of periodic distributions, defined as the topological dual to the space of infinitely differentiable periodic functions $C^\infty(\mathbb{T}^d)$ (see Section 4.1.1 in [34]). In that case, $\langle \psi_{j,k,e}, g \rangle$ is interpreted as the action of $g \in \mathcal{D}^*(\mathbb{T}^d)$ on the function $\psi_{j,k,e}$.

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

$$\mathcal{F}[g](\xi) := \int_{\mathbb{T}^d} g(x) e^{-2\pi i \xi x}\, dx, \quad \xi \in \mathbb{Z}^d. \tag{14}$$

The Fourier transform of a function $g \in L^1(\mathbb{R}^d)$ is defined as in (14) extending the integration over $\mathbb{R}^d$. The formal definition of the Fourier transform is as usual extended to functions in L^2 and, by duality, to distributions $\mathcal{D}^*(\mathbb{T}^d)$ (see, e.g., Section 4.1.1 in [34]).

2.2. *Main result.* The main ingredient of the estimator (4) is the dictionary Φ, on which we impose the following assumptions.

ASSUMPTION 1. Φ is of the form $\Phi = \{\phi_\omega \mid \omega \in \Omega\} \subset L^2$ 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 $cn^\Gamma \leq \#\Omega_n \leq Q(n)$ for all n for a polynomial Q and constants $c, \Gamma > 0$. The sets Ω_n are assumed to satisfy the inequality (6) for any $g \in L^\infty$.

EXAMPLES. (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 wavelets (see Proposition 1 below).

(b) Another family Φ satisfying Assumption 1 is given by translations and rescalings of (the smooth approximation to) the indicator function of a cube. In Section 2.3.2, we verify the assumption for such a system, that has been used previously as a dictionary for function estimation (see [37]).

(c) In Section 2.3.3, 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), and let Y_ω be as in (3) the projections of the white noise model onto a dictionary Φ satisfying Assumption 1. We denote the estimator in (4) as *frame-constrained TV-estimator* with respect to the dictionary Φ, where we minimize over the set

$$X_n := \{g \in BV \cap L^\infty \mid \|g\|_{L^\infty} \leq \log n\}. \tag{15}$$

We use the convention in (4) that, whenever the arg min is taken over the empty set, $\hat{f}_\Phi$ is the constant zero function.

In the following, we assume that $n \geq 2$ in order to avoid $\log 1 = 0$. The reason for the additional constraint $\|g\|_{L^\infty} \leq \log n$ is technical: We will need upper bounds on the supremum norm of $\hat{f}_\Phi$. As it turns out, the upper bound $\log n$ will not affect the minimax polynomial rate of convergence of the estimator (but it yields additional logarithmic factors). Alternatively, if we knew an upper bound L for the supremum norm of f, we could choose $X_n = \{g \in BV \cap L^\infty \mid \|g\|_{L^\infty} \leq L\}$. In that case, the risk bounds in Theorem 1 would improve in some logarithmic factors (see Remark 2).

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

(a) *Let γ_n be as in (5) 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 (4) with parameter γ_n satisfies*

$$\|\hat{f}_\Phi - f\|_{L^q} \leq Cn^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}} (\log n)^{3-\min\{d,2\}} \tag{16}$$

with probability at least $1 - (\#\Omega_n)^{1-\kappa^2}$ uniformly over $f \in BV_L$.

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

$$\sup_{f \in BV_L} \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q}\big] \leq Cn^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}} (\log n)^{3-\min\{d,2\}} \tag{17}$$

holds for n large enough and a constant $C > 0$ independent of n.

REMARK 1. (a) Notice that part (a) of the theorem implies that (16) holds asymptotically almost surely if $\kappa^2 > 2$.

(b) By the assumption that $\|\phi_\omega\|_{L^2} = 1\ \forall \omega \in \Omega$, we have the tail bound

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

for any $n \in \mathbb{N}$ and $t \geq 0$, where dW denotes the white noise process in $L^2(\mathbb{T}^d)$. This bound follows from Chernoff's inequality and the union bound, and it will play an important role for bounding the stochastic estimation error of the estimator $\hat{f}_\Phi$.

(c) The constants in the right-hand side of the theorem depend on the noise level σ like $C_\sigma \asymp \max\{\sigma^2, 1\}^{\min\{\frac{1}{d+2}, \frac{1}{dq}\}}$. This matches the lower bound in Theorem 2 for $\sigma \geq 1$.

REMARK 2. The logarithmic factors in (16) and (17) 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 X_n in (15)). 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 Proposition 5 in Section 6 for an explanation of the different factors in $d = 1$ and $d \geq 2$.

Regarding the optimality of the log factors, we distinguish two cases:

- Case $q < 1 + 2/d$. It is known that the minimax rate for BV functions is exactly of order $n^{-1/(d+2)}$, see [52] (where the same rate is shown for a larger Besov class $B^1_{1,\infty}$). Here we do lose a logarithmic factor, and it is unclear whether this is due to our analysis or the method itself.
- Case $q \geq 1 + 2/d$. It is not known whether our lower bounds in Theorem 2 are optimal or not. We conjecture that the minimax optimal rate should have some logarithmic factors, as it is seen in the discrete setting [71], Theorem 2, when $q = 2$ and $d \geq 2$. In this case, the logarithmic factors in our results might be potentially optimal.

REMARK 3. Recall that our parameter set BV_L 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 Ω_n index the wavelet coefficients up to level $J = \lfloor \frac{1}{d}\frac{\log n}{\log 2} \rfloor$. In the proof of Theorem 1 we need a relaxed form of Assumption 1, namely an inequality of the form

$$\max_{(j,k,e)\in\Omega}\left|\langle\psi_{j,k,e}, g\rangle\right| \leq \max_{(j,k,e)\in\Omega_n}\left|\langle\psi_{j,k,e}, g\rangle\right| + C2^{-Jd/2} \quad \forall J \in \mathbb{N} \tag{18}$$

for sufficiently smooth g. But this inequality for all $J \in \mathbb{N}$ is equivalent to $\|g\|_{B^0_{\infty,\infty}(\mathbb{T}^d)} \leq C$ (see Berstein-type inequalities for Besov spaces, e.g., in Section 3.4 in [13]). 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 $\widetilde{BV}_L := \{g \in BV \mid |g|_{BV} \leq L, \|g\|_{B^0_{\infty,\infty}} \leq L\}$.

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 (see (13)), 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 (18) 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 functions, or Sobolev $W^{k,p}$ functions with $k > d/p$, so the assumption $f \in L^\infty$ is implicit. Alternatively, we refer to Section 3 in [53] and to [20] 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 the main result of this paper, which is a direct consequence of Theorem 1.

THEOREM 2. *Under the assumptions of Theorem 1, the estimator $\hat{f}_\Phi$ is minimax optimal up to logarithmic factors over the parameter set BV_L defined in (7) with respect to the L^q-risk for $q \in [1, \infty)$ in any dimension $d \in \mathbb{N}$, that is,*

$$\inf_{\hat{f}} \sup_{f \in BV_L} \mathbb{E}\big[\|\hat{f} - f\|_{L^q}\big] \geq C\big(\sigma^2/n\big)^{\min\{\frac{1}{d+2}, \frac{1}{dq}\}}$$

for any $q \in [1, \infty)$, where the infimum runs over all measurable functions from the sample space of dY in (1) to the reals.

The proof of Theorem 2 is given in Section 6.2. It consists of proving a lower bound for the minimax risk over BV_L, which we show agrees with the upper bound proven in Theorem 1.

2.3. *Examples.*

2.3.1. *Wavelet-based estimator.* For $S \in \mathbb{N}$, let $\Phi = \{\psi_{j,k,e} \mid (j,k,e) \in \Omega\}$ be an S-regular wavelet basis of $L^2(\mathbb{T}^d)$ as described in Section 2.1. For $n \in \mathbb{N}$, $n \geq 2^d$, define the subset

$$\Omega_n := \big\{(j,k,e) \in \Omega \mid j = 0, \dots, J-1\big\}, \tag{19}$$

with $J = \lfloor \frac{1}{d} \frac{\log n}{\log 2} \rfloor$. Note that $2^{-d} n \leq \#\Omega_n = 2^{Jd} \leq n$ for any $n \geq 2^d$.

PROPOSITION 1. *An S-regular wavelet basis of L^2 as in Section 2.1 with $S > \max\{1, d/2\}$ satisfies Assumption 1 with the sets Ω_n in (19), a linear polynomial $Q(x) = x$ and parameter $\Gamma = 1$.*

For the proof, see Section 2.1 in the Supplementary Material [19]. A direct consequence of this proposition and of Theorem 1 is that the frame-constrained TV-estimator with the wavelet basis above is nearly minimax optimal over BV_L.

REMARK 4. In dimension $d = 1$, [24] 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 a control on the BV-seminorm: this second aspect is crucial in higher dimensions. Indeed, in the proof of Theorem 1 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 (4) 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 perform optimally over Besov spaces $B^s_{p,t}$ for $s > d(1/p - 1/2)$ only (see, e.g., [20]). 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 satisfies 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.

2.3.2. *m-Adic multiscale systems*. We construct the multiscale TV-estimator by choosing Φ to be a family of smooth functions supported in cubes of different sizes at different locations. Assumption 2 makes this precise. For notational simplicity, we sometimes index the set functions in Φ by the cube $B \subset [0,1)^d$ in which they are supported, and the set of all cubes considered is denoted by Ω.

ASSUMPTION 2. The system of functions $\Phi = \{\phi_B \mid B \in \Omega\}$ satisfies the following conditions:

(a) for fixed $m \in \mathbb{N}$, $m \geq 2$, the set Ω consists of the intersections with $[0,1)^d$ of all m-adic cubes at m-adic positions contained in $[0,2)^d$. For each $n \in \mathbb{N}$ with $n \geq m^d$, define $J = \lceil \frac{1}{d} \frac{\log n}{\log m} \rceil$, $R = J \max\{1, \frac{d}{2}\}$ and

$$\mathcal{D}_R := \{\overline{k} = (k_1 m^{-R}, \ldots, k_d m^{-R}) \mid k_i = 0, \ldots, m^R - 1, i = 1, \ldots, d\},$$

$$\Omega_n := \{(\overline{k} + [0, m^{-j})^d) \cap [0,1)^d \mid j = 0, \ldots, J-1, \overline{k} \in \mathcal{D}_R\};$$

(b) there is a function $K \in C^\infty(\mathbb{R}^d)$ with supp $K \subseteq [0,1)^d$, $|\mathcal{F}[K](\xi)| > 0$ in $|\xi| < 2$ and $\|K\|_{L^2(\mathbb{R}^d)} = 1$, $\|K\|_{L^\infty(\mathbb{R}^d)} \leq 2$ such that all functions $\phi_B \in \Phi$ are given by translation, dilation and rescaling of K. More precisely, for each cube $B \in \Omega$ of the form $B = \overline{k}_B + [0, |B|^{1/d})^d$, the function $\phi_B \in \Phi$ is given by

$$\phi_B(z) = |B|^{-1/2} K\big(|B|^{-1/d}(z - \overline{k}_B)\big).$$

REMARK 5. (a) An example of a function K 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 $|\mathcal{F}[1_{[a,b]}](\xi)| > 0$ for $|\xi \cdot (b-a)| < 1$. Taking K to be a smooth approximation to an indicator function, the estimator (4) is reminiscent of that proposed by [30].

(b) For given $m \geq 2$ and $n \in \mathbb{N}$ with $n \geq m^d$, $\#\Omega_n = J m^{dR} = J m^{dJ \max\{1, d/2\}}$, whence

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

PROPOSITION 2. *Let $\Phi = \{\phi_B \mid B \in \Omega\}$ satisfy Assumption 2. Then it satisfies Assumption 1 with polynomial $Q(x) = x^{\max\{1,d/2\}+1}$ and $\Gamma = \max\{1, d/2\}$.*

See Section 2.2 of the Supplementary Material [19] for the proof of Proposition 2. We remark that part of the proof of Proposition 2 is based on a characterizations of Besov spaces via local means [75]. Again this proposition together with Theorem 1 proves near minimax optimality for the multiscale TV-estimator.

2.3.3. *Shearlet and curvelet estimators*. Another relevant example of the estimator in (4) in $d \geq 2$ corresponds to the case when Φ contains a frame of shearlets or curvelets. While classical curvelets are defined for $d = 2$ (see, e.g., [5]), there are several extensions to higher dimensions. In order to simplify and unify the analysis, in this paper we will work with the construction of shearlets in Section 3 of [47], and the curvelet frame from Section 7 of [2]. The reason for working with these constructions is that they are defined in all dimensions by a partition of frequency space, thus simplifying the notation. We nevertheless remark that the analysis presented here can be easily adapted to other curvelet and shearlet constructions.

Let $\{\overline{\varphi}_{j,\tilde{\theta}} \mid (j,\tilde{\theta}) \in \Xi\}$ denote either the tight shearlet frame or the tight curvelet frame mentioned above. Then $\{\varphi_{j,\tilde{\theta}} \mid (j,\tilde{\theta}) \in \Theta\}$ consists of the normalized periodizations of the

elements $\overline{\varphi}_{j,\tilde{\theta}}$ that have a nonzero overlap with the indicator function of the unit cube, that is, $\int_{[0,1]^d} \overline{\varphi}_{j,\tilde{\theta}}(z)\,dz \neq 0$. For simplicity of the notation, we index the elements by $(j,\tilde{\theta}) \in \Theta \subset \mathbb{N}_0 \times \tilde{\Theta}$, where $j \geq 0$ plays the role of a scale index, and $\tilde{\theta}$ indexes the position and orientation of the frame elements (see the references above for the precise construction in each case). In the rest of this section, we will consider frames of $L^2(\mathbb{T}^d)$ that contain the set $\{\varphi_{j,\tilde{\theta}} \mid (j,\tilde{\theta}) \in \Theta\}$.

ASSUMPTION 3. Let $\{\psi_{j,k,e} \mid (j,k,e) \in \Theta^W\}$ denote an S-regular wavelet basis of $L^2(\mathbb{T}^d)$ with $S > \max\{1, d/2\}$, and let $\{\varphi_{j,\tilde{\theta}} \mid (j,\tilde{\theta}) \in \Theta\}$ denote the set of functions constructed above. Then define $\Phi := \{\psi_{j,k,e} \mid (j,k,e) \in \Theta^W\} \cup \{\varphi_{j,\tilde{\theta}} \mid (j,\tilde{\theta}) \in \Theta\}$. Further, for $n \in \mathbb{N}$ define $J = \lfloor \frac{1}{d} \frac{\log n}{\log 2} \rfloor$ and let $\Phi_n := \{\psi_{j,k,e} \mid (j,k,e) \in \Theta_n^W\} \cup \{\varphi_{j,\tilde{\theta}} \mid (j,\tilde{\theta}) \in \Theta_n\}$, where

$$\Theta_n^W := \big\{(j,k,e) \in \Theta^W \mid j = 0, \dots, J-1\big\},$$
$$\Theta_n := \big\{(j,\tilde{\theta}) \in \Theta \mid j = 0, \dots, \tilde{J}-1\big\},$$

where $\tilde{J} \in \mathbb{N}$ is the largest possible natural number such that $2^{d(J-1)} \leq \#\Theta_n \leq 2^{dJ}$. For consistency with the notation in the previous sections, we define the joint index set $\Omega_n := \Theta_n^W \cup \Theta_n$.

REMARK 6. (a) The assumption that Φ contains a wavelet basis as well as 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 on a curvelet frame only, the embeddings in Lemma 9 in [2] together with classical embeddings of Besov spaces (see Remark 4 of Section 3.5.4 in [73]) would give the bound

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

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

$$\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}(\mathbb{T}^d)} \leq C \frac{n^{\delta'}}{\sqrt{n}} \operatorname{Polylog}_{d,\delta'}(n)$$

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 space associated with the curvelet frame. The situation for the shearlet frame is analogous, as its associated decomposition space equals that of the curvelet frame (see Proposition 4.4 in [47]).

(b) We make the assumption that $\#\Theta_n \leq 2^{dJ}$ for any $n \in \mathbb{N}$ and $J = \lfloor \frac{1}{d} \frac{\log n}{\log 2} \rfloor$ in order to simplify subsequent computations. The assumption is justified, since the cardinality of Θ_n behaves indeed like $O(2^{dJ})$. In fact, the number of curvelet (or shearlet) frame elements at scale 2^{-j} that have a nonzero overlap with the unit cube behaves as 2^{dj}, since there are $O(2^{j+\frac{d-1}{2}j})$ positions and $O(2^{\frac{d-1}{2}j})$ orientations. We refer to Section 8.2 in [6] and [2] for the details. The claim for the shearlet frame follows from that of the curvelet frame by the comparison in Section 4.4 in [47].

The constructions of tight curvelet frames in [2] and of shearlet frames in [47] yield smooth frame elements that are exponentially decaying in space. We use this to show that the family Φ satisfies Assumption 1.

PROPOSITION 3. *Let Φ satisfy Assumption 3 with either the shearlet or the curvelet frame. Then it satisfies Assumption 1 with $Q(x) = 2x$ and $\Gamma = 1$.*

The proof of Proposition 3 is given in Section 2.3 of the Supplementary Material [19]. As a consequence, we conclude from Theorem 1 that the curvelet TV-estimator is nearly minimax optimal for estimating BV_L functions.

We close this section presenting some dictionaries Φ that do not satisfy Assumption 1, where hence Theorem 1 does not apply.

(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 with $S > \max\{1, d/2\}$ (see Section 4.3 in [34]).

(b) For the multiscale TV-estimator in Section 2.3.2 we considered a dictionary Φ consisting on *smoothed* indicator functions of cubes in $[0,1]^d$. The smoothing 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 2.2 of the Supplementary Material [19]).

(c) As argued in part (a) of 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 [2]) do not match Besov spaces exactly, whence Assumption 1 does not hold.

3. Extension to nonperiodic functions. In this section, we show how our approach generalizes to the estimation of nonperiodic functions. For simplicity, we consider functions supported on $[0,1]^d$, but our approach applies to functions supported on bounded sets whose boundary has bounded $(d-1)$-Lebesgue measure. Consider observations from the white noise model

$$dY(x) = f(x)\,dx + n^{-1/2}\,dW(x), \quad x \in [0,1]^d, \tag{20}$$

where

$$f \in BV_L\big([0,1]^d\big) := \big\{ f \in BV\big([0,1]^d\big) \mid |f|_{BV([0,1]^d)} \le L, \|f\|_{L^\infty} \le L \big\},$$

and the $BV([0,1]^d)$ seminorm is defined as in (11) with the difference that h runs over $C^1_c([0,1]^d; \mathbb{R}^d)$.

In a nutshell, our approach to estimate f in (20) is to embed that model into the periodic model (1), apply the periodic estimator there, and then restrict back to $[0,1]^d$. In the following we explain this approach.

3.1. *Embedding into the periodic setting.* Given the data (20), we define new observations by "padding" dY with periodic white noise, that is,

$$d\widetilde{Y}(x) = \begin{cases} dY(x) & \text{if } x \in [0,1]^d, \\ d\widetilde{W}(x) & \text{if } x \in [-\varepsilon, 1+\varepsilon]^d \backslash [0,1]^d \end{cases} \tag{21}$$

for a fixed $\varepsilon > 0$, where $d\widetilde{W}$ is a Gaussian white noise process over the torus $\mathbb{T}^d$, which we identify with $[-\varepsilon, 1+\varepsilon]^d$ with periodic boundary conditions.

Defining the function $f_{\text{ext}} = f 1_{[-\varepsilon,1+\varepsilon]^d}$ and its periodic extension f_{per}, we easily see that

$$\begin{aligned} \|f_{\text{per}}\|_{L^\infty(\mathbb{T}^d)} &= \|f\|_{L^\infty([0,1]^d)} \le L, \\ |f_{\text{per}}|_{BV(\mathbb{T}^d)} &= |f 1_{[-\varepsilon,1+\varepsilon]^d}|_{BV(\mathbb{R}^d)} \le |f|_{BV([0,1]^d)} + |\partial [0,1]^d| \|f\|_{L^\infty([0,1]^d)} \\ &\le (1 + 2^d) L. \end{aligned}$$

These last inequalities follow from the fact that, by extending f by zero outside of $[0,1]^d$, we are potentially introducing jumps at the boundary of $[0,1]^d$. Therefore, the total variation of f over $\mathbb{R}^d$ is bounded by its total variation inside the hypercube plus the worst case variation at the boundary, which is bounded by the maximum jump size (i.e., $\|f\|_{L^\infty}$) times the $d-1$-Lebesgue measure of the boundary, which equals 2^d for $[0,1]^d$. This implies that $f_{\mathrm{per}} \in BV_L(\mathbb{T}^d)$ as defined in (7) (up to relabeling of the constant L).

3.2. *Nonperiodic estimator.* Given observations $d\widetilde{Y}$ as in (21), which follow model (1) with drift f_{per}, we compute the estimator $\hat{f}_\Phi$ as in (4), which for suitable frame and threshold γ_n attains the optimal convergence rate. In order to estimate the original function f, we remove the padding of $\hat{f}_\Phi$, that is,

$$\hat{f}(x) := \hat{f}_\Phi(x) 1_{[0,1]^d}(x), \quad x \in [0,1]^d. \tag{22}$$

THEOREM 3. *Let $d \in \mathbb{N}$, $q \in [1,\infty)$ and assume model (20). Let Φ satisfy Assumption 1, and γ_n be as in (5) with $\kappa^2 > 1 + \frac{1}{(d+2)\Gamma}$. Let the estimator $\hat{f}_\Phi$ be as in (4) with parameter γ_n, and define $\hat{f}$ as in (22). We have that*

$$\sup_{f \in BV_L([0,1]^d)} \mathbb{E}\big[\|\hat{f} - f\|_{L^q([0,1]^d)}\big] \le C n^{-\min\{\frac{1}{d+2}, \frac{1}{dq}\}} (\log n)^{3-\min\{d,2\}}$$

holds for n large enough and a constant $C > 0$ independent of n.

The proof of Theorem 3 is given in Section 3 of the Supplementary Material [19].

4. Extension to discretized model. In this section, we show how the frame-constrained TV estimator can be extended to the discrete nonparametric regression model, and prove a convergence result for it. Consider observations from the model

$$Y_i = f(x_i) + \sigma \varepsilon_i, \quad x_i \in \Gamma_n, i = 1, \dots, n, \tag{23}$$

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

$$\Gamma_n := \left\{ \left(\frac{k_1}{m}, \dots, \frac{k_d}{m} \right) \,\middle|\, k_i \in \{1, \dots, m\}, i = 1, \dots, d \right\} \tag{24}$$

is the observation grid. Of course, different grids may be used. In (23), ε_i are independent standard normal random variables, and $\sigma > 0$ plays the role of the standard deviation of the noise. Additionally, we have to assume that f is well defined on the grid Γ_n and that it satisfies a form of continuity there:

$$\begin{aligned} f(x) &= \lim_{r \to 0} \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y)\, dy, \\ Df(x) &= \lim_{r \to 0} \frac{1}{|B(x,r)|} \int_{B(x,r)} Df(y)\, dy \quad \text{and} \end{aligned} \tag{25}$$

$$\lim_{r \to 0} \frac{|[Df]_s|(B(x,r))}{r^d} = 0 \quad \text{for any } x \in \Gamma_n, \tag{26}$$

where Df is the Lebesgue continuous part of the gradient of f, and $[Df]_s$ is its singular part, and $B(x,r)$ is a ball of radius r around x. We remark that these equalities hold for almost any x, since f is of bounded variation. Define now the parameter space

$$BV_{L,n}(\mathbb{T}^d) = \{g \in BV_L \mid g \text{ satisfies (25) and (26)}\},$$

where BV_L is defined in (7). In particular, we allow functions in $BV_{L,n}$ to have discontinuities, provided they do not occur at the sampling points.

In order to apply our multiscale methodology here, we need a discrete frame on which to project the data. For that, let $\Phi_n = \{\phi_\omega^n \mid \omega \in \Omega_n\}$ be a dictionary of discretized elements, that is, each ϕ_ω^n is a vector of n values

$$(\phi_\omega^n)_i = n^{-1/2}\phi_\omega(x_i) \quad \text{for } i = 1, \dots, n, \tag{27}$$

which are the evaluations of ϕ_ω at the grid points. The scaling factor $n^{-1/2}$ is chosen so that $\|\phi_\omega^n\|_{\ell^2} \to \|\phi_\omega\|_{L^2} = 1$ as $n \to \infty$. Define now

$$\hat{f}_D \in \operatorname*{arg\,min}_{g \in X_n} |g|_{BV} \quad \text{subject to } \max_{\omega \in \Omega_n} \left| \sum_{x_i \in \Gamma_n} (\phi_\omega^n)_i \big(g(x_i) - Y_i\big) \right| \le \kappa\sigma\sqrt{2\log \#\Omega_n}. \tag{28}$$

THEOREM 4. *Let $d \in \mathbb{N}$ and $q \in [1, \infty)$. Let Φ satisfy Assumption 1, and γ_n be as in (5) with $\kappa^2 > 1 + \frac{1}{(d+2)\Gamma}$. Let the estimator $\hat{f}_D$ be as in (28) based on observations (23). We have that*

$$\sup_{f \in BV_{L,n}(\mathbb{T}^d)} \mathbb{E}\big[\|\hat{f}_D - f\|_{L^q}\big] \le C\big(\max\{n^{-\frac{1}{2}}, n^{-\frac{1}{d}}\}\big)^{\min\{\frac{2}{d+2}, \frac{2}{dq}\}} (\log n)^{3-\min\{d,2\}} \tag{29}$$

holds for n large enough and a constant $C > 0$ independent of n.

The proof of Theorem 4 is given in Section 4 of the Supplementary Material [19]. Notice that the convergence rate consists of two terms: the term $n^{-1/2}$ is the same as in the white noise model, and arises from the stochastic fluctuations of the observations. The term $n^{-1/d}$ is however not present in the white noise model, and it is the discretization error in which we incur by observing discretized data only. Some remarks are due:

(a) The discretization error is unavoidable, and it is so large because of the roughness of BV functions. Indeed, approximating the frame coefficients of a BV functions by its discrete coefficients is subject to an error of order $O(n^{-1/d})$. For comparison, that discretization error for a β-Hölder function is of order $O(n^{-\beta/d})$: this is consistent with the fact that, for $\beta > d/2$, the white noise model and nonparametric regression models are Le Cam equivalent for β-Hölder functions [69].

(b) The rate in Theorem 4 equals that in the white noise model for $d = 1, 2$, while the discretization error $n^{-1/d}$ dominates for $d \ge 3$.

(c) As discussed in the Introduction, the only other works dealing with the minimax estimation of functions of bounded variation in higher dimensions are [44] and [71]. A crucial difference with our setting is that they consider a discretized model and measure convergence with a *discrete* ℓ^2 risk. This amounts effectively to ignoring discretization effects, which is why they do not observe the discretization error $O(n^{-1/d})$. It is however well known that discretization effects are often present in practice (e.g., in nanoscopy), and the error bound in (29) faithfully represents that.

5. Computational aspects. In this section, we discuss how the multiscale TV-estimator can be computed efficiently. For that, we encounter two challenges: one is the discretization of the BV seminorm, the other is the numerical solution of the optimization problem in (4).

5.1. *Discretization of the problem.* We discretize (4) as follows:

(i) We do not minimize over functions $g \in X_n$ defined on the continuous domain $[0,1]^d$, but over vectors of function values on a particular grid. Which particular grid we take (e.g., Cartesian, polar, etc.) is not crucial (see however point (iii) below). In fact, in applications, the discretization of the signals is determined by the measurement process. In the following, we denote the discretization of a function g on a grid $\Gamma_n = \{x_i, i = 1, \dots, n\}$ by $g_n := \{g(x_i), x_i \in \Gamma_n\}$. We denote by V the space of all such vectors of function values, which we identify with $\mathbb{R}^n$.

(ii) We discretize the frame $\{\phi_\omega\}$ onto which we project g_n. We do so by associating to each function ϕ_ω a vector of function values ϕ_ω^n as defined in (27). We also discretize the inner product in (4) in order to have

$$\langle g_n, \phi_\omega^n \rangle_{\ell^2} = n^{-1} \sum_{x_i \in \Gamma_n} g(x_i)\phi_\omega(x_i).$$

(iii) Finally, we discretize the BV seminorm in (4). While there are many ways of discretizing it (see, e.g., [8] and [12]), we choose the discretization in Cartesian coordinates, also known as isotropic discretization. We remark however that our theory also covers the improved discretization proposed in [15] (see Remark 7 for a comparison of the different discretizations). Finally, we stress that the discretization of the BV seminorm can be unstable in some situations, but in our case we can guarantee stability (see Remark 8 below).

REMARK 7 (Different discretizations of the BV seminorm). Here we discuss different possible discretizations of the BV seminorm and its practical effects on the final reconstruction. First of all, we distinguish between the *anisotropic* and the *isotropic* discretization, which are given by

$$BV_{\text{anis}}(g_n) = \sum_{x_i \in \Gamma_n} \sum_{i=1}^{d} |\Delta_i g(x)| \quad \text{and} \quad BV_{\text{iso}}(g_n) = \sum_{x_i \in \Gamma_n} \sqrt{\sum_{i=1}^{d} |\Delta_i g(x)|^2},$$

where $\Delta_i g(x)$ is a discrete approximation to the partial derivative of g along the ith direction by finite differences. Notice that our definition of the BV seminorm in (11) corresponds to the isotropic discretization (since we use the Euclidean norm of the vector field $h(x)$ in (11)). However, we mention the anisotropic discretization here for two reasons. The first one is that, in the mathematical statistics literature, the only convergence rates for TV-regularized least squares in dimension $d \geq 2$ have been proven for the anisotropic discretization (see [44] and [71]). The second reason is that, in the numerical analysis literature, the isotropic discretization is known to be superior to the anisotropic one [15]. The reason for that is that the anisotropic discretization detects variations along the Cartesian axes well, but it performs poorly for variations along different directions, for instance along curves. We illustrate this difference in Figure 3.

On the other hand, the isotropic discretization has difficulties in detecting variations along the diagonal directions, and refined discretizations of the BV seminorm have been proposed in order to solve that issue. A remarkable example of such a refinement was proposed by Condat [15], and it consists of discretizing the vector fields $h(x)$ in the definition (11) on a grid *finer* than Γ_n, which allows the discretized TV functional to distinguish finer directional information (we refer to [15] for the details). In Figure 3, we show a comparison of the anisotropic, the isotropic BV_{iso} and Condat's isotropic discretizations.

Finally, we stress once again the value of deriving our theory in the continuous setting, since then we can discretize the BV functional as we want (e.g., with Condat's discretization) in order to achieve better results.

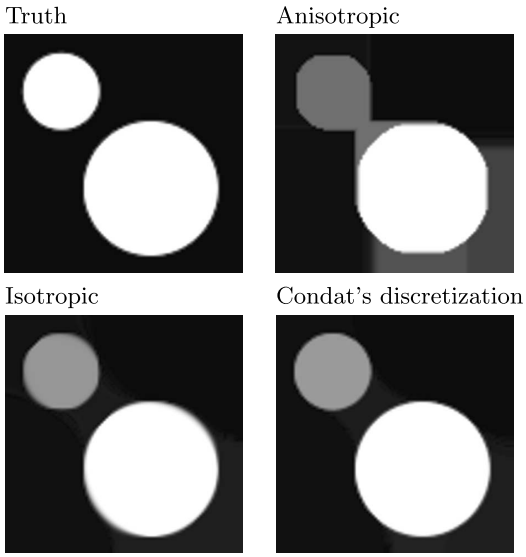

FIG. 3. *Reconstructions by TV-regularization for different discretizations of the TV seminorm.*

REMARK 8 (Stability of the discretization). Since we have discretized the BV seminorm in order solve the problem (4), we have necessarily introduced an error, and we should ask how far we are from the solution of the original problem. The answer is that, as $n \to \infty$, we do not lose anything, as the properly rescaled discretized BV functional Γ-converges to the BV-seminorm [8]. Indeed, as explained in Section 5.2 below, we will iteratively use Chambolle's algorithm in the discretized model, which was shown to produce reconstructions that converge to the minimizer of the continuous model in the limit $n \to \infty$ [8].

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 the setting of Bayesian inverse problems, [48] and [49] proved that imposing a *discretized BV* prior (analogous to regularizing with the BV seminorm) shows the following phenomenon: as the level of discretization grows finer, the posterior mean estimator converges to the posterior mean corresponding to a *Sobolev* H^1 prior (Theorem 5.1 in [49]). Further, [48] show that Besov $B^1_{1,1}$ priors do not show this effect. This is a major computational difference between the BV and the Besov $B^1_{1,1}$ or Sobolev seminorms: the former is not discretization invariant, while the latter are.

5.2. *Optimization algorithm for* (4). After discretizing the problem (4), we can write it in the form

$$\min_{v \in V} F(Kv) + BV_{\text{dis}}(v), \tag{30}$$

where V is the set of vectors of $\mathbb{R}^n$ described in point (i) in Section 5.1, BV_{dis} is a discretization of the BV seminorm as discussed in point (iii) of Section 5.1, and K and F are defined as follows:

(1) The operator $K : \mathbb{R}^n \to \mathbb{R}^{\#\Omega_n}$ maps an vector g_n to its coefficients with respect to the (discretized) frame ϕ^n_ω, that is,

$$K : g_n \mapsto \{\langle \phi^n_\omega, g_n \rangle, \omega \in \Omega_n\} := \left\{ \frac{1}{n} \sum_{x_i \in \Gamma_n} \phi_\omega(x_i) g(x_i), \omega \in \Omega_n \right\}.$$

(2) The functional F is given by

$$F(z) = 1_{\leq 0}(v - Y - \gamma_n) + 1_{\leq 0}(-v + Y - \gamma_n) \quad \text{for } v \in \mathbb{R}^{\#\Omega_n},$$

where γ_n is the threshold in (4), and $Y = \{Y_\omega, \omega \in \Omega_n\} \in \mathbb{R}^{\#\Omega_n}$ denotes the vector of coefficients of Y with respect to the frame $\{\phi_\omega\}$. The indicator functions in the above display are defined as

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

The first observation is that (30) is a large-scale, convex, nonsmooth optimization problem, which renders standard techniques such as interior point methods inapplicable or unfeasible. Instead, we used the Chambolle–Pock (CP) primal-dual algorithm [10] for nonsmooth optimization. In order to do so, we write (30) as a saddle-point problem

$$\min_{v \in V} \max_{w \in W} \langle Kv, w \rangle - F^\star(w) + G(v), \tag{31}$$

where $F^\star$ is the convex conjugate of F. The CP algorithm solves (30) by writing the optimality conditions for (31) in terms of the proximal operators of G and $F^\star$, and solving these optimality conditions iteratively (we refer to [10] for the details). This means that the computational complexity of the CP algorithm is driven by the complexity of the proximal operators of G and $F^\star$. In our setting, these proximal operators can be computed as follows:

(1) The convex conjugate of F, $F^\star$, is given by

$$F^\star(z) = \sum_{\omega \in \Omega_n} z_\omega Y_\omega + \gamma_n |z_\omega|.$$

The proximal operator of $F^\star$ is given by

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

whose solution is soft-thresholding of $w - \tau Y$ by the threshold $\tau\gamma_n$.

(2) The proximal operator of G is given by the TV-L^2 functional, that is,

$$\arg\min_{g_n \in \mathbb{R}^{\Gamma_n}} \frac{\|g_n - v\|^2}{2\tau} + G(g_n).$$

This minimization problem can be solved efficiently by Chambolle's algorithm [8], which carefully employs the characterization of the BV seminorm as a supremum over differentiable vector fields.

In this setting, we can apply the CP algorithm to compute the solution to (30) efficiently. As an example of the runtime, in our implementation in Matlab the computation on a $n = 512$ one-dimensional signal takes of the order of 0.3 seconds, and around 3.5 minutes on a $n = 256 \times 256$ two-dimensional image. One comment is due regarding the choice of free parameters in our problem, that is, regarding the choice of the threshold γ_n and of the frame Φ. On one hand, the threshold should be chosen as in Theorem 1 in order to guarantee optimality. On the other and, regarding the frame Φ, in our simulations we have considered several choices. We have worked with the discretization of frames given by orthonormal bases of Daubechies wavelets, with mixed frames of Daubechies wavelets and curvelets (as in Section 2.3.3), and with redundant sets of indicator functions at dyadic positions (as in Section 2.3.2). We remark that there is no optimal choice of the frame, but that particular frames may give better empirical results for certain functions. For example, if f has elongated features or filaments, a curvelet frame should produce good reconstructions, whereas a multiscale system as in Section 2.3.2 will work well for piecewise constant functions.

REMARK 9 (Alternative algorithms). In [30], the optimization problem (4) was solved by an Alternating Direction Method of Multipliers (ADMM) approach, which 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 due to the large amount and redundancy of the constraints. Additionally, the convergence guarantees for the projection step can be quite slow if the sets in which we project intersect with a small angle. Instead, by using the CP algorithm (which is a preconditioned version of ADMM algorithm) and using a different splitting, we circumvent the projection step and replace it by the soft-thresholding step mentioned above.

6. Proof of the main theorems.

6.1. *Proof of Theorem 1.* We begin with a preparation.

PROPOSITION 4. *Let Φ satisfy Assumption 1 and, for $n \in \mathbb{N}$, let $\hat{f}_\Phi$ be the estimator defined in (4) with γ_n given by (5). Then conditionally on the event A_n in (32) we have*:

$$\text{(i)} \quad \|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}(\mathbb{T}^d)} \le C\gamma_n + C\frac{\|f\|_{L^\infty(\mathbb{T}^d)} + \log n}{\sqrt{n}},$$

$$\text{(ii)} \quad \|\hat{f}_\Phi - f\|_{BV(\mathbb{T}^d)} \le \|f\|_{L^\infty(\mathbb{T}^d)} + 2|f|_{BV(\mathbb{T}^d)} + \log n,$$

for any $f \in BV \cap L^\infty$, and a constant $C > 0$ independent of n, f and $\hat{f}_\Phi$.

PROOF. For part (i), apply Assumption 1 to $g = \hat{f}_\Phi - f$, which yields

$$\|\hat{f}_\Phi - f\|_{B^{-d/2}_{\infty,\infty}(\mathbb{T}^d)} \le C \max_{\omega\in\Omega_n} |\langle \phi_\omega, \hat{f}_\Phi - f\rangle| + C\frac{\|\hat{f}_\Phi - f\|_{L^\infty(\mathbb{T}^d)}}{\sqrt{n}}.$$

The numerator in the second term can be bounded by $\|f\|_{L^\infty(\mathbb{T}^d)} + \log n$ by construction of $\hat{f}_\Phi$, while the first term can be bounded as

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

conditionally on A_n, where in the second inequality we used the definition of $\hat{f}_\Phi$. This completes the proof of (i). For (ii), we have

$$\begin{aligned}\|\hat{f}_\Phi - f\|_{BV(\mathbb{T}^d)} &\le \|\hat{f}_\Phi - f\|_{L^1(\mathbb{T}^d)} + |\hat{f}_\Phi - f|_{BV(\mathbb{T}^d)} \\ &\le \|\hat{f}_\Phi - f\|_{L^\infty(\mathbb{T}^d)} + |\hat{f}_\Phi - f|_{BV(\mathbb{T}^d)}.\end{aligned}$$

The first term is bounded by $\|f\|_{L^\infty(\mathbb{T}^d)} + \log n$, while the second is bounded by $|\hat{f}_\Phi|_{BV(\mathbb{T}^d)} + |f|_{BV(\mathbb{T}^d)}$. Finally, conditionally on A_n we have $|\hat{f}_\Phi|_{BV(\mathbb{T}^d)} \le |f|_{BV(\mathbb{T}^d)}$. This is so because $\hat{f}_\Phi$ is defined as the minimizer of the bounded variation seminorm among the functions satisfying $\max_{\omega\in\Omega_n} |\langle \phi_\omega, g\rangle - Y_\omega| \le \gamma_n$. Note that, conditionally on A_n, the function f satisfies this constraint, and hence f is an admissible function for the minimization problem defining $\hat{f}_\Phi$, whence $|\hat{f}_\Phi|_{BV(\mathbb{T}^d)} \le |f|_{BV(\mathbb{T}^d)}$. This completes the proof. □

The proof of Theorem 1 relies heavily on results from the theory of function spaces. In particular, we use the following interpolation inequalities.

PROPOSITION 5 (Interpolation inequalities). (a) *For $d = 1$ and $q \in [1, 3]$, there is a constant $C > 0$ such that*

$$\|g\|_{L^q} \le C(\log n)\|g\|^{2/3}_{B^{-1/2}_{\infty,\infty}}\|g\|^{1/3}_{BV} + Cn^{-1}\|g\|^{2/3}_{L^\infty}\|g\|^{1/3}_{BV}$$

holds for any $n \in \mathbb{N}$ and any $g \in L^\infty \cap BV(\mathbb{T}^d)$.

(b) *Let $d \ge 2$ and $q \in [1, \frac{d+2}{d}]$. Then there is a constant $C > 0$ such that*

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

holds for any $g \in B^{-d/2}_{\infty,\infty} \cap BV(\mathbb{T}^d)$.

We give the proof of Proposition 5 in Section 1 of the Supplementary Material [19]. It is the generalization to periodic functions of a result by [14]. The different results in $d = 1$ and $d \ge 2$ in Proposition 5 are due to the nature of certain embeddings between Besov and L^q spaces.

PROOF OF PART (a) OF THEOREM 1. We prove the claim of part (a) Theorem 1 conditionally on the event

$$A_n := \left\{\max_{\omega\in\Omega_n}\left|\int_{\mathbb{T}^d}\phi_\omega(x)\,dW(x)\right| \le \frac{\sqrt{n}}{\sigma}\gamma_n\right\}. \tag{32}$$

By the choice of γ_n in (5) and part (b) of Remark 1, we have $\mathbb{P}(A_n) \ge 1 - (\#\Omega_n)^{1-\kappa^2}$, which tends to one as $n \to \infty$.

Consider first the case $q \le 1 + 2/d$. For $d \ge 2$, part (b) of Proposition 5 applies and gives the interpolation inequality

$$\|\hat f_\Phi - f\|_{L^q(\mathbb{T}^d)} \le C\|\hat f_\Phi - f\|^{\frac{2}{d+2}}_{B^{-d/2}_{\infty,\infty}(\mathbb{T}^d)}\|\hat f_\Phi - f\|^{\frac{d}{d+2}}_{BV(\mathbb{T}^d)}.$$

Conditionally on A_n, Proposition 4 gives us bounds for the terms in the right-hand side, and using that $f \in BV_L$ gives

$$\begin{aligned}\|\hat f_\Phi - f\|_{L^q(\mathbb{T}^d)} &\le C\left(\gamma_n + C\frac{\|f\|_{L^\infty(\mathbb{T}^d)} + \log n}{\sqrt{n}}\right)^{\frac{2}{d+2}}\left(\|f\|_{L^\infty(\mathbb{T}^d)} + 2|f|_{BV(\mathbb{T}^d)} + \log n\right)^{\frac{d}{d+2}}\\ &\le Cn^{-\frac{1}{d+2}}\left(\sigma\sqrt{\log\#\Omega_n} + L + \log n\right)^{\frac{2}{d+2}}(L + \log n)^{\frac{d}{d+2}}\\ &\le C\left(1 \vee \sigma^2\right)^{\frac{1}{d+2}}n^{-\frac{1}{d+2}}\log n.\end{aligned}$$

Since $\#\Omega_n \le Q(n)$ grows at most polynomially in n, the claim follows.

For the case $d = 1$, we use part (a) of Proposition 5, which yields

$$\|g\|_{L^q} \le C(\log n)\|g\|^{2/3}_{B^{-1/2}_{\infty,\infty}}\|g\|^{1/3}_{BV} + Cn^{-1}\|g\|^{2/3}_{L^\infty}\|g\|^{1/3}_{BV}$$

for $g = \hat f_\Phi - f$ and $q \in [1, 3]$. Proposition 4 now gives, conditionally on A_n,

$$\|\hat f_\Phi - f\|_{L^q} \le C\left(1 \vee \sigma^2\right)^{1/3}n^{-1/3}(\log n)^2 + Cn^{-1}\log n,$$

which yields the claim.

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

PROOF OF PART (b) OF THEOREM 1. Using the convergence conditionally on A_n proved in part (a), we can bound the expected risk as

$$\begin{aligned}\mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q(\mathbb{T}^d)}\big] &= \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q(\mathbb{T}^d)} 1_{A_n}\big] + \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q(\mathbb{T}^d)} 1_{A_n^c}\big] \\ &\le C r_n \mathbb{P}(A_n) + \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q(\mathbb{T}^d)} 1_{A_n^c}\big] \\ &\le C r_n + \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q(\mathbb{T}^d)} 1_{A_n^c}\big],\end{aligned} \tag{33}$$

where $r_n = ((1 \vee \sigma^2)/n)^{\min\{\frac{1}{d+2}, \frac{1}{dq}\}} (\log n)^{3-\min\{d,2\}}$. The rest of the proof consists in showing that the second term behaves as $o(n^{-1/2})$ for $\kappa^2 > 1 + \frac{1}{(d+2)\Gamma}$. By assumption we have the bounds $\|f\|_{L^\infty} \le L$ and $\|\hat{f}_\Phi\|_{L^\infty} \le \log n$, so we can bound the second term as

$$\mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^q(\mathbb{T}^d)} 1_{A_n^c}\big] \le \mathbb{E}\big[\|\hat{f}_\Phi - f\|_{L^\infty(\mathbb{T}^d)} 1_{A_n^c}\big] \le (L + \log n)\mathbb{P}(A_n^c).$$

By part (b) of Remark 1 we have $\mathbb{P}(A_n^c) \le (\#\Omega)^{1-\kappa^2}$, and inserting this back in (33) yields

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

Choosing $\kappa^2 > 1 + 1/((d+2)\Gamma)$ yields the claim. □

6.2. *Minimax rate over BV*. Here we prove Theorem 2 by showing a lower bound for the minimax risk over Besov spaces $B^1_{1,1}$ with respect to the L^q-risk. This implies a lower bound for the minimax risk over BV_L, since

$$BV_L \supset \big(B^1_{1,1} \cap L^\infty\big)_L := \big\{g \in B^1_{1,1} \mid \|g\|_{B^1_{1,1}} \le L, \|g\|_{L^\infty} \le L\big\}.$$

The minimax L^q-risk for $q \le 1 + 2/d$ (dense case) is well understood, and the associated minimax rates have been known for a while to be $n^{-\frac{1}{d+2}}$. Its proof follows the classical strategy of constructing a set of alternatives in $(B^1_{1,1} \cap L^\infty)_L$ that are well separated in the L^q-norm, and applying an information inequality (e.g., Fano's inequality). It can be found in Chapter 10 of [43], so we do not reproduce it here.

On the other hand, the regime $q \ge 1 + 2/d$ is far less popular, and we have not found any proof of what the minimax rate is there. The difficulty here is that $B^1_{1,1}$ is a Besov space with "$s \le d/p$", and the literature has focused mainly on the case $s > d/p$ (with some exceptions, see [36] and [52]). Our proof that the minimax rate is $O(n^{-\frac{1}{dq}})$ in that regime follows the same idea as in the other regimes: we construct a set of well separated alternatives and show that no statistical procedure can distinguish them perfectly. As in the dense regime, our construction is based on Assouad's cube [1].

PROOF OF THEOREM 2. Our proof follows the proof of Theorem 10.3 in [43] closely. We structure it in several steps.

Construction of alternatives: Let $g_0 \in B^1_{1,1} \cap L^\infty$ satisfy

$$\|g_0\|_{B^1_{1,1}} \le L/2 \quad \text{and} \quad \|g_0\|_{L^\infty} \le L/2.$$

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

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

Since Daubechies wavelets are compactly supported, there are at most $O(2^{jd})$ such wavelets with nonoverlapping supports. We will not need all of them, but only a subset of cardinality

$\#R_j = S_j = \lfloor 2^{j\Delta} \rfloor$ for a real number $\Delta \in [0, d]$ to be chosen later. Consider now vectors $\varepsilon \in \{-1, +1\}^{S_j}$ with components indexed by $(k, e) \in R_j$. Our alternatives will have the form

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

for $\gamma > 0$ to be chosen later. Define the set $\mathcal{G} := \{g^{\varepsilon} \mid \varepsilon \in \{-1, +1\}^{S_j}\}$. Notice that all functions in this set satisfy $\|g^{\varepsilon}\|_{B^1_{1,1}} \le L$ and $\|g^{\varepsilon}\|_{L^\infty} \le L$ provided that

$$\gamma \le \frac{L}{2} 2^{-j(1-d/2+\Delta)} \quad \text{and} \quad \gamma \le \frac{L}{2\|\psi\|_{L^\infty}} 2^{-jd/2}, \tag{34}$$

respectively. In the following, we choose $\Delta = d - 1$ in order to balance these two terms. Finally, the L^q-separation between these alternatives is

$$\delta := \inf_{\varepsilon \neq \varepsilon'} \|g^{\varepsilon} - g^{\varepsilon'}\|_{L^q} = 2^{1/q} \|\gamma \psi_{j,k,e}\|_{L^q} = 2^{1/q} \gamma 2^{jd(\frac{1}{2}-\frac{1}{q})} \|\psi\|_{L^q}, \tag{35}$$

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 $(B^1_{1,1} \cap L^\infty)_L$. We reproduce the claim (Lemma 10.2 in [43]) for completeness.

LEMMA 1. *For $\varepsilon \in \{-1, +1\}^{S_j}$ and $(k, e) \in R_j$, define $\varepsilon_{*k} := (\varepsilon'_{(k_1,e_1)}, \ldots, \varepsilon'_{(k_{S_j},e_{S_j})})$, 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 λ, $p_0 > 0$ such that

$$\mathbb{P}_{g^{\varepsilon}}\big(LR(g^{\varepsilon_{*k}}, g^{\varepsilon}) > e^{-\lambda}\big) \ge p_0 \quad \forall \varepsilon, n, \tag{36}$$

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

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

where δ is defined in (35).

Verification of (36): The condition (36) is easily verified in our setting with Gaussian observations under the condition that $n\gamma^2 \le c$ for n large enough (see Section 10.5 in [43]). Indeed, by Markov's inequality we have

$$\mathbb{P}_{g^{\varepsilon}}\big(LR(g^{\varepsilon_{*k}}, g^{\varepsilon}) > e^{-\lambda}\big) \ge 1 - \frac{1}{\log e^{\lambda}} \mathbb{E}_{g^{\varepsilon}} |\log LR(g^{\varepsilon_{*k}}, g^{\varepsilon})|,$$

and using Proposition 6.1.7 in [34] to bound the expectation by the Kullback–Leibler divergence we get

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

Using the Cameron–Martin theorem to interpret the Gaussian probability measures (see Theorem 2.6.13 in [34]), the Kullback–Leibler divergence between Gaussian measures is easily computer and gives

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

Hence, choosing $\gamma = t_0 \sigma n^{-1/2}$ for a small enough constant $t_0 > 0$ gives (36).

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 $\mathcal{G}$, that is,

$$\sup_{f \in (B^1_{1,1} \cap L^\infty)_L} \mathbb{E}_f \|\hat{f} - f\|_{L^q} \geq \sup_{g^\varepsilon \in \mathcal{G}} \mathbb{E}_{g^\varepsilon} \|\hat{f} - g^\varepsilon\|_{L^q} \geq \frac{e^{-\lambda} p_0}{2} \delta 2^{j\Delta/q} \tag{37}$$

for any estimator $\hat{f}$. It remains to choose the scale parameter $j \geq 0$. Recall that we have chosen $\gamma = t_0 \sigma n^{-1/2}$. Further, by (34) we also need $\gamma \leq c 2^{-j(1-d/2+\Delta)} = c 2^{-jd/2}$, for the choice $\Delta = d - 1$. We choose j such that $2^{-jd/2} = c\sigma n^{-1/2}$, which gives the bound in (37)

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

7. Summary and outlook. We presented a family of estimators in the Gaussian white noise model defined by minimization of the BV-seminorm under a constraint on the frame coefficients of the residuals. Under conditions on the frame that amount to a certain compatibility with the Besov space $B^{-d/2}_{\infty,\infty}$, we show that these estimators attain the minimax optimal rate of convergence in any dimension up to logarithmic factors. There are still several open questions regarding extensions of our estimator. First, in the extension to a nonparametric regression model with discretely sampled data presented in Section 4. There we showed that an additional discretization error appears that slows down the rate as compared with the white noise model. We do not know whether this rate is sharp in a minimax sense (up to logarithmic factors). Notice that the asymptotic equivalence of the white noise and the multivariate nonparametric regression models derived by [69] does not apply for functions of bounded variation, so the minimax rates need not be the same in the two models. It is thinkable that the discretization error arises from the use of a continuous L^q-risk in the discretized model, in which case the discrete ℓ^q risk might be better suited. We leave the clarification of this question for future research.

A second question concerns the relation between the multiscale data-fidelity and statistical testing. In fact, our use of dictionary elements with L^2-norm equal to one is analogous to the multiplicative scaling used by [27] to correctly weight their multiresolution test statistics. This raises the question of whether an *additive scaling* in our data-fidelity is necessary in our setting, as it is in theirs. The answer is that such an additive scaling would help us remove some (but not all) of the logarithmic terms in the error bound in Theorem 1. However, it would imply additional difficulties in the theoretical analysis of the estimator, since the constraint would no longer match the Besov scale exactly. Alternatively, a different multiplicative scaling could be used to link the multiscale data-fidelity with the *logarithmic* Besov spaces (see Section 4.4 in [34]). We leave as an open question whether these modified data-fidelities and Besov spaces could yield an improved performance.

Another interesting question concerns the choice of the risk functional. We have proven convergence rates with respect to the L^q-risk, which measures the *global* error made by the estimator. In contrast, the use of multiscale risk functionals has been proposed as an alternative quality measure which takes spatial adaptation into account (see, e.g., [4] and [55]). We expect that estimators like (4) should perform particularly well with respect to such multiscale risks, and postpone the answer to that question for future work.

The extension of our theory to statistical inverse problems is particularly attractive, since in many applications one only has access to a transformed version of the object of interest (see, e.g., [31] and [65] for applications of TV-regularization to microscopy and tomography, respectively). The analysis done in the present paper is expected to be adaptable to inverse problems if the operator is assumed to have "good" mapping properties in

the Besov scale $B^s_{\infty,\infty}$. The modification would essentially involve a constraint of the form $\max_{\omega\in\Omega_n} |\langle \phi_\omega, Tg\rangle - Y_\omega| \le \gamma_n$ in (4), where T is the forward operator (see [31] and [55] for examples and analysis of such an estimator). From this constraint, it is apparent that the dictionary Φ has to depend on the forward operator T (see [68] for a similar construction). Finally, the extension to non-Gaussian noise models is of interest in many applications. In that respect, note that the analysis of the estimator (4) depends on the tail behavior of the statistic $\max_{\omega\in\Omega_n} |\langle \phi_\omega, dW\rangle|$ being *sub-Gaussian*. Finally, the extension to SDE-based models (see, e.g., [35]) appears to us of interest.

Acknowledgments. We thank two anonymous referees for their insightful comments that improved the scope and quality of the paper.

Funding. The first author was supported by DFG RTG 2088-B2.

The second and third authors were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2067/1-390729940.

The second author was supported by DFG CRC 937-A10.

The third author was supported by DFG CRC 755-A4.

SUPPLEMENTARY MATERIAL

Supplement to "Frame-constrained total variation regularization for white noise regression" (DOI: 10.1214/20-AOS2001SUPP; .pdf). This Supplement is organized as follows. In Section 1, we prove the interpolation inequalities of Proposition 5. In Section 2, we prove Propositions 1, 2 and 3. In Sections 3 and 4, we prove Theorems 3 and 4, respectively.

REFERENCES

[1] ASSOUAD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.* **296** 1021–1024. MR0777600

[2] BORUP, L. and NIELSEN, M. (2007). Frame decomposition of decomposition spaces. *J. Fourier Anal. Appl.* **13** 39–70. MR2296727 https://doi.org/10.1007/s00041-006-6024-y

[3] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958 https://doi.org/10.1214/aos/1032181159

[4] CAI, T. T. and LOW, M. G. (2005). Nonparametric estimation over shrinking neighborhoods: Superefficiency and adaptation. *Ann. Statist.* **33** 184–213. MR2157801 https://doi.org/10.1214/009053604000000832

[5] CANDÈS, E. J. and DONOHO, D. L. (2000). Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Technical report, Stanford Univ., California, Dept. of Statistics.

[6] CANDÈS, E. J. and DONOHO, D. L. (2004). New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.* **57** 219–266. MR2012649 https://doi.org/10.1002/cpa.10116

[7] CANDÈS, E. J. and GUO, F. (2002). New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction. *Signal Process.* **82** 1519–1543.

[8] CHAMBOLLE, A. (2004). An algorithm for total variation minimization and applications. *J. Math. Imaging Vision* **20** 89–97. MR2049783 https://doi.org/10.1023/B:JMIV.0000011320.81911.38

[9] CHAMBOLLE, A. and LIONS, P.-L. (1997). Image recovery via total variation minimization and related problems. *Numer. Math.* **76** 167–188. MR1440119 https://doi.org/10.1007/s002110050258

[10] CHAMBOLLE, A. and POCK, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40** 120–145. MR2782122 https://doi.org/10.1007/s10851-010-0251-1

[11] 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** 199–231. MR2657627 https://doi.org/10.1137/090758003

[12] CLASON, C., KRUSE, F. and KUNISCH, K. (2018). Total variation regularization of multi-material topology optimization. *ESAIM Math. Model. Numer. Anal.* **52** 275–303. MR3808161 https://doi.org/10.1051/m2an/2017061

[13] COHEN, A. (2003). *Numerical Analysis of Wavelet Methods. Studies in Mathematics and Its Applications* **32**. North-Holland, Amsterdam. MR1990555

[14] COHEN, A., DAHMEN, W., DAUBECHIES, I. and DEVORE, R. (2003). Harmonic analysis of the space BV. *Rev. Mat. Iberoam.* **19** 235–263. MR1993422 https://doi.org/10.4171/RMI/345

[15] CONDAT, L. (2017). Discrete total variation: New definition and minimization. *SIAM J. Imaging Sci.* **10** 1258–1290. MR3684410 https://doi.org/10.1137/16M1075247

[16] DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. MR3556784 https://doi.org/10.3150/15-BEJ756

[17] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. SIAM, Philadelphia, PA. MR1162107 https://doi.org/10.1137/1.9781611970104

[18] DAVIES, P. L. and KOVAC, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29** 1–65. MR1833958 https://doi.org/10.1214/aos/996986501

[19] DEL ÁLAMO, M., LI, H. and MUNK, A. (2021). Supplement to "Frame-constrained total variation regularization for white noise regression." https://doi.org/10.1214/20-AOS2001SUPP

[20] DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. MR1400080 https://doi.org/10.1006/acha.1996.0017

[21] DENG, H. and ZHANG, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs. *Ann. Statist.* To appear.

[22] 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** 82–104. MR2782120 https://doi.org/10.1007/s10851-010-0248-9

[23] DONOHO, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.* **1** 100–115. MR1256530 https://doi.org/10.1006/acha.1993.1008

[24] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414 https://doi.org/10.1214/aos/1024691081

[25] STARCK, J. L., DONOHO, D. L. and CANDÈS, E. J. (2001). Very high quality image restoration by combining wavelets and curvelets. In *Wavelet Applications in Signal and Image Processing IX* **4478** 9–19.

[26] DÜMBGEN, L. and KOVAC, A. (2009). Extensions of smoothing via taut strings. *Electron. J. Stat.* **3** 41–75. MR2471586 https://doi.org/10.1214/08-EJS216

[27] DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961 https://doi.org/10.1214/aos/996986504

[28] DURAND, S. and FROMENT, J. (2001). Artifact free signal denoising with wavelets. In *Acoustics, Speech, and Signal Processing*, 2001. *Proceedings.* (*ICASSP'*01). 2001 *IEEE International Conference on* **6** 3685–3688. IEEE, New York.

[29] EVANS, L. C. and GARIEPY, R. F. (2015). *Measure Theory and Fine Properties of Functions. Textbooks in Mathematics*. CRC Press, Boca Raton, FL. MR3409135

[30] 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. MR2988407 https://doi.org/10.1214/12-EJS671

[31] 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** 370–387. MR3068534 https://doi.org/10.1007/s10851-012-0368-5

[32] FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

[33] GARNETT, J. B., LE, T. M., MEYER, Y. and VESE, L. A. (2007). Image decompositions using bounded variation and generalized homogeneous Besov spaces. *Appl. Comput. Harmon. Anal.* **23** 25–56. MR2333827 https://doi.org/10.1016/j.acha.2007.01.005

[34] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[35] GOBET, E., HOFFMANN, M. and REISS, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* **32** 2223–2253. MR2102509 https://doi.org/10.1214/009053604000000797

[36] GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields* **159** 479–543. MR3230001 https://doi.org/10.1007/s00440-013-0512-1

[37] GRASMAIR, M., LI, H. and MUNK, A. (2018). Variational multiscale nonparametric regression: Smooth functions. *Ann. Inst. Henri Poincaré Probab. Stat.* **54** 1058–1097. MR3795077 https://doi.org/10.1214/17-AIHP832

[38] GUNTUBOYINA, A., LIEU, D., CHATTERJEE, S. and SEN, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** 205–229. MR4065159 https://doi.org/10.1214/18-AOS1799

[39] GUO, K., KUTYNIOK, G. and LABATE, D. (2006). Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines*: *Athens* 2005. *Mod. Methods Math.* 189–201. Nashboro Press, Brentwood, TN. MR2233452

[40] HADDAD, A. and MEYER, Y. (2007). An improvement of Rudin–Osher–Fatemi model. *Appl. Comput. Harmon. Anal.* **22** 319–334. MR2311857 https://doi.org/10.1016/j.acha.2006.09.001

[41] HALTMEIER, M. and MUNK, A. (2014). Extreme value analysis of empirical frame coefficients and implications for denoising by soft-thresholding. *Appl. Comput. Harmon. Anal.* **36** 434–460. MR3175087 https://doi.org/10.1016/j.acha.2013.07.004

[42] HAN, Q., WANG, T., CHATTERJEE, S. and SAMWORTH, R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[43] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (2012). *Wavelets, Approximation, and Statistical Applications*. *Lecture Notes in Statistics* **129**. Springer, New York. MR1618204 https://doi.org/10.1007/978-1-4612-2222-4

[44] HÜTTER, J. C. and RIGOLLET, P. (2016). Optimal rates for total variation denoising. In *Conference on Learning Theory* 1115–1146.

[45] JIANG, H. (2014). *Photoacoustic Tomography*. CRC Press, Boca Raton.

[46] LABATE, D., LIM, W. Q., KUTYNIOK, G. and WEISS, G. (2005). Sparse multidimensional representation using shearlets. In *Wavelets XI* **5914** 59140U. International Society for Optics and Photonics.

[47] LABATE, D., MANTOVANI, L. and NEGI, P. (2013). Shearlet smoothness spaces. *J. Fourier Anal. Appl.* **19** 577–611. MR3048591 https://doi.org/10.1007/s00041-013-9261-x

[48] LASSAS, M., SAKSMAN, E. and SILTANEN, S. (2009). Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging* **3** 87–122. MR2558305 https://doi.org/10.3934/ipi.2009.3.87

[49] LASSAS, M. and SILTANEN, S. (2004). Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl.* **20** 1537–1563. MR2109134 https://doi.org/10.1088/0266-5611/20/5/013

[50] LEDOUX, M. (2003). On improved Sobolev embedding theorems. *Math. Res. Lett.* **10** 659–669. MR2024723 https://doi.org/10.4310/MRL.2003.v10.n5.a9

[51] FANG, B., GUNTUBOYINA, A. and SEN, B. (2020). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *Ann. Statist.* To appear.

[52] LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.* **43** 1178–1242. MR3346701 https://doi.org/10.1214/14-AOS1306

[53] LEPSKI, 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. *Ann. Statist.* **25** 929–947. MR1447734 https://doi.org/10.1214/aos/1069362731

[54] LEPSKII, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

[55] LI, H. (2016). Variational estimators in statistical multiscale analysis. Ph.D. thesis, Georg-August-Universität Göttingen.

[56] LI, H., GUO, Q. and MUNK, A. (2019). Multiscale change-point segmentation: Beyond step functions. *Electron. J. Stat.* **13** 3254–3296. MR4010980 https://doi.org/10.1214/19-ejs1608

[57] LI, H., HALTMEIER, M., ZHANG, S., FRAHM, J. and MUNK, A. (2014). Aggregated motion estimation for real-time MRI reconstruction. *Magn. Reson. Med.* **72** 1039–1048.

[58] MALGOUYRES, F. (2001). A unified framework for image restoration. Technical report, Univ. of California, Los Angeles.

[59] MALGOUYRES, F. (2002). Mathematical analysis of a model which combines total variation and wavelet for image restoration. *J. Inf. Process.* **2** 1–10.

[60] MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

[61] MEYER, Y. (2001). *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*: *The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures*. *University Lecture Series* **22**. Amer. Math. Soc., Providence, RI. MR1852741 https://doi.org/10.1090/ulect/022

[62] 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** 19–41. MR2136637 https://doi.org/10.1111/j.1467-9868.2005.00486.x

[63] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Springer, Berlin. MR1775640

[64] NEMIROVSKIY, A. S. (1985). Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk SSSR Tekhn. Kibernet*. **3** 50–60. MR0844292

[65] NIINIMÄKI, K., LASSAS, M., HÄMÄLÄINEN, K., KALLONEN, A., KOLEHMAINEN, V., NIEMI, E. and SILTANEN, S. (2016). Multiresolution parameter choice method for total variation regularized tomography. *SIAM J. Imaging Sci*. **9** 938–974. MR3521540 https://doi.org/10.1137/15M1034076

[66] NIRENBERG, L. (1959). On elliptic partial differential equations. *Ann. Sc. Norm. Super. Pisa Cl. Sci*. (3) **13** 115–162. MR0109940

[67] OSHER, S., SOLÉ, A. and VESE, L. (2003). Image decomposition and restoration using total variation minimization and the H^{-1} norm. *Multiscale Model. Simul*. **1** 349–370. MR2030155 https://doi.org/10.1137/S1540345902416247

[68] PROKSCH, K., WERNER, F. and MUNK, A. (2018). Multiscale scanning in inverse problems. *Ann. Statist*. **46** 3569–3602. MR3852662 https://doi.org/10.1214/17-AOS1669

[69] REISS, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist*. **36** 1957–1982. MR2435461 https://doi.org/10.1214/07-AOS525

[70] RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D, Nonlinear Phenom*. **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

[71] 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* 3513–3521.

[72] SCHERZER, O., GRASMAIR, M., GROSSAUER, H., HALTMEIER, M. and LENZEN, F. (2009). *Variational Methods in Imaging*. *Applied Mathematical Sciences* **167**. Springer, New York. MR2455620

[73] SCHMEISSER, H.-J. and TRIEBEL, H. (1987). *Topics in Fourier Analysis and Function Spaces*. *A Wiley-Interscience Publication*. Wiley, Chichester. MR0891189

[74] SPOKOINY, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivariate Anal*. **82** 111–133. MR1918617 https://doi.org/10.1006/jmva.2001.2023

[75] TRIEBEL, H. (1988). Characterizations of Besov–Hardy–Sobolev spaces: A unified approach. *J. Approx. Theory* **52** 162–203. MR0929302 https://doi.org/10.1016/0021-9045(88)90055-X

[76] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[77] WAHBA, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal*. **14** 651–667. MR0471299 https://doi.org/10.1137/0714044

[78] WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res*. **17** Paper No. 105, 41. MR3543511