Variational Multiscale Nonparametric Regression: Smooth Functions

Markus Grasmair
Department of Mathematical Sciences
Norwegian University of Science and Technology, Trondheim, Norway

Housen Li, and Axel Munk
Institute for Mathematical Stochastics, University of Göttingen
and Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

March 27, 2018

Abstract

For the problem of nonparametric regression of smooth functions, we reconsider and analyze a constrained variational approach, which we call the MultIscale Nemirovski-Dantzig (MIND) estimator. This can be viewed as a multiscale extension of the Dantzig selector (*Ann. Statist.*, 35(6): 2313–51, 2009) based on early ideas of Nemirovski (*J. Comput. System Sci.*, 23:1–11, 1986). MIND minimizes a homogeneous Sobolev norm under the constraint that the multiresolution norm of the residual is bounded by a universal threshold. The main contribution of this paper is the derivation of convergence rates of MIND with respect to L^q-loss, $1 \leq q \leq \infty$, both almost surely and in expectation. To this end, we introduce the method of approximate source conditions. For a one-dimensional signal, these can be translated into approximation properties of B-splines. A remarkable consequence is that MIND attains almost minimax optimal rates simultaneously for a large range of Sobolev and Besov classes, which provides certain adaptation. Complimentary to the asymptotic analysis, we examine the finite sample performance of MIND by numerical simulations.

Keywords: Nonparametric regression; adaptation; convergence rates; minimax optimality; multiresolution norm; approximate source conditions.

1 Introduction

In this paper, we will consider the nonparametric regression problem to estimate a smooth function $f\colon [0,1]^d \to \mathbb{R}$ from n measurements

$$y_n(x) = f(x) + \xi_n(x) \quad \text{for } x \in \Gamma_n, \tag{1}$$

where Γ_n is the regular grid on $[0,1]^d$ containing n equidistant points, and $\{\xi_n(x) : x \in \Gamma_n\}$ a set of independent, identically distributed (i.i.d.) centered sub-Gaussian random variables with scale parameter

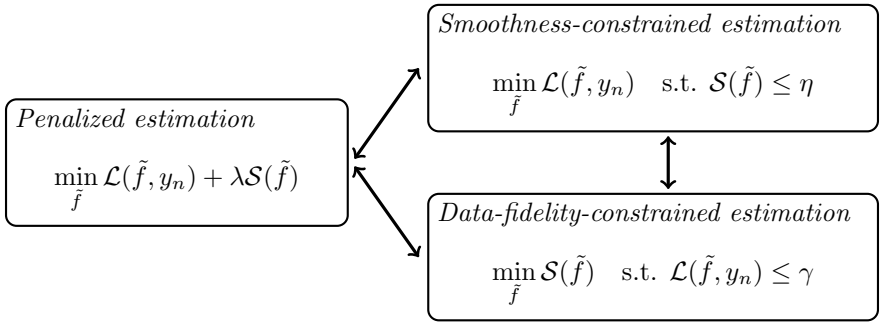

Figure 1: Variational statistical estimation.

σ, i.e., the common distribution function Φ satisfies

$$\int e^{\tau t} \Phi(dt) \le e^{(\tau\sigma)^2/2} \quad \text{for every } \tau \in \mathbb{R}. \tag{2}$$

For simplicity, we assume that the truth f can be extended periodically to $\mathbb{R}^d$ to avoid boundary effects, and that the noise level σ is known.

1.1 Variational statistical estimation

Since the fundamental work of (Nadaraya, 1964; Stone, 1984) and many others, the literature on non-parametric regression techniques has become enormously rich and diverse, and has found its way into many textbooks, see (Green and Silverman, 1994; Fan and Gijbels, 1996; Györfi et al., 2002; Tsybakov, 2009; Korostelev and Korosteleva, 2011) for example. A prodigious amount of these estimation methods can be casted in a variational framework, which can be roughly categorized into three different formulations: *penalized estimation*, *smoothness-constrained estimation*, and *data-fidelity-constrained estimation*, see Figure 1.

Penalized estimation is a solution of the Lagrangian variational problem (also known as generalized Tikhonov regularization)

$$\min_{f} \mathcal{L}(f, y_n) + \lambda \mathcal{S}(f). \tag{3}$$

The *regularization term* $\mathcal{S}(f)$ accounts for a-priori assumptions of the truth f, such as smoothness, sparsity, etc. The *data fidelity term* $\mathcal{L}(f, y_n)$ measures the deviation from the data y_n. If $\mathcal{L}(\cdot, y_n)$ is the log-likelihood function of the model, this amounts to penalized maximum-likelihood regression (see e.g. van de Geer, 1988; Eggermont and LaRiccia, 2009, for general exposition). Prominent examples include smoothing splines (Wahba, 1990), local polynomial estimators (Fan and Gijbels, 1996), and locally adaptive splines (Mammen and van de Geer, 1997). It is known that the choice of the balancing parameter λ is in general subtle, although there are nowadays many data driven strategies, such as (generalized) cross validation (Wahba, 1977), or Lepskiĭ's balancing principle (Lepskiĭ, 1990), to mention a few. The latter even provides adaptation over a range of generalized Sobolev scales, see e.g. (Goldenshluger and Nemirovski, 1997; Lepski et al., 1997).

Smoothness-constrained estimation is to minimize the data fidelity term $\mathcal{L}$ under the regularization constraint $\mathcal{S}$,

$$\min_f \mathcal{L}(f, y_n) \qquad \text{subject to } \mathcal{S}(f) \leq \eta. \tag{4}$$

It includes the well-known lasso (Tibshirani, 1996) for $\mathcal{L} = \|\cdot - y_n\|_{\ell^2}$ and $\mathcal{S} = \|\cdot\|_{\ell^1}$ as a special case. Another example is Nemirovski's (1985) estimator $\hat{f}_{p,\eta}$ defined by

$$\hat{f}_{p,\eta} \in \arg\min_f \|S_n f - y_n\|_{\mathcal{B}} \qquad \text{subject to } \|D^k f\|_{L^p} \leq \eta, \tag{5}$$

where S_n denotes the *sampling operator* on the grid Γ_n, and the *multiresolution norm* $\|\cdot\|_{\mathcal{B}}$ measures the maximum of normalized local averages (see Section 3 for a formal definition). The estimator $\hat{f}_{p,\eta}$ is known to be minimax optimal (up to at most a log-factor) over Sobolev ellipsoids $\{f; \|D^k f\|_{L^p} \leq \eta\} \subset W^{k,p}$, see (Nemirovski, 1985, 2000). This indicates one drawback of this type of estimator: the choice of the threshold η determines a priori the smoothness information (measured by $\mathcal{S}$) of the truth f, which is often unavailable in reality.

Data-fidelity-constrained estimation results from the "reverse" formulation of (4), given by

$$\min_f \mathcal{S}(f) \qquad \text{subject to } \mathcal{L}(f, y_n) \leq \gamma. \tag{6}$$

Many basis (or dictionary) based thresholding-type methods, such as soft-thresholding (Donoho, 1995a), and block thresholding (Hall et al., 1997; Cai, 1999, 2002; Cai and Zhou, 2009; Chesneau et al., 2010), can be written this way. Here $\gamma = \gamma_n$ can be chosen as a universal threshold, not depending on the data. For example, proper wavelet thresholding provides spatial adaptivity, and is known to be minimax optimal for smooth functions, see (Donoho and Johnstone, 1994; Donoho et al., 1995, 1996; Härdle et al., 1998), while at the same time computationally fast as the thresholding is applied to each empirical wavelet coefficient, separately. Such adaptivity of wavelet based methods is also known for more general settings, such as linear inverse problems, see e.g. (Donoho, 1995b; Cavalier et al., 2002; Cohen et al., 2004; Hoffmann and Reiss, 2008). The Dantzig selector (Candès and Tao, 2007) is also a particular data-fidelity-constrained estimator, originally introduced for linear models. In the nonparametric setting (1), it has the form

$$\min_{f \in \mathbb{R}^{\Gamma_n}} \|f\|_{\ell^1} \qquad \text{subject to } \|f - y_n\|_{\ell^\infty} \leq \gamma. \tag{7}$$

Many other ℓ^1-minimization approaches for recovering sparse signals also take the form of (6), see (Donoho et al., 2006; Cai et al., 2010) for example.

From a convex analysis point of view, all three estimation methods in Figure 1 can be viewed as equivalent, as under weak assumptions (Ivanov et al., 2002) each estimator in (3), (4), (6) can be obtained as a solution of the other optimization problems via Fenchel duality (cf. Bickel et al., 2009, for this in the case of the lasso and the Dantzig selector). The correspondence between the parameters λ, η, γ, however, is not given explicitly, and depends on the data y_n. It is exactly the lack of this explicit correspondence that makes the different statistical nature of these estimations. From this perspective, the data-fidelity-constrained estimation (6) has a certain appeal, since its threshold parameter can be chosen universally, i.e. only determined by the noise characteristics and the sample size n, and still allows for a sound statistical interpretation. For instance, it can often be chosen in such a way that the truth f satisfies the constraint on the r.h.s. of (6) with probability at least $1 - \alpha$, which immediately leads to the so called *smoothness guarantee* of the estimate $\hat{f}$ in (6),

$$\inf_f \mathbb{P}\left\{\mathcal{S}(\hat{f}) \leq \mathcal{S}(f)\right\} \geq 1 - \alpha.$$

1.2 MIND estimator

In the literature, multiscale data-fidelity-constrained methods which do not explicitly rely on a specific basis or dictionary and hence do not allow for component or blockwise thresholding have also been around for some while. For example, Nemirovski (1985) briefly discussed the "reverse" of his estimator (5) as well, which is given by

$$\min_f \|D^k f\|_{L^p} \qquad \text{subject to } \|S_n f - y_n\|_{\mathcal{B}} \leq \gamma. \tag{8}$$

These estimators all combine variational minimization with so called multiscale testing statistics. Empirically, they have been found to perform very well and even outperform those explicit methods based on wavelets or dictionaries (cf. Candès and Guo, 2002; Dong et al., 2011; Frick et al., 2013). In fact, the latter methods, as signal-to-noise ratio decreases, often show visually disturbing artifacts because of missing band pass information (Candès and Guo, 2002). On the other hand, the computation of such multiscale data-fidelity-constrained estimators, in general, leads to a high dimensional non-smooth convex optimization problem, remaining a burden for a long time. However, recently certain progress has been made in the development of algorithms for this type of problems (see Beck and Teboulle, 2009; Chambolle and Pock, 2011; Frick et al., 2012, for example). In the one dimensional case, fast algorithmic computation is sometimes feasible for specific functionals $\mathcal{S}$ (e.g. Davies and Kovac, 2001; Davies et al., 2009; Dümbgen and Kovac, 2009; Frick et al., 2014). In contrast to these computational achievements, the underlying statistical theory for these methods is currently not well understood, in particular with regard to their asymptotic convergence behavior. In fact, there is only a small number of results in this direction we are aware of: for fixed $k \in \mathbb{N}$ and $p \in [1, \infty]$, and under the somewhat artificial assumption that the truth f lies in the constraint on the r.h.s. of (8), Nemirovski (1985) derived the convergence rate of (8) (i.e. $\mathcal{S} := \|D^k \cdot\|_{L^p}$) which coincides with the minimax rate over Sobolev ellipsoids in $W^{k,p}$ up to a log-factor. Special cases of this result have also appeared in (Davies and Meise, 2008) for $k = p = 2$, and in (Davies et al., 2009) for $k = 2$, $p = \infty$. In particular, adaptation of this type of estimators has not been provided so far, to the best of our knowledge. Intending to fill such gap, we focus on the "reverse" Nemirovski estimator (8) with $p = 2$, that is,

$$\hat{f}_{\gamma_n} = \arg\min_f \frac{1}{2}\|D^k f\|_{L^2}^2 \qquad \text{subject to } \|S_n f - y_n\|_{\mathcal{B}} \leq \gamma_n, \tag{9}$$

which we call the *MultIscale Nemirovski-Dantzig estimator* (MIND). The choice of the name credits the fact that it is a particular "reverse" Nemirovski's estimator (8), and the r.h.s. is a (multiscale) extension of the Dantzig estimator (7).

The main contribution of this work is trifold. First, we introduce the approximate source conditions (Hofmann and Yamamoto, 2005; Hofmann, 2006) from regularization theory and inverse problems into the statistical analysis of nonparametric regression. By combining them with an improved interpolation inequality of the multiresolution norm and Sobolev norms, we are able to translate the statistical analysis into a deterministic approximation problem. The approximate source condition is essentially equivalent to smoothness concepts in terms of (approximate) variational inequalities (cf. Hofmann et al., 2007; Scherzer et al., 2009; Flemming and Hofmann, 2010) via Fenchel duality, see (Flemming, 2012); and conditions of this kind are fundamental for convergence analysis in inverse problems (see e.g. Engl et al., 1996, Section 3.2).

Second, we present both the L^q-risk convergence rate ($1 \leq q \leq \infty$) and the almost sure convergence rate for MIND, provided that an estimate of the approximate source condition is known. It is worth noting that the derivation of the L^q-risk convergence rate is more involved, for which one has to bound

the size of MIND, when the truth does not lie in the multiscale constraint, which notably extends Nemirovski (1985)'s technique. Our analysis for such situation is built on the observation that the MIND estimator is always close to the data, which leads us to an upper bound on its L^q-loss in terms of the multiresolution norm of the noise. The latter can be easily controlled because it has a sub-Gaussian tail.

Third, we show a *partial adaptation* property of MIND in one dimension, in the sense that for a fixed $k \in \mathbb{N}$, it attains minimax optimality (up to a log-factor) simultaneously over Sobolev ellipsoids in $W^{s,p}$ and Besov ellipsoids in $B^{s,p'}_p$ for all $(s,p) \in [1,k] \times \{\infty\} \cup [k+1, 2k] \times [2,\infty]$ and $p' \in [1,\infty]$. These results explain to some extent the remarkably good multiscale reconstruction properties of MIND empirically found in various signal recovery and imaging applications, see Section 6 and (Candès and Guo, 2002; Davies et al., 2009; Frick et al., 2013).

1.3 Organization of the paper

The rest of the paper is organized as follows: In Section 3, we present the multiresolution norm together with its deterministic and stochastic properties. Section 4 is devoted to approximate source conditions and so called distance functions, which provide methods for analyzing the L^q-loss ($1 \le q \le \infty$) of MIND. Combining such general results and an estimate of the distance functions, we obtain explicit convergence rates for smooth functions, in the one dimensional case, in Section 5. These rates are further shown to be minimax optimal up to a log-factor simultaneously over a large range of Besov and Sobolev classes. In addition to the asymptotic results, the finite sample behavior, as well as choices of the tuning parameter, of MIND is examined empirically on simulated examples in Section 6. The paper ends with discussions and open questions in Section 7. Technical proofs are given in the appendix.

2 A heuristic explanation to MIND

Before going into technical details, we illustrate the intuition behind MIND's ability to recover features of the truth in a multiscale fashion by a toy example.

Example 1. Let us consider the estimation of a smooth function $f\colon [0,1] \to \mathbb{R}$ from measurements

$$y_i = f\big(\frac{i}{n}\big) + \varepsilon_i \quad \text{for } i = 0, \dots, n-1,$$

with independent standard Gaussian error ε_i. Assume now that we have an estimator $\hat{f} \equiv \hat{f}_{s,t,a}$, such that

$$\hat{f}_{s,t,a}\big(\frac{i}{n}\big) := f\big(\frac{i}{n}\big) + s\varphi_a\big(\frac{i}{n}\big) + t\varepsilon_i \qquad \text{for } i = 0, \dots, n-1,$$

where $s, t \ge 0$, $a > 0$, $\varphi_a(x) := \sqrt{a}\varphi\left(a(x - 1/2)\right)$ and

$$\varphi(x) := \begin{cases} Ce^{\frac{1}{x^2-1}} & \text{if } |x| < 1 \\ 0 & \text{if } |x| \ge 1 \end{cases},$$

with the constant C satisfying $\|\varphi\|_{L^2} = 1$. That is, the estimator $\hat{f}$ differs from the truth f only by a deterministic distortion φ_a of scale a and a random perturbation $t\varepsilon$. By elementary computations one can show that

$$\Big|t - \frac{s}{\sqrt{a}}\Big| n \lesssim \|\hat{f} - f\|_{\ell^1} \lesssim \big(t + \frac{s}{\sqrt{a}}\big) n,$$

$$\|\hat{f} - f\|_{\ell^2} \sim (t+s)\sqrt{n},$$
$$\|\hat{f} - f\|_{\ell^\infty} \sim t\sqrt{\log n} + s\sqrt{a},$$

hold almost surely as $n \to \infty$.

These estimates indicate that the difference between f and the estimator $\hat{f}$ measured with respect to the ℓ^1-norm depends on the level of the random perturbation as well as the level and the scale of the deterministic distortion. Moreover, both the random and the deterministic part of the difference scale linearly with n, which indicates that the ℓ^1-norm is incapable of distinguishing random from deterministic deviations. For the ℓ^2-norm the situation is similar. In contrast, in case of the ℓ^∞-norm, the deterministic and the random part scale asymptotically differently, and thus the ℓ^∞-norm can, in principle, distinguish between these distortions. However, it also depends on the scale of the deterministic distortion; if the scale of the deterministic distortion is of order $\log n$, then again it is indistinguishable from random noise.

Now note that one can also show that

$$\frac{1}{\sqrt{2/(an)}}\Big| \sum_{\substack{i\in\mathbb{N} \\ -\frac{1}{a}\le\frac{i}{n}-\frac{1}{2}\le\frac{1}{a}}} \hat{f}(\frac{i}{n}) - f(\frac{i}{n})\Big| \sim t\sqrt{\log n} + s\sqrt{n}, \tag{10}$$

holds almost surely as $n \to \infty$. Here, the deterministic and the random parts scale differently, and the scale of the deterministic distortion does not influence the right hand side of (10). These favorable properties are, however, based on the prior knowledge of the support of the deterministic distortion φ_a, which explicitly appears on the left hand side of (10). Still, it is possible to use the local averages in (10) by taking the supremum over all possible scales and locations of deterministic perturbation, which, basically, results in the multiresolution norm. Later on we will see that this approach results in the same asymptotic estimate as (10). Therefore, the multiscale constraint of MIND with $\gamma_n \sim \log n$ guarantees that every feasible candidate contains no deterministic distortion, and the smoothness-enforcing regularization term selects the one without random distortion. The combination of both ensures that MIND avoids both deterministic and random distortions, thus close to the truth.

3 Multiresolution Norm

Define a *cube* B to be a subset of $[0,1]^d$ of the form $B = \prod_{i=1}^d [a_i, a_i + h)$, where $a_i \in \mathbb{R}$, $i = 1, \dots, d$, and $0 < h \le 1$. By $|B|$ we denote its d-dimensional volume h^d. As in (1), let Γ_n be the regular grid on $[0,1]^d$,

$$\Gamma_n := \Big\{ (\frac{\tau_1}{n^{1/d}}, \dots, \frac{\tau_d}{n^{1/d}}); \tau_i = 0, \dots, n^{1/d} - 1, \text{ for } i = 1, \dots, d \Big\}. \tag{11}$$

Definition 3.1. The *multiresolution norm* $\|\cdot\|_{\mathcal{B}}$ on $\mathbb{R}^{\Gamma_n}$ with respect to a non-empty system of cubes $\mathcal{B}$ is defined by

$$\|y\|_{\mathcal{B}} := \sup_{B\in\mathcal{B}} \frac{1}{\sqrt{n(B)}}\Big| \sum_{x\in\Gamma_n\cap B} y(x)\Big| \quad \text{for } y = \big(y(x)\big)_{x\in\Gamma_n} \in \mathbb{R}^{\Gamma_n}. \tag{12}$$

Here

$$n(B) := \#\Gamma_n \cap B.$$

Moreover, in the case of $n(B) = 0$, that is, $\Gamma_n \cap B = \emptyset$, we set the term $\left|\sum_{x\in\Gamma_n\cap B} y(x)\right|/\sqrt{n(B)}$ to zero.

Our main tool for the analysis of MIND will be estimates both for $\|\xi_n\|_{\mathcal{B}}$ and for $\|S_n f\|_{\mathcal{B}}$ that hold for sufficiently rich systems of cubes. Here and in the following S_n denotes the sampling operator on the set Γ_n in (11), provided that f is continuous.

Definition 3.2 (Nemirovski (1985)). A system $\mathcal{B}$ of distinct cubes is called *normal*, if there exists $c > 1$ such that for every cube $B \subseteq [0,1]^d$ there exists a cube $\tilde{B} \in \mathcal{B}$ such that $\tilde{B} \subseteq B$ and $|\tilde{B}| \geq |B|/c$.

In order to estimate the convergence rate of MIND for functions that are smoother than imposed by the regularization term, it is necessary to impose an additional regularity condition on the system B.

Definition 3.3. The system of cubes $\mathcal{B}$ is called *regular* (or *m-regular*) for some $m \in \mathbb{N}$, $m \geq 2$, if it contains at least the *m-partition system*, which is defined as all sets of the form $[\ell m^{-j}, (\ell+1)m^{-j})$ for all $\ell \in \mathbb{N}^d$, $j \in \mathbb{N}$.

Remark 3.4. Formally, a normal or regular system $\mathcal{B}$ is independent of the grid Γ_n. Given a grid Γ_n, the value of the multiresolution norm, however, depends on the intersection of the cubes in $\mathcal{B}$ with Γ_n. In particular, it is the *(effective) cardinality* of distinct cubes of $\mathcal{B}$ on Γ_n, that is, the number $\#\{B\cap\Gamma_n; B \in \mathcal{B}\}$, that determines the computational complexity of the evaluation of the multiresolution norm. In order to obtain numerically feasible algorithms, one therefore would like to choose this effective cardinality as small as possible while still ensuring normality or regularity of $\mathcal{B}$.

a) The system of all distinct cubes is clearly normal and regular. Its corresponding multiresolution norm also appears as a particular scan statistics (maximum likelihood ratio statistic in the Gaussian setting), which examines the signal at every scale and location. This is a standard tool for detecting a deterministic signal with unknown spatial extent against a noisy background, see e.g. (Glaz and Balakrishnan, 1999; Siegmund and Yakir, 2000; Dümbgen and Spokoiny, 2001). However, the cardinality of all the cubes on Γ_n is $\mathcal{O}\left(n^2\right)$, making it computationally impractical for large scale problems. In practice, sparser normal systems, while retaining multiresolution nature, are therefore preferable. Some examples are given below.

b) The system of cubes with dyadic edge lengths is normal, and of effective cardinality $\mathcal{O}(n \log n)$ on Γ_n. It is easy to see that this system is 2-regular.

c) Sparse system with optimal detection power. In one dimension, a normal system of $\mathcal{O}(n)$ intervals can be constructed from the system introduced in (Rivera and Walther, 2013), by including some intervals of small scales (i.e. length $\leq \log(n)/n$). This system is still sufficiently rich to be statistically optimal, in the setting of bump detection in the intensity of a Poisson process or in a density, but it is not regular. The heuristics behind is that after considering one interval, not much is gained by looking at intervals of similar scales and similar locations (see also Chan and Walther, 2013). For higher dimensions, such system can be constructed similarly, see (Walther, 2010; Sharpnack and Arias-Castro, 2014).

d) The m-partition system has effective cardinality $\mathcal{O}(n)$ on Γ_n, and is normal and obviously m-regular. As will be shown in Section 5, it is rich enough for nearly optimal estimation of smooth functions (see Section 6 for its practical performance). In particular, for $m = 2$, it corresponds to the support set of the wavelet multiresolution scheme.

It is clear that every regular system of cubes is necessarily normal. The converse, however, need not hold. That is, there exist normal systems of cubes that are not m-regular for any $m \in \mathbb{N}$ (cf. the second example above).

Finally, we note that the multiresolution norm is, actually, not necessarily a norm but always a semi-norm. That is, it can happen that $\|y\|_{\mathcal{B}} = 0$ although the vector $y \in \mathbb{R}^{\Gamma_n}$ is different from zero. Obviously this is the case if $B \cap \Gamma_n = \emptyset$ for all $B \in \mathcal{B}$, in which case $\|\cdot\|_{\mathcal{B}}$ is identically zero. If the system $\mathcal{B}$ is normal, however, this situation cannot occur for n sufficiently large: the normality of $\mathcal{B}$ implies in particular that $\mathcal{B}$ contains a cube of volume at least $1/c$, which, for $n > c$ necessarily has a non-empty intersection with the grid Γ_n. Still it is possible that $\|y\|_{\mathcal{B}} = 0$ for some non-zero y. On the other hand, if $\mathcal{B}$ is normal and $f\colon [0,1]^d \to \mathbb{R}$ is continuous and non-zero, then there exists some $n_0 \in \mathbb{N}$ such that $\|S_n f\|_{\mathcal{B}} \neq 0$ for all $n \geq n_0$, which means that the multiresolution norm of the point evaluation of a continuous non-zero function will eventually become non-zero. For simplicity, we will consider only systems $\mathcal{B}$, s.t. $\|\cdot\|_{\mathcal{B}}$ is a norm, which allows us later to define its dual norm. Moreover, in the important case of the m-partition system, it is easy to see that $\|\cdot\|_{\mathcal{B}}$ is indeed a norm, and that it can be bounded below by the maximum norm on $\mathbb{R}^{\Gamma_n}$.

The main property of the multiresolution norm is that it allows to distinguish between random noise and smooth functions, see Example 1. As the number n of sampling points increases, the multiresolution norm of a smooth function increases with a rate of $n^{1/2}$. In contrast, the multiresolution norm of i.i.d. Gaussian noise can be expected to grow only with a rate of $\sqrt{\log n}$. More precisely, the multiresolution norm has the following properties:

Proposition 3.5. *Let $\theta > 0$, $\mathcal{B}$ be a system of cubes and $\xi_n := \{\xi_n(x) : x \in \Gamma_n\}$ a set of i.i.d. sub-Gaussian random variables* (2) *with parameter $\sigma > 0$. Then there exists a constant C_θ such that*

$$\mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \geq t\right\} \leq \min\left\{1, 2n^2 e^{-\frac{t^2}{2\sigma^2}}\right\},$$
$$\mathbb{E}\left[\|\xi_n\|_{\mathcal{B}}^\theta\right] \leq C_\theta \left(\sigma\sqrt{\log n}\right)^\theta \qquad \text{for every } n > 1.$$

The first result follows from a simple union bound (Nemirovski, 1985), and the second from the first using

$$\mathbb{E}\left[\|\xi\|_{\mathcal{B}}^\theta\right] = \int_0^\infty \theta t^{\theta-1} \mathbb{P}\left\{\|\xi\|_{\mathcal{B}} \geq t\right\} dt.$$

The next result provides an interpolation inequality for the L^q-norm of a function in terms of its multiresolution norm and the norm of its k-th order derivative. For $k > d/2$, k, $d \in \mathbb{N}$ and $1 \leq q \leq \infty$, let

$$\vartheta = \vartheta(k,d,q) := \begin{cases} \dfrac{k}{2k+d} & \text{if } q \leq \dfrac{4k+2d}{d}, \\ \dfrac{k - d/2 + d/q}{2k} & \text{if } q \geq \dfrac{4k+2d}{d}. \end{cases} \tag{13}$$

Proposition 3.6. *Let $\mathcal{B}$ be a normal system of cubes, $k, d \in \mathbb{N}$, $k > d/2$ and $1 \leq q \leq \infty$. Then there exist constants C and n_0, both depending only on k and d, such that for every f and for $n \geq n_0$,*

$$\|f\|_{L^q} \leq C \max\left\{ \frac{\|S_n f\|_{\mathcal{B}}^{2\vartheta}}{n^\vartheta} \|D^k f\|_{L^2}^{1-2\vartheta},\ \frac{\|S_n f\|_{\mathcal{B}}}{n^{1/2}},\ \frac{\|D^k f\|_{L^2}}{n^{\vartheta'}} \right\}, \tag{14}$$

where $\vartheta = \vartheta(k,d,q)$ is given by (13) *and $\vartheta' := 2k\vartheta/d$.*

Remark 3.7. The inequality (14) actually holds for general $\|D^k f\|_{L^p}$ $(1 \leq p \leq \infty)$ in place of $\|D^k f\|_{L^2}$, provided that $k > d/p$ or $k \geq d$ and $p = 1$, and proper choices of ϑ, ϑ'. This is a generalization of the interpolation inequality by Nemirovski (1985). The original version holds only for normal systems with $c = 6$ (cf. Definition 3.2) and $p > d$. In fact, one can prove the general inequality (14) similarly as (Nemirovski, 1985, 2000), while replacing the use of Taylor polynomials and Vitali cover by that of averaged Taylor polynomials (Brenner and Scott, 2008, Chapter 4) and Besicovitch cover (Besicovitch, 1945, 1946). Thus, the proof is omitted here.

Note, that $k > d/p$ or $k \geq d$ and $p = 1$ is the weakest condition to ensure the continuity of f, which guarantees that the evaluation S_n and the multiresolution norm $\|\cdot\|_{\mathcal{B}}$ are well-defined. In this sense, Proposition 3.6 is already in its most general form.

4 Approximate Source Conditions

For technical simplicity, we will make the following assumption (see, however, Section 7).

Assumption 1. Every function f is defined on the d-dimensional torus $\mathbb{T}^d \sim \mathbb{R}^d/\mathbb{Z}^d$ (or equivalently *periodic*), and has *mean zero*, i.e.

$$\int_{\mathbb{T}^d} f(x)\,dx = 0.$$

We denote by $L_0^p(\mathbb{T}^d)$ with $1 \leq p \leq \infty$ the space of all L^p-functions with zero mean. We note that $L_0^p(\mathbb{T}^d)$ is a closed subspace of $L^p(\mathbb{T}^d)$ for all p. Similarly, we will denote by $H_0^k(\mathbb{T}^d) := H^k(\mathbb{T}^d) \cap L_0^2(\mathbb{T}^d)$ the space of all k-th order Sobolev functions with zero mean. We define the *homogeneous Sobolev norm* $\|D^k \cdot\|_{L^2} (=: \|\cdot\|_{H_0^k})$ as the norm in $H_0^k(\mathbb{T}^d)$.

In order to derive convergence rates for MIND, we will now introduce more recent techniques from regularization theory and inverse problems, which have not been applied in a statistical context so far, to the best of our knowledge. To that end we interpret the problem of nonparametric regression as the inverse problem of solving the equation $S_n f = y_n$ for f, where we see S_n as a mapping from $H_0^k(\mathbb{T}^d)$ to $\mathbb{R}^{\Gamma_n}$, see also (Bissantz et al., 2007). If $k > d/2$, which we always assume, it follows from the Sobolev embedding theorem (see e.g. Adams and Fournier, 2003, Theorem 4.12) that $H_0^k(\mathbb{T}^d)$ is continuously embedded in the space of all continuous functions, which in turn implies that the mapping S_n is bounded. Typical conditions in regularization theory that allow the derivation of estimates of the quality of the reconstruction in dependence of the actually realized noise level on y_n are so called *source conditions*. In this setting, they would usually be formulated as the condition that $f = S_n^* \omega$ for some *source element* $\omega \in \mathbb{R}^{\Gamma_n}$, where $S_n^* \colon \mathbb{R}^{\Gamma_n} \to H_0^k(\mathbb{T}^d)$ denotes the adjoint of the sampling operator S_n with respect to the norm on $H_0^k(\mathbb{T}^d)$ (see Groetsch, 1984; Engl et al., 1996).

Such an assumption, however, is quite restrictive in this setting; for instance, for $d = 1$, it basically implies that the function f is a spline. Therefore we use a different, but related, approach based on *approximate source conditions* (see Hofmann and Yamamoto, 2005; Hofmann, 2006). Here, the idea is to measure how well the function f can be approximated by functions of the form $S_n^* \omega$ for approximate source elements ω of given norm $t \geq 0$; we thus obtain a function $d \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, which measures for each $t \geq 0$ the distance between f and the image of the ball of radius t under S_n^*. In (Hofmann and Yamamoto, 2005), this function d has been called *distance function*. Its asymptotic properties, as the deterministic "noise level" goes to zero, have been used to obtain convergence rates for the solution of inverse problem.

In order to apply this approach to nonparametric regression using the multiresolution norm, we have to consider two refinements. First, we are interested in the asymptotics as $n \to \infty$, which means that the operator S_n we are considering changes as well. Therefore, we will have to regard instead of a single distance function a whole family of distance functions $d_n \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, one for each possible grid size. Second, since we are measuring the defect of the solution not with respect to the usual Euclidean norm but rather with respect to the multiresolution norm, we have to measure the approximation quality of an approximate source element in terms of the dual multiresolution norm (see Hein, 2008, for a similar argumentation in the case of Banach space regularization). This complicates the theory considerably, since the (dual) multiresolution norm is neither uniformly smooth nor uniformly convex.

4.1 Distance Function for Multiscale Regression

We start by considering the dual $\|\cdot\|_{\mathcal{B}^*}$ of the multiresolution norm on $\mathbb{R}^{\Gamma_n}$ with respect to the set of cubes $\mathcal{B}$. This norm is defined as

$$\|\omega\|_{\mathcal{B}^*} = \max\left\{ \sum_{x \in \Gamma_n} \omega(x) v(x) : v \in \mathbb{R}^{\Gamma_n},\ \|v\|_{\mathcal{B}} \leq 1 \right\}.$$

From the definition of the multiresolution norm in (12) it readily follows that for proper real numbers $(c_B)_{B \in \mathcal{B}}$

$$\|\omega\|_{\mathcal{B}^*} = \min\Big\{ \sum_{B \in \mathcal{B}} |c_B| \sqrt{n(B)} : \omega(x) = \sum_{B \ni x} c_B \text{ for all } x \in \Gamma_n \Big\}. \tag{15}$$

Next note that, since $S_n \colon H_0^k(\mathbb{T}^d) \to \mathbb{R}^{\Gamma_n}$ is bounded linear, it has an adjoint $S_n^* \colon \mathbb{R}^{\Gamma_n} \to H_0^k(\mathbb{T}^d)$, which is defined by the equation

$$\sum_{x \in \Gamma_n} f(x)\omega(x) = \langle f, S_n^* \omega \rangle_{H_0^k} = \langle D^k f, D^k S_n^* \omega \rangle_{L^2} = \int_{\mathbb{T}^d} D^k f \, D^k S_n^* \omega \, dx.$$

Definition 4.1. The *multiscale distance function* for f is defined as

$$d_n(t) := \min_{\|\omega\|_{\mathcal{B}^*} \leq t} \|D^k S_n^* \omega - D^k f\|_{L^2} = \min_{\|\omega\|_{\mathcal{B}^*} \leq t} \|S_n^* \omega - f\|_{H_0^k}.$$

Thus the function $d_n(t)$ measures the distance between f and the image of the ball of radius t with respect to $\|\cdot\|_{\mathcal{B}^*}$ under the mapping S_n^*. Put differently, it describes how well the function f can be approximated (with respect to the homogeneous k-th order Sobolev norm, $\|D^k \cdot\|_{L^2} \equiv \|\cdot\|_{H_0^k}$) by functions in the range of S_n^*.

In the following we will provide some description of the mapping S_n^*. To that end we denote for every $x \in \Gamma_n$ by $e_x \in \mathbb{R}^{\Gamma_n}$ the standard basis vector at x defined by

$$e_x(z) = \begin{cases} 1 & \text{if } z = x, \\ 0 & \text{else.} \end{cases}$$

Moreover we define

$$\varphi_x := S_n^* e_x.$$

Then we have for every $f \in H_0^k(\mathbb{T}^d)$ the equality

$$f(x) = \int_{\mathbb{T}^d} D^k u \, D^k \varphi_x \, dy.$$

Now let $f \in H^k(\mathbb{T}^d)$ be arbitrary. Then $f - \int_{\mathbb{T}^d} f \, dz \in H_0^k(\mathbb{T}^d)$ and therefore,

$$f(x) - \int_{\mathbb{T}^d} f \, dz = \int_{\mathbb{T}^d} D^k f \, D^k \varphi_x \, dz = (-1)^k \int_{\mathbb{T}^d} f \, \Delta^k \varphi_x \, dz = \langle f, (-1)^k \Delta^k \varphi_x \rangle_{L^2}$$

for every $f \in H^k(\mathbb{T}^d)$. Since $f(x) = \langle f, \delta_x \rangle$, we obtain that $\varphi_x = S_n^* e_x$ is the unique weak solution in $H_0^k(\mathbb{T}^d)$ of the equation[1]

$$(-1)^k \Delta^k \varphi_x = \delta_x - 1.$$

Moreover we have for general $\omega \in \mathbb{R}^{\Gamma_n}$, $\omega = \sum_{x \in \Gamma_n} \omega_x e_x$, the representation

$$S_n^* \omega = \sum_{x \in \Gamma_n} \omega_x \varphi_x.$$

Then the definition of $d_n(t)$ implies that

$$d_n(t) = \min\Big\{ \|f - \sum_{x \in \Gamma_n} c_x \varphi_x\|_{H_0^k} : \|(c_x)_{x \in \Gamma_n}\|_{\mathcal{B}^*} \le t \Big\}. \tag{16}$$

Because of the definition of the dual multiresolution norm, we can further rewrite this by introducing the functions

$$\varphi_B := \sum_{x \in B \cap \Gamma_n} \varphi_x \qquad \text{for } B \in \mathcal{B}.$$

We then obtain the representation

$$d_n(t) = \min\Big\{ \|f - \sum_{B \in \mathcal{B}} c_B \varphi_B\|_{H_0^k} : \sum_{B \in \mathcal{B}} |c_B| \sqrt{n(B)} \le t \Big\}. \tag{17}$$

4.2 Abstract Convergence Rates

We consider now the MIND estimator $\hat{f}_{\gamma_n}$, which is defined as the solution of the optimization problem given in (9). Our first result provides an estimate of the accuracy of MIND, measured in terms of an L^q-norm, under the assumption that the multiresolution norm of the error is bounded by γ_n. While the result is purely deterministic, it immediately allows for the derivation of almost sure convergence rates by adapting the parameter γ_n to the number of measurements.

Theorem 4.2. *Let k, $d \in \mathbb{N}$, $k > d/2$ and $1 \le q \le \infty$. Assume that $\mathcal{B}$ is normal and the inequality*

$$\|\xi_n\| = \|S_n f - y_n\|_{\mathcal{B}} \le \gamma_n$$

[1] There is a solution of series form:

$$\varphi_x(z) = \sum_{\lambda \in \mathbb{Z}^d \setminus \{0\}} (2\pi \|\lambda\|)^{-2k} e^{2\pi i \lambda \cdot (z-x)} = \sum_{\lambda \in \mathbb{N}_0^d \setminus \{0\}} 2 (2\pi \|\lambda\|)^{-2k} \cos\left(2\pi \lambda \cdot (z-x)\right).$$

is satisfied, and denote by $\hat{f}_{\gamma_n}$ *the MIND estimator* (9)*. In addition, define*

$$c_n := \min_{t\geq 0}\big(d_n(t) + (\gamma_n t)^{1/2}\big).$$

Then there exist constants $C > 0$ *and* $n_0 \in \mathbb{N}$*, both depending only on* k *and* d*, such that*

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} \leq C \max\Big\{\frac{\gamma_n^{2\vartheta} c_n^{1-2\vartheta}}{n^{\vartheta}},\ \frac{\gamma_n}{n^{1/2}},\ \frac{c_n}{n^{\vartheta'}}\Big\} \quad \text{for } n \geq n_0, \tag{18}$$

where $\vartheta = \vartheta(k,d,q)$ *is given by* (13) *and* $\vartheta' := 2k\vartheta/d$*.*

Proof. See Appendix A. □

As a direct consequence of the previous result and the fact that the multiresolution norm of independent sub-Gaussian noise with high probability increases at most logarithmically, we obtain an asymptotic convergence rate almost surely for MIND with properly chosen γ_n, i.e.

$$\gamma_n = C(\log n)^r \quad \text{for some } r \geq \frac{1}{2} \text{ and } C > \begin{cases} 0 & \text{if } r > \frac{1}{2}, \\ \sigma\sqrt{5 + \frac{2k}{d}} & \text{if } r = \frac{1}{2}. \end{cases} \tag{19}$$

We emphasize that such choice of γ_n is universal, in the sense that it is independent of the smoothness of the truth f, and the system of cubes $\mathcal{B}$. In particular, when $r > 1/2$, γ_n depends on n only.

Corollary 4.3. *Assume that* $\mathcal{B}$ *is normal, that* γ_n *is chosen as in* (19)*, and that*

$$\min_{t\geq 0}\big(d_n(t) + (\log n)^{r/2} t^{1/2}\big) = \mathcal{O}\left(n^{-\mu}\right)$$

for some $0 \leq \mu < 1/2$*. Then there exists a constant* C *such that the MIND estimator* $\hat{f}_{\gamma_n}$ *satisfies the estimate*

$$\limsup_{n\to\infty} \left(n^{\mu(1-2\vartheta)+\vartheta}(\log n)^{-2r\vartheta}\|\hat{f}_{\gamma_n} - f\|_{L^q}\right) \leq C \quad a.s. \tag{20}$$

with ϑ *given in* (13)*.*

Proof. With the given choice of γ_n, Proposition 3.5 implies that

$$\mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} > \gamma_n\right\} \to 0 \qquad \text{as } n \to \infty.$$

As a consequence, the probability that the estimate in Theorem 4.2 applies tends to 1 as $n \to \infty$. Noting that, for n sufficiently large and $0 \leq \mu < 1/2$, the first term on the right hand side of (18) is always dominant, we obtain (20). □

Moreover, we obtain under the same assumptions also the same convergence rate in expectation. The proof of this result, however, is more involved, because it requires an estimate for the error $\|\hat{f}_{\gamma_n} - f\|_{L^q}$ in the high noise case $\|\xi_n\|_{\mathcal{B}} > \gamma_n$, in which case the estimate from Theorem 4.2 does not apply. Thus it is relegated to the appendix.

Theorem 4.4. *Assume the setting of Theorem 4.2 and Corollary 4.3. Then the MIND estimator* $\hat{f}_{\gamma_n}$ *satisfies*

$$\mathbb{E}\left[\|\hat{f}_{\gamma_n} - f\|_{L^q}\right] = \mathcal{O}\left(n^{-\mu(1-2\vartheta)-\vartheta}(\log n)^{2r\vartheta}\right) \tag{21}$$

as $n \to \infty$*, with* ϑ *given in* (13)*.*

Proof. See Appendix B. □

Remark 4.5. We note that the inequality

$$c_n = \min_{t\geq 0}\big(d_n(t) + (\log n)^{r/2}t^{1/2}\big) \leq d_n(0) = \|D^k f\|_{L^2}$$

always holds. Under the assumption that

$$f \in W_0^{k,p}(\mathbb{T}^d) \qquad \text{for some } p \in [2,\infty],$$

we therefore always obtain with the parameter choice given in Corollary 4.3 and Theorem 4.4, respectively, for MIND

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} = \mathcal{O}\left(n^{-\vartheta}(\log n)^{2r\vartheta}\right),$$

almost surely and in expectation.

We conclude this section by emphasizing that the introduced multiscale distance functions $d_n(t)$ transform the convergence analysis of MIND into the study of approximation property of the bases $\big(\varphi_x\big)_{x\in\Gamma_n}$, or the frames $\big(\varphi_B\big)_{B\in\mathcal{B}}$, with the size of coefficients controlled in certain sense, see (16) and (17). In one dimension, we are able to derive sharp error bounds for such approximation problem (see Section 5), using the well-developed theory of B-splines, see the next section. However, in higher dimensions, this approximation problem (16) or (17) remains still open. Note that there exist some results on the approximation error of bases $\big(\varphi_x\big)_{x\in\Gamma_n}$ (see Dyn et al., 1999; Narcowich et al., 2002, 2003), but we are not aware of any results about the size of the coefficients.

5 Convergence Rates for $d = 1$

As a first step, we show that the range of the adjoint S_n^* of the sampling operator consists basically of splines. Moreover, it is possible to obtain estimates for the dual multiresolution norm of splines provided that the system of intervals on which the multiresolution norm is based is sufficiently regular. The desired approximate source conditions follow then from approximation results for splines. In the following, we will introduce the necessary notation and state our main theorems, while the major proofs are, again, postponed to the appendix.

5.1 Notation

We denote the forward and backward differences of a function $f\colon \mathbb{T} \to \mathbb{R}$ by

$$D_{h,+}f(\cdot) = f(\cdot + h) - f(\cdot), \text{ and } D_{h,-}f(\cdot) = f(\cdot) - f(\cdot - h) \text{ with } h \in \mathbb{R},$$

and that of a sequence $\{a_i\}_{0\leq i\leq n-1}$ by

$$(D_+a)_i = a_{i+1} - a_i, \text{ and } (D_-a)_i = a_i - a_{i-1}.$$

Here we define $(D_+a)_{n-1} = a_0 - a_{n-1}$ and $(D_-a)_0 = a_0 - a_{n-1}$. We note that the adjoints of these mappings are given, respectively, by

$$D_{h,+}^* = -D_{h,-} \text{ and } D_+^* = -D_-.$$

In the following, we will give a brief introduction to Besov spaces, as it is required here, and refer the interested readers to (Triebel, 1983, 1992, 1995; Adams and Fournier, 2003) for further details. First we define the r-th *modulus of smoothness* of f in $L^p(\mathbb{T})$, $1 \le p \le \infty$, as

$$\varpi_r(f;t)_p := \sup_{0\le|h|\le t} \|D_{h,+}^r f\|_{L^p}.$$

Based on it, we define the *Besov norm* $\|\cdot\|_{B_{p,0}^{s,p'}}$, with $s > 0, 1 \le p, p' \le \infty$, as

$$\|f\|_{B_{p,0}^{s,p'}} := \|f\|_{L^p} + |f|_{s,p,p',r},$$

where $s < r$, $r \in \mathbb{N}$, and

$$|f|_{s,p,p',r} := \begin{cases} \left(\int_{\mathbb{T}} (t^{-s}\varpi_r(f;t)_p)^{p'} \frac{dt}{t}\right)^{1/p'} & \text{if } 1 \le p' < \infty \\ \operatorname{ess\,sup}_{t>0} t^{-s}\varpi_r(f;t)_p & \text{if } p' = \infty. \end{cases}$$

The *Besov space* $B_{p,0}^{s,p'}(\mathbb{T})$ is then defined as the Banach space consisting of functions with bounded Besov norm, that is,

$$B_{p,0}^{s,p'}(\mathbb{T}) := \{f \in L_0^p(T); \|f\|_{B_{p,0}^{s,p'}} < \infty\}.$$

An equivalent definition of Besov spaces is based on interpolation theory of Banach spaces, for instance using the K-method, see e.g. (Triebel, 1995). Note that the Sobolev space $W_0^{s,p}(\mathbb{T}^d)$ equals the Besov space $B_{p,0}^{s,p}(\mathbb{T}^d)$ for every non-integer $s \in (0,\infty)$, but $W_0^{k,p}(\mathbb{T}^d) \ne B_{p,0}^{k,p}(\mathbb{T}^d)$ for $k \in \mathbb{N}$.

Given $m \in \mathbb{N}$, denote by $\mathcal{P}_m$ the space of polynomials of order m (i.e. of degree $\le m-1$), that is,

$$\mathcal{P}_m := \Big\{\sum_{i=1}^m a_i x^{i-1} : a_i \in \mathbb{R},\, i = 1,\dots,m\Big\}.$$

Now assume that $\Gamma \subset \mathbb{T}$ is a discrete subset. The space of piecewise polynomials of order m on $\mathbb{T}$ with knots in Γ is defined by

$$\mathcal{PP}_m(\Gamma;\mathbb{T}) := \Big\{p\colon \mathbb{T}\to\mathbb{R} : \text{for all } (x,y) \subset \mathbb{T}\setminus\Gamma \text{ there exists } q \in \mathcal{P}_m \text{ s.t. } p(t) = q(t) \text{ for all } t \in (x,y)\Big\}.$$

Then we define the space of m-order splines on $\mathbb{T}$ with simple knots in Γ as

$$\mathcal{S}_m(\Gamma;\mathbb{T}) := \mathcal{PP}_m(\Gamma;\mathbb{T}) \cap \mathcal{C}^{m-2}(\mathbb{T}).$$

Let $Q_0^m \in \mathcal{S}_m(\Gamma_n;\mathbb{T})$ be given by

$$Q_0^m(x) := \frac{n^{m-1}}{(m-1)!}\sum_{i=0}^m (-1)^i \binom{m}{i}\Big(x - \frac{i}{n}\Big)_+^{m-1} \quad \text{for } x \in [0,1),$$

where $(x)_+ := \max\{x, 0\}$. Then $\{Q_i^m(x) := Q_0^m(x - i/n) : i = 0,\dots,n-1\}$ forms a basis of $\mathcal{S}_m(\Gamma_n;\mathbb{T})$, which is called the basis of *normalized B-splines*. More details can be found in (Wahba, 1990; Schumaker, 2007) for example.

5.2 Dual operator and splines

For each $i = 0, 1, \dots, n-1$, we denote by $\varphi_{i,n}$ the unique weak solution of the differential equation

$$(-1)^k \varphi_{i,n}^{(2k)} = \delta\Big(\frac{i}{n} - \cdot\Big) - 1,\ \varphi_{i,n} \in H_0^k(\mathbb{T}). \tag{22}$$

As demonstrated in Section 4.1, it follows that $S_n^* e_{i/n} = \varphi_{i,n}$. We will show in the following that the span of the functions $\varphi_{i,n}$ in particular contains the space of all splines of order $2k$ on Γ_n with zero mean.

To that end, let us first define $\chi_n \in L^2(\mathbb{T})$ by

$$\chi_n(y) := \begin{cases} 1, & \text{if } y \in [0, \frac{1}{n}), \\ 0, & \text{if } y \in [\frac{1}{n}, 1). \end{cases}$$

By integrating both sides of (22) and respecting the zero mean we obtain

$$(-1)^{k-1} \varphi_{i,n}^{(2k-1)}(z) = \begin{cases} z - \frac{i}{n} + \frac{1}{2}, & \text{if } 0 \le z < \frac{i}{n}, \\ z - \frac{i}{n} - \frac{1}{2}, & \text{if } \frac{i}{n} \le z < 1. \end{cases}$$

Therefore

$$(-1)^k D_{\frac{1}{n},-} \varphi_{i,n}^{(2k-1)}(z) = \chi_n\Big(z - \frac{i}{n}\Big) - \frac{1}{n}.$$

Repeating this procedure m times (with $m \le 2k$), we see that

$$(-1)^k D_{\frac{1}{n},-}^m \varphi_{i,n}^{(2k-m)}(z) = (\chi_n *^{m-1} \chi_n)\Big(z - \frac{i}{n}\Big) - \frac{1}{n^m}.$$

As a consequence, it follows that

$$\psi_{i,n}^m(z) \ := \ (-1)^k n^{m-1} D_{\frac{1}{n},-}^m \varphi_{i,n}^{(2k-m)}(z) \ = \ n^{m-1}(\chi_n *^{m-1} \chi_n)(z - \frac{i}{n}) - \frac{1}{n} \ = \ Q_i^m(z) - \frac{1}{n} \tag{23}$$

is the L^2-projection of the normalized B-spline Q_i^m onto $L_0^2(\mathbb{T})$. We do note here that the functions $\psi_{i,n}^m$ are not linearly independent, their sum being zero.

Now assume that

$$h = \sum_{i=0}^{n-1} \tilde{c}_i \psi_{i,n}^m$$

for some coefficients $\tilde{c}_i \in \mathbb{R}$. Noting that

$$D_{1/n,-} \varphi_{i,n}^{(k)}(z) = \varphi_{i,n}^{(k)}(z) - \varphi_{i,n}^{(k)}\Big(z - \frac{1}{n}\Big) = \varphi_{i,n}^{(k)}(z) - \varphi_{i+1,n}^{(k)}(z),$$

we see that

$$h = (-1)^k n^{m-1} \sum_{i=0}^{n-1} \tilde{c}_i D_{\frac{1}{n},-}^m \varphi_{i,n}^{(2k-m)}$$

$$=(-1)^k n^{m-1} \sum_{i=0}^{n-1} \tilde{c}_i (D^{m-1}_{\frac{1}{n},-} \varphi^{(2k-m)}_{i,n} - D^{m-1}_{\frac{1}{n},-} \varphi^{(2k-m)}_{i+1,n})$$
$$=(-1)^k n^{m-1} \sum_{i=0}^{n-1} (D_- \tilde{c})_i D^{m-1}_{\frac{1}{n},-} \varphi^{(2k-m)}_{i,n}.$$

Repeating this argumentation m times, we obtain

$$h = (-1)^k n^{m-1} \sum_{i=0}^{n-1} \tilde{c}_i D^m_{\frac{1}{n},-} \varphi^{(2k-m)}_{i,n} = (-1)^k n^{m-1} \sum_{i=0}^{n-1} (D^m_- \tilde{c})_i \varphi^{(2k-m)}_{i,n}.$$

This shows that, indeed, the span of the functions $\psi^m_{i,n}$ is contained in the span of the functions $\varphi^{(2k-m)}_{i,n}$, and that the change of coefficients with respect to the different spanning sets is given by the linear mapping $\tilde{c} \mapsto (-1)^k n^{m-1} D^m_- \tilde{c}$.

Remark 5.1. It is still interesting to know that

$$\varphi_{i,n}(x) \equiv (-1)^{k-1} B_{2k}(x - \frac{i}{n}),$$

with B_{2k} the Bernoulli polynomial (see e.g. Kress, 1998, Section 9.4),

$$B_{2k}(x) := 2(-1)^{k-1} \sum_{l=1}^{\infty} \frac{\cos(2\pi l x)}{(2\pi l)^{2k}},$$

although this fact is not needed in this paper. One can easily see this by means of Fourier series.

5.3 Convergence Rates

We now derive the main results of this paper, where we prove convergence rates in the one-dimensional case for f contained in various Sobolev and Besov spaces.

Our first main result in the one-dimensional setting is concerned with the high regularity situation, where the function f actually is of higher smoothness than assumed by the regularization term $\|D^k \hat{f}\|^2_{L^2}$. In this case, it turns out that indeed a higher order convergence rate is obtained than the one discussed in Remark 4.5. For this to hold, however, we have to assume that the system of intervals $\mathcal{B}$ is regular (see Definition 3.3), which implies its normality. The proof of this result, mainly postponed to the appendix, relies on estimates for the multiscale distance function d_n, which in turn follow from various approximation results with splines.

Proposition 5.2. *Assume that $d = 1$, $r \geq 1/2$, that $\mathcal{B}$ is regular, and that*

$$f \in B^{k+s,p'}_{p,0}(\mathbb{T})$$

for some $1 \leq s \leq k$ and $1 \leq p, p' \leq \infty$. Then

$$\min_{t \geq 0} \big(d_n(t) + (\log n)^{r/2} t^{1/2} \big) = \mathcal{O}\left(n^{-\mu} (\log n)^{2r\mu} \right)$$

with

$$\mu = \frac{s - \left(\frac{1}{p} - \frac{1}{2}\right)_+}{2s + 2k + 1 - 2\left(\frac{1}{p} - \frac{1}{2}\right)_+}. \tag{24}$$

The same result holds for $f \in W_0^{k+s,p}(\mathbb{T})$ with $1 \leq s \leq k$ and $1 \leq p \leq \infty$.

Proof. See Appendix C. □

Theorem 5.3. *Assume that $d = 1$, that $\mathcal{B}$ is regular, and that*

$$f \in B_{p,0}^{k+s,p'}(\mathbb{T})$$

for some $1 \leq s \leq k$ and $1 \leq p, p' \leq \infty$. Then the MIND estimator $\hat{f}_{\gamma_n}$ satisfies, with a parameter choice γ_n given by (19),

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} = \mathcal{O}\left(n^{-\mu(1-2\vartheta)-\vartheta}(\log n)^{2r\mu(1-2\vartheta)+2r\vartheta}\right) \qquad \text{as } n \to \infty,$$

a.s. and in expectation, with ϑ in (13) and μ in (24). The same result holds for $f \in W_0^{k+s,p}(\mathbb{T})$ with $1 \leq s \leq k$ and $1 \leq p \leq \infty$.

Proof. This is a direct consequence of Proposition 5.2, Corollary 4.3, and Theorem 4.4 □

Remark 5.4. Note that the rate obtained in the previous result greatly simplifies in the case where $p \geq 2$ and $q \leq 4k + 2$. Then, a short computation shows that it can be written as

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} = \mathcal{O}\left(n^{-\frac{k+s}{2k+2s+1}}(\log n)^{\frac{2k+2s}{2k+2s+1}r}\right) \qquad \text{as } n \to \infty.$$

In the one-dimensional case, it is also possible to obtain convergence rates in the case where the regularity of the function f is overestimated by the regularization term. In this case, the approach based on the multiscale distance function does not readily apply, because it is inherently based on the assumption that $f \in H_0^k(\mathbb{T})$. Instead, it is possible to approximate f by a sufficiently regular function, for which then the higher order results can be applied. The final convergence rate then follows from a combination of these higher order rates and the approximation error.

Theorem 5.5 (Over-smoothing). *Let $\mathcal{B}$ be normal, $d = 1$, $k \in \mathbb{N}$, $1 \leq q \leq 4k + 2$ and*

$$f \in W_0^{s,\infty}(\mathbb{T}) \text{ or } B_{\infty,0}^{s,p'}(\mathbb{T}) \qquad \text{with } s \in [1, k] \text{ and } p' \in [1, \infty].$$

Let also $\hat{f}_{\gamma_n}$ be the MIND estimator by (9) with the homogeneous Sobolev norm of order k, $\|D^k \cdot\|_{L^2}$, and the threshold γ_n in (19). Then it holds that,

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} = \mathcal{O}\left(n^{-s/(2s+1)}(\log n)^{\epsilon + s/(2s+1)}\right) \qquad \text{as } n \to \infty,$$

a.s. and in expectation, for any $\epsilon > \frac{(2r-1)k}{2k+1}$ with r in (19).

Proof. See Appendix D. □

Remark 5.6. For simplicity, the convergence rates results of Theorems 5.3 and 5.5 were only given in $\mathcal{O}$ notation. However, it is worth pointing out that the proofs, if followed closely, actually also provide the constants in these rates. Most importantly, one can show that the constant only depends on the norm of f in the corresponding Besov or Sobolev space. More precisely, it can be shown that the constant can, in the Besov space case, be written in the form $C\|f\|^{1-2\vartheta}_{B^{k+s,p'}_{p,0}}$ with $C > 0$ only depending on k, s, p, p', and q, and the analogous result holds for the Sobolev space case. As we will see in the next subsection, this observation leads to the partial adaptation property of the MIND estimator, in minimax sense.

5.4 Comparison with Minimax Rates

Given a class $\mathcal{F}$ of continuous functions, we define the minimax L^q-risk of nonparametric regression (1) over $\mathcal{F}$ by

$$\mathcal{R}_q(n;\mathcal{F}) := \inf\Big\{\sup_{f\in\mathcal{F}} \mathbb{E}\left[\|\hat{f} - f\|_{L^q}\right] : \hat{f} \text{ is an estimator}\Big\}.$$

In other words, we measure for each estimator, the maximal expected error over all functions $f \in \mathcal{F}$, and then compute the infimum of this maximal error over the class of all estimators.

In the case of $\mathcal{F}$ consisting of Sobolev or Besov functions of a certain regularity, it is possible to derive explicit lower bounds for the minimax risk $\mathcal{R}_q$. To that end, we introduce, for $s \geq 0$, $1 \leq p \leq \infty$, and $L > 0$ the Sobolev ball

$$S^{s,p}_L := \left\{f \in W^{s,p}_0(\mathbb{T}) : \|f\|_{W^{k,p}_0} \leq L\right\}, \tag{25}$$

and for $s \geq 0$, $1 \leq p, p' \leq \infty$, and $L > 0$ the Besov ball

$$B^{s,p,p'}_L := \left\{f \in B^{s,p'}_{p,0}(\mathbb{T}) : \|f\|_{B^{k,p'}_{p,0}} \leq L\right\}. \tag{26}$$

In (Nemirovski, 1985) it has been shown that, for $s \in \mathbb{N}$, and n sufficiently large, there exists a constant $C > 0$ depending only on s such that

$$\mathcal{R}_q(n; S^{s,p}_L) \geq C \begin{cases} \left(\frac{\sigma^2}{n}\right)^{\beta} L^{1-2\beta} & \text{if } q < (2s+1)p, \\ \left(\frac{\sigma^2 \log n}{n}\right)^{\beta} L^{1-2\beta} & \text{if } q \geq (2s+1)p, \end{cases} \tag{27}$$

where

$$\beta := \begin{cases} \frac{s}{2s+1} & \text{if } q < (2s+1)p, \\ \frac{s-1/p+1/q}{2s+1-2/p} & \text{if } q \geq (2s+1)p. \end{cases}$$

Following the proof in (Nemirovski, 1985), one can show that this result still holds for non-integer $s > 1/p$, and also for all the Besov balls $B^{s,p,p'}_L$ with $s > 1/p$. Even more, in the case of $q = (2s+1)p$, the lower bound can be tightened to include the logarithmic factor $(\log n)^{(1-p/\min\{p,p'\})_+/q}$, see (Donoho et al., 1996, Theorem 1) for details.

5.4.1 Partial adaptation

Comparing these minimax L^q-risks with the convergence rates of MIND in Theorems 5.3 and 5.5, and Remark 4.5, we see that, for $1 \leq q \leq 4k+2$, the polynomial part of our rates coincides with the

polynomial part of the minimax risk in case either the function f is contained in the Sobolev space $W_0^{s,p}(\mathbb{T})$ with either $1 \leq s \leq k$ and $p = \infty$, $s = k$ and $2 \leq p \leq \infty$, or $k+1 \leq s \leq 2k$ and $p \geq 2$ (see Figure 2). In other words, in all of these cases, the convergence rates we obtain with MIND are optimal up to a logarithmic factor.

We want to stress here that our convergence rates do not rely on a precise knowledge of the smoothness class of the function f. In contrast, the regularization parameter γ_n does only depend on the sample size, and the smoothing order of the regularization term need only be a rough guess of the actual smoothness of f. Neither in the case where the smoothness of f is overestimated nor in the case where it is slightly underestimated do we obtain results that are, asymptotically, far from being optimal. The method MIND automatically adapts to the smoothness of the function f independent of our prior guess.

Note that the adaptation range of MIND scales with the smoothness order of regularization k. This suggests that we should choose k as large as possible. The minimization problem in (9) becomes, however, more numerically unstable as k increases. Thus, the choice of k should balance the performance and the numerical stability. In practice, we found that it works fine for $k = 1, 2, 3$ (cf. Section 6.3).

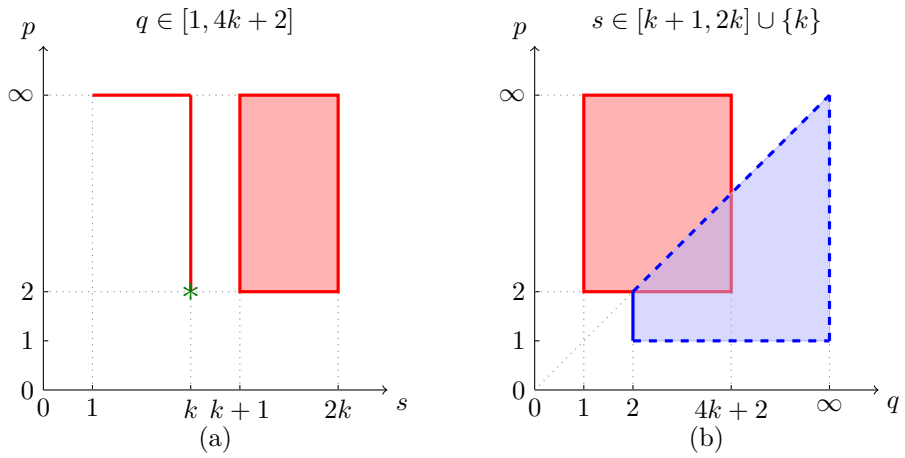

Figure 2: Adaptive minimax optimality over balls in $W_0^{s,p}(\mathbb{T})$ or $B_{p,0}^{s,p'}(\mathbb{T})$. (a) For $q \in [1, 4k+2]$, the MIND estimator with homogeneous Sobolev norm of order k attains minimax optimal rates in terms of L^q-risk up to a log-factor, simultaneously in all spaces $W_0^{s,p}(\mathbb{T})$ with smoothness parameters s, p within the red region ("partial adaptation"). By contrast, the Nemirovski's estimator $\hat{f}_{2,\eta}$ in (5) is minimax optimal up to a log-factor only for $W_0^{k,2}(\mathbb{T})$, marked by a green asterisk. (b) For $s \in [k+1, 2k] \cup \{k\}$, MIND is optimal up to a log-factor in terms of L^q-risk over $S_L^{s,p}$ or $B_L^{s,p,p'}$ with parameters p, q within the red region. Note that no linear estimator is optimal for parameters p, q in the blue region.

6 Numerical Examples

The MIND estimator defined by (9) is the solution to a high dimensional non-smooth convex optimization problem, due to the multiresolution norm. As mentioned in the introduction, there are several efficient algorithms nowadays that are able to tackle such a problem. In this paper, we have chosen the alternating

direction method of multipliers (ADMM). It decomposes the original problem (9) into two sub-problems, the first being the smoothing penalization problem, and the other the projection onto the multiresolution ball. The latter one can be solved by the Dykstra algorithm as introduced in (Boyle and Dykstra, 1986). We refer to (Frick et al., 2012) for further details. It can be shown that the ADMM for this problem converges linearly (cf. Deng and Yin, 2015), although the theoretical rate might be very slow for large k, the smoothing order of the regularization. We note that the problem (9) can also be formulated as a quadratic programming problem and solved, for instance, by the interior point method (cf. Nesterov et al., 1994, for example).

There are two practical concerns: The first is the choice of the system of cubes $\mathcal{B}$. The general convergence rate results (cf. Theorem 4.2) only require that the system should be normal, see Remark 3.4 for some examples. In the case of $d = 1$, the concrete rates impose an additional (but very weak) condition, namely that it should contain an m-partition system for some m, i.e., m-regularity of $\mathcal{B}$, see Theorem 5.3. For the examples in our numerical simulations, we found that the system of all cubes, the system of all cubes with dyadic edge lengths, and the 2-partition system perform comparably. Therefore we display the results for the 2-partition system in the following numerical experiments for the sake of computational efficiency.

The other concern is the choice of the threshold γ_n. The asymptotic theory only requires that γ_n satisfies the condition (19), which is independent of the interval system $\mathcal{B}$, and the smoothness of the truth. In the finite sample situation, we recommend a refined choice, which has a direct statistical interpretation, cf. Section 1.1. It selects γ_n as the α-quantile of the multiscale statistic, i.e.,

$$\gamma_n(\alpha) := \inf \left\{ \gamma : \mathbb{P} \left\{ \|\xi_n\|_{\mathcal{B}} > \gamma \right\} \leq \alpha \right\}. \tag{28}$$

This ensures that f lies in the confidence set defined by the multiscale constraint in (9) with probability at least $1 - \alpha$. Thus we have

$$\mathbb{P} \left\{ \|D^k \hat{f}_{\gamma_n}\|_{L^2} \leq \|D^k f\|_{L^2} \right\} \geq 1 - \alpha, \tag{29}$$

that is, the MIND estimator is smoother than the truth with probability at least $1 - \alpha$. The asymptotic distribution of $\|\xi_n\|_{\mathcal{B}}$ is, under general assumptions, a Gumbel law (after proper rescaling), see (Kabluchko, 2011; Haltmeier and Munk, 2013). If $\mathcal{B}$ consists of all the cubes and ξ_n is standard Gaussian, then

$$\gamma_n(\alpha) \sim \sqrt{2d \log n} + \frac{\log(d \log n) + \log J_d - 2 \log \log(1/\alpha)}{2\sqrt{2d \log n}} \quad \text{as } n \to \infty,$$

where $J_d \in (0, \infty)$ is a constant. Although this violates the condition (19) when $d = 1$, the asymptotic analysis in this paper still holds for $\gamma_n(\alpha_n)$ if $\alpha_n \to 0$ sufficiently fast, which might even possibly improve the rates, in terms of the log-factor and the constant. The estimation of γ_n can be done by Monte-Carlo simulations and is needed only once for a fixed size of measurements and a fixed system of cubes. The number of Monte-Carlo simulations is chosen as 10^5 in the following examples.

In the following simulations, we only consider the one-dimensional case $d = 1$, and assume the noise is i.i.d. Gaussian with a known variance σ^2. In practice, one can easily pre-estimate σ^2, see e.g. (Rice, 1984; Hall et al., 1990; Dette et al., 1998) among other references.

6.1 Comparison study

We now investigate the performance of MIND $\hat{f}_{\gamma_n(\alpha)}$ on spatially variable functions, Bumps, HeaviSine, and Doppler (Donoho and Johnstone, 1994), and compare it with the smoothing spline estimator (SS)

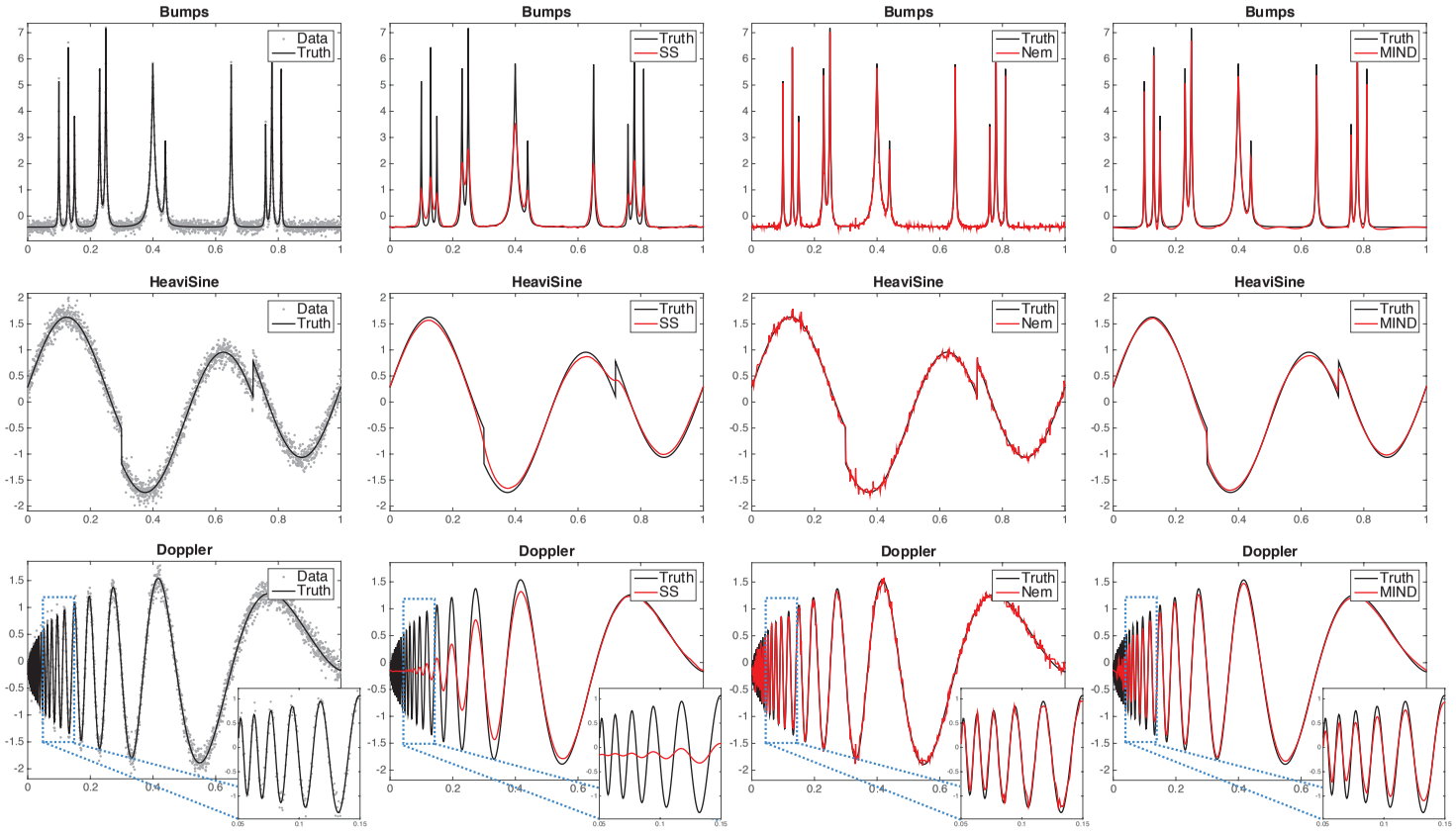

Figure 3: Comparison of Nem in (5), SS in (30), and the MIND in (9) (number of samples $n = 2^{11}$, noise level $\sigma = 0.12\|f\|_{L^2}$).

$\hat{f}_\lambda$, defined as the solution of

$$\min_f \|S_n f - y_n\|_{\ell^2} + \lambda \|D^k f\|_{L^2}, \tag{30}$$

and the Nemirovski's estimator (Nem) $\hat{f}_{2,\eta}$ in (5) with $p = 2$. We choose $k = 1$ for all three estimators. The parameter α in MIND is set to 0.1, λ in SS is tuned manually to give the best visual quality, and η in Nem is chosen as the oracle $\|Df\|_{L^2} (=: \eta_0)$, which is numerically estimated using discrete Fourier transform.

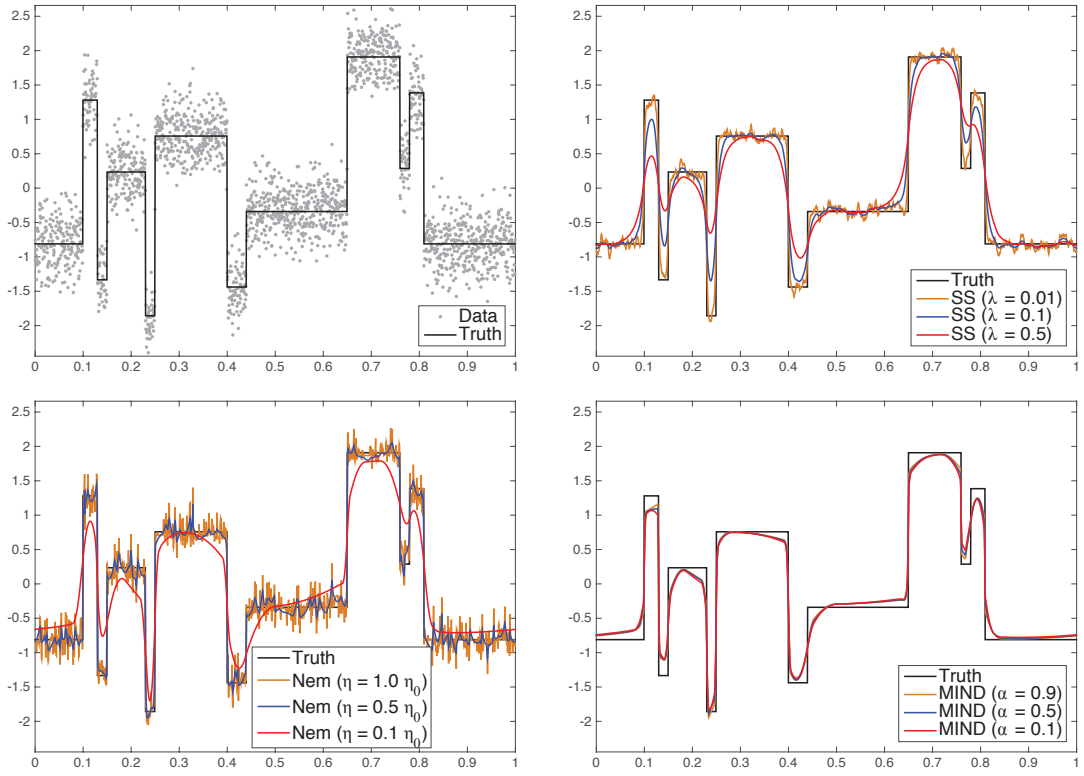

Figure 4: Impact of parameter choices: various η for Nem $\hat{f}_{2,\eta}$ in (5), various λ for SS $\hat{f}_\lambda$ in (30), and various α for the MIND $\hat{f}_{\gamma_n}$ in (9) with $\gamma_n = \gamma_n(\alpha)$ in (28). (number of samples $n = 2^{11}$, noise level $\sigma = 0.3\|f\|_{L^2}$, and $\eta_0 = \|D^k f\|_{L^2}$).

The simulation results are summarized in Figure 3. One can see that MIND detects a large number of features at various scales of smoothness, and performs best on all the test signals. By contrast, SS with the "optimal" parameter recovers only a narrow range of scales of smoothness; for instance, on the Doppler signal, it works well for the smoother part (on $[0.5, 1]$), but deteriorates fast as the signal gets more oscillatory. The Nem with oracle $\eta (= \eta_0)$ is still very noisy on each test signal. We note that convex duality (cf. Section 1.1) implies that there is a one-to-one correspondence between MIND and Nem as long as the different parameters are not unreasonably large. The Nem will reproduce the results by MIND if we choose as the threshold η, $0.8\eta_0$ for Bumps, $0.3\eta_0$ for HeaviSine, and $0.6\eta_0$ for Doppler.

This means that, even if $\eta_0 \equiv \|Df\|_{L^2}$ is known exactly, one cannot find a universal threshold η for Nem, which explains our numerical findings.

6.2 Robustness and stability

We examine the robustness of Nem, SS, and MIND, with respect to the choice of parameters, and the smoothness assumption, on the Blocks signal (Donoho and Johnstone, 1994), which is not even continuous, and hence falls not into the domain of our estimator. From Figure 4 we find that the MIND estimator is rather robust to the choice of significance level α, while Nem and SS are much more sensitive. Besides, MIND recovers the truth quite well with the correct number of local extrema, and slight distortion near change-points. As we already noted before, the performance SS is restricted to some fixed scale of smoothness. In contrast, Nem with a proper choice of threshold η adapts to a wider range of smoothness scales, which is due to its relation to MIND via duality. Thus, this study again confirms that MIND is practically preferable over Nem and SS.

Now, we continue to consider the impact of significance level on the performance of the MIND estimator. Exemplarily, we choose Bumps as the test signal for different noise levels. In Figure 5, it shows that MIND with various choices of significance levels perform almost identically well in the case of low noise level ($\sigma = 0.1$) and medium noise level ($\sigma = 0.5$). However, in the high noise level ($\sigma = 1.2$) case, MIND with larger α tends to detect more bumps. For example, MIND ($\alpha = 0.9$) recovers 6 more bumps than MIND ($\alpha = 0.1$), four out of which are actually correct (marked by vertical blue lines), while 2 false bumps are detected (marked by vertical red dashed lines, in the bottom panel of Figure 5). Recall that the significance level α can be interpreted as an error control in the sense of (29). Thus, the additional power by an increased significance level comes at the expense of a lower confidence about the inference.

Note additionally that all the test signals considered so far are not strictly periodic, so the simulations also reveal that MIND is not too sensitive to the periodicity assumption. In practice, one can extend a non-periodic function to a periodic one by symmetric extension, see for instance (Mallat, 2009).

6.3 Choice of smoothness order

Now we explore the choice of smoothness parameter k in the regularization term for the MIND estimator. The Doppler with symmetric extension (see Figure 6) is chosen as the test signal. The significance level for MIND is set to $\alpha = 0.1$. Figure 6 shows that MIND detects more features of different smoothness scales as k increases, namely 24 peaks for $k = 1$, 26 for $k = 2$, and 27 for $k = 3$. Meanwhile, the height of the peaks gets more accurate for larger k. This is in accordance with our theoretical finding that the adaptation range increases with k, see Section 5.4.1. As already mentioned, one should, however, notice that the optimization problem becomes numerically more ill-conditioned as k increases.

7 Discussion

In this paper, we have introduced a constrained variational estimator, MIND, which minimizes the L^2-norm of the k-th order derivatives, k being the anticipated smoothness of the function to be recovered, subject to the constraint that the multiresolution norm of the residual is bounded by some parameter γ_n depending on the sample size n. The idea behind this approach is that this norm effectively allows to differentiate between smooth functions and noise, as the multiresolution norm of a continuous function

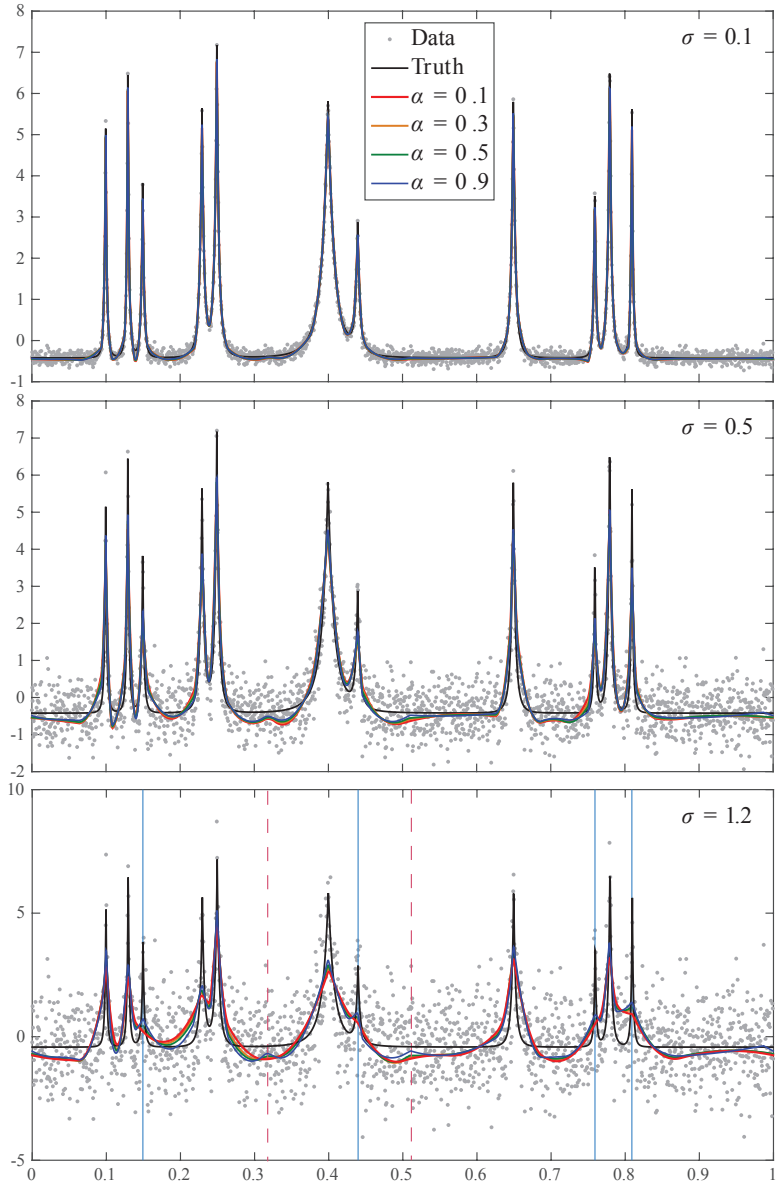

Figure 5: Stability of MIND in significance level α and noise level σ. The reconstructions by MIND $\hat{f}_{\gamma_n}$ with $\gamma_n = \gamma_n(\alpha)$ for a range of α's are shown, together with the true signal and noisy data, in the cases of different noise levels (number of samples $n = 2^{11}$).

is of the order of $\sqrt{n}$, while the expected multiresolution norm of a sample of independent sub-Gaussian noise is of the order of $\sqrt{\log n}$. If we therefore use a threshold parameter $\gamma_n \sim (\log n)^r$ with $r > 1/2$

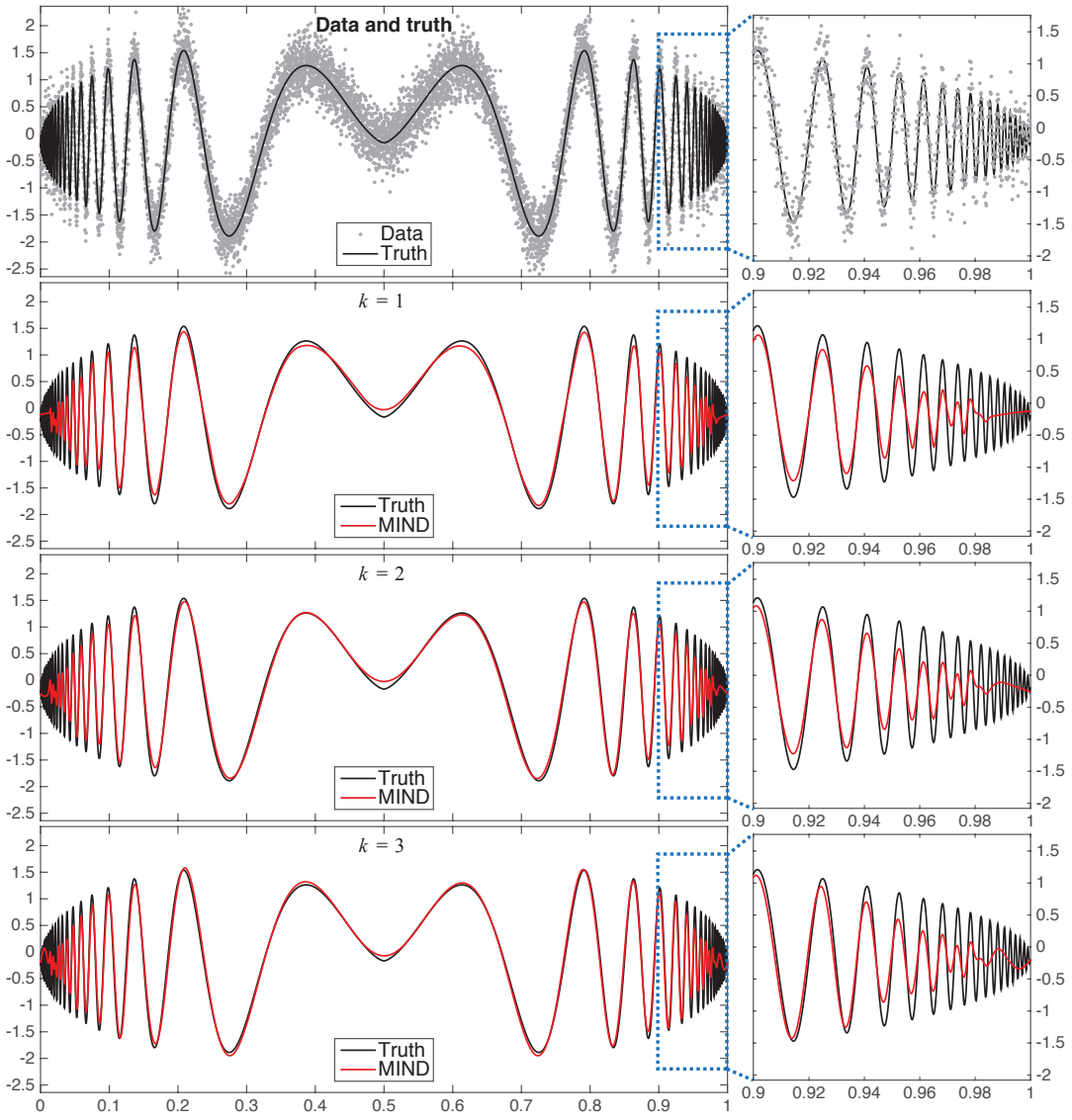

Figure 6: Various choices of k for MIND in (9) (number of samples $n = 2^{13}$, noise level $\sigma = 0.3\|f\|_{L^2}$).

we can expect that, for a sufficiently large sample size, our estimator, MIND, will be close to the true function, while the residuals consist mostly of noise.

The main theoretical contribution of this paper was to underpin the already known empirically good performance of MIND in several special cases by some theoretical evidence. For general dimension d, from an interpolation inequality for the multiresolution norm and Sobolev norms (Proposition 3.6), we derive asymptotic convergence rates provided that $f \in H_0^k(\mathbb{T}^d)$ and one applies regularization with the

homogeneous H^k-norm (see Remark 4.5). Moreover, these rates turn out to be minimax optimal up to a logarithmic factor. In order to derive convergence rates for different smoothness classes, we have adapted the concept of approximate source conditions, to our statistical setting. These are known to be a useful tool for the derivation of rates for deterministic inverse problems. However, these conditions are quite abstract, and it is not immediately clear how they relate to more tangible properties of f.

In the one-dimensional setting, a much more detailed analysis is possible. Here the abstract conditions for convergence rates can be related to approximation properties of splines. Mainly we have shown that the rates depend on how well the k-th derivative of the function f can be approximated by B-splines with coefficients that are small with respect to the dual multiresolution norm. Using results from approximation theory, we were able to translate the approximate source conditions into very general smoothness conditions for the function f. Mainly this gives us optimal convergence rates for a function $f \in H_0^s(\mathbb{T})$ with $k+1 \leq s \leq 2k$. More general, we have obtained with this argumentation convergence rates for functions f contained in the fractional order Sobolev space $W_0^{s,p}(\mathbb{T})$ with $k+1 \leq s \leq 2k$, and the rates are again optimal as long as $p \geq 2$. Moreover, the same results hold for comparable Besov spaces. While these results are only concerned with functions f that are of higher regularity than assumed a-priori, it is also possible to derive rates for the case where f is of lower regularity, that is, where the prior assumption that $f \in H_0^k(\mathbb{T})$ fails. The idea here is to approximate f by a spline of higher regularity and then to apply the higher order convergence rate results to this spline. The final rate then results from a trade-off between the approximation power of the spline and the higher order convergence rate. With this technique, one obtains optimal convergence rate for the lower order setting $f \in W_0^{s,\infty}(\mathbb{T})$ with $1 \leq s \leq k$.

It is important to note here that the choice of the parameter γ_n is independent of the actual smoothness of f. This is why MIND yields (up to a logarithmic factor) simultaneously optimal convergence rates for a range of smoothness classes (with smoothness order $s \in [1,k] \cup [k+1,2k]$), making it truly an adaptive method. Additionally, the numerical results indicate that MIND appears to be fairly robust with respect to the actual choice of the parameter γ_n for a given sample size n, further enhancing its practical applicability.

There are several questions still open concerning MIND for nonparametric regression. First of all, almost all concrete results concerning convergence rates in this paper were derived for a one-dimensional setting. In higher dimensions we only have the (somehow generic) result mentioned in Remark 4.5 that gives us an optimal convergence rate if our guess for the smoothness class of f is correct. It is, however, not at all obvious how to obtain rates for higher order smoothness classes. In the one-dimensional case, the method we used relied both on approximation results using B-splines and on estimates for the dual multiresolution norm of the coefficients of these B-splines. In higher dimensions, we expect that similar results for polyharmonic splines would be required, but it is not clear which basis splines have to be used (cf. (16) and (17)). Moreover, in the literature, there is few results on the size of approximation coefficients, which are necessary for our analysis. Similarly, the method we have used for the derivation of the lower order convergence rates relies intrinsically on spline approximation, which, again, makes the generalization to higher dimensions difficult.

Also in the one-dimensional case there are several interesting open questions. Our results only apply to a periodic setting with functions that have zero mean. The main reason for the restriction to periodic functions is that this avoids having to deal with boundary conditions that would have to be taken into account in non-periodic cases. The restriction to functions with zero mean on the other hand is to simplify norms of Sobolev spaces. Dropping the requirement of zero mean is possible; for instance, one can instead use $\hat{f}^0_{\gamma_n} + \bar{y}_n$, with $\hat{f}^0_{\gamma_n}$ the MIND (9) applied to $(y_n - \bar{y}_n)$. The same convergence rates still hold true by the analysis presented in this paper, and by the fact that $\bar{y}_n$ converges to the mean of the

truth, $\int_{[0,1)} f(x)dx$, at a parametric rate, $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$. Most importantly, we are concerned with the gap between $H_0^k(\mathbb{T})$ and $H_0^{k+1}(\mathbb{T})$. It seems reasonable to assume that MIND is asymptotically optimal also for functions in $H_0^s(\mathbb{T})$ with $k < s < k+1$, but the methods we have used for the derivation of the different rates appear not to be applicable to this case. Also, while we do obtain convergence rates for functions $f \in W_0^{s,p}(\mathbb{T})$ with $p < 2$, these rates are not optimal. Here we suspect that this is due to the fact that we use the L^2-norm of the k-th order derivative for regularization and that better rates could be obtained by using the L^1-norm instead.

In our future work, we will try to extend our results to the cases mentioned above, in particular to higher dimensional and non-periodic settings. Additionally, we plan to consider a generalization to the solution of ill-posed operator equations. In particular for deconvolution problems we expect that very similar results can be obtained as in this paper, as long as the convolution kernel has sufficiently slowly decreasing Fourier coefficients.

A Proof of Theorem 4.2

The assumption $\|S_n f - y_n\|_{\mathcal{B}} \leq \gamma_n$ implies that f is admissible for the minimization problem (9), which in turn implies that

$$\frac{1}{2}\|D^k \hat{f}_{\gamma_n}\|_{L^2}^2 \leq \frac{1}{2}\|D^k f\|_{L^2}^2.$$

As a consequence, we obtain the estimate

$$\begin{aligned}
\frac{1}{2}\|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2}^2 &= \frac{1}{2}\|D^k \hat{f}_{\gamma_n}\|_{L^2}^2 - \frac{1}{2}\|D^k f\|_{L^2}^2 - \langle f, \hat{f}_{\gamma_n} - f\rangle_{H_0^k} \\
&\leq -\langle f, \hat{f}_{\gamma_n} - f\rangle_{H_0^k} \\
&= \min_t \min_{\|\omega\|_{\mathcal{B}^*} \leq t} \Big(\langle S_n^*\omega - f, \hat{f}_{\gamma_n} - f\rangle_{H_0^k} - \langle S_n^*\omega, \hat{f}_{\gamma_n} - f\rangle_{H_0^k} \Big) \\
&\leq \min_t \min_{\|\omega\|_{\mathcal{B}^*} \leq t} \Big(\|D^k S_n^*\omega - D^k f\|_{L^2} \|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2} \\
&\qquad + \|\omega\|_{\mathcal{B}^*} \|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}} \Big) \\
&\leq \min_t \Big(d_n(t) \|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2} + t \|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}} \Big).
\end{aligned}$$

Thus we have for every $t \geq 0$ the inequality

$$\frac{1}{2}\|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2}^2 \leq d_n(t) \|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2} + t \|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}}.$$

Since

$$\|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}} \leq 2\gamma_n,$$

we obtain the inequality

$$\begin{aligned}
\|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2} &\leq d_n(t) + \sqrt{d_n(t)^2 + 2t\|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}}} \\
&\leq 2d_n(t) + (2t)^{1/2} \|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}}^{1/2} \\
&\leq 2d_n(t) + 2(\gamma_n t)^{1/2}
\end{aligned} \tag{31}$$

for every $t \geq 0$. We now recall the interpolation inequality (see Proposition 3.6)

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} \leq C \max\Bigg\{ \frac{\|S_n(\hat{f}_{\gamma_n} - f)\|_{\mathcal{B}}^{2\vartheta}}{n^{\vartheta}} \|D^k(\hat{f}_{\gamma_n} - f)\|_{L^2}^{1-2\vartheta}, \\ \frac{\|S_n(\hat{f}_{\gamma_n} - f)\|_{\mathcal{B}}}{n^{1/2}}, \frac{\|D^k(\hat{f}_{\gamma_n} - f)\|_{L^2}}{n^{\vartheta'}} \Bigg\} \quad (32)$$

for n sufficiently large, with some $C > 0$ and $\vartheta' = 2k\vartheta/d$.

If the maximum in (32) is attained at the first term, the estimate (31) implies that

$$\begin{aligned} \|\hat{f}_{\gamma_n} - f\|_{L^q} &\leq C \frac{\|S_n \hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}}^{2\vartheta}}{n^{\vartheta}} \|D^k \hat{f}_{\gamma_n} - D^k f\|_{L^2}^{1-2\vartheta} \\ &\leq C \frac{(2\gamma_n)^{2\vartheta}}{n^{\vartheta}} \min_t \big(2d_n(t) + 2(\gamma_n t)^{1/2}\big)^{1-2\vartheta} \\ &\leq 2C \frac{\gamma_n^{2\vartheta}}{n^{\vartheta}} \min_t \big(d_n(t) + (\gamma_n t)^{1/2}\big)^{1-2\vartheta}. \end{aligned}$$

On the other hand, if the maximum in (32) is attained at the second term, we have

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} \leq C \frac{\|S_n(\hat{f}_{\gamma_n} - f)\|_{\mathcal{B}}}{n^{1/2}} \leq 2C \frac{\gamma_n}{n^{1/2}}.$$

Finally, if the maximum in (32) is attained at the third term, we have

$$\begin{aligned} \|\hat{f}_{\gamma_n} - f\|_{L^q} \leq C \frac{\|D^k(\hat{f}_{\gamma_n} - f)\|_{L^2}}{n^{\vartheta'}} &\leq C \min_t \frac{2d_n(t) + 2(\gamma_n t)^{1/2}}{n^{\vartheta'}} \\ &\leq 2C \frac{1}{n^{\vartheta'}} \min_t \big(d_n(t) + (\gamma_n t)^{1/2}\big). \end{aligned}$$

□

B Proof of Theorem 4.4

Denote in the following

$$p(t) := \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \leq t\right\}.$$

Using Theorem 4.2, we see that we can estimate, for n sufficiently large,

$$\mathbb{E}\left[\|\hat{f}_{\gamma_n} - f\|_{L^q}\right] \leq \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \leq \gamma_n\right\} C n^{-\mu(1-2\vartheta)-\vartheta} (\log n)^{r(1+2\vartheta)/2} \\ + \int_{\gamma_n}^{\infty} \sup\left\{\|\hat{f}_{\gamma_n} - f\|_{L^q} : \|\xi_n\|_{\mathcal{B}} = t\right\} dp(t). \quad (33)$$

In the following, we will show that the second term on the right hand side of (33) tends to zero faster as $n \to \infty$. To that end, we observe first that the Sobolev embedding theorem (Adams and Fournier, 2003, Theorem 4.12) and the Poincaré inequality (Ziemer, 1989, Theorem 4.4.2) imply that

$$\|\hat{f}_{\gamma_n} - f\|_{L^q} \leq \|\hat{f}_{\gamma_n}\|_{L^q} + \|f\|_{L^q} \leq C\|D^k \hat{f}_{\gamma_n}\|_{L^2} + \|f\|_{L^\infty}$$

for some constant C depending only on q, d and k. Moreover, by construction, we have

$$\|D^k \hat{f}_{\gamma_n}\|_{L^2} \leq \|D^k g\|_{L^2} \text{ for all } g \text{ satisfying } \|S_n g - S_n f - \xi_n\|_{\mathcal{B}} \leq \gamma_n.$$

Now let $h \in H^k(\mathbb{R}^d)$ be such that $h(0) = 1$, $\int_{\mathbb{R}^d} h(x)\,dx = 0$, and $\operatorname{supp} h \subset [-1/2, 1/2]^d$. Define moreover, for $n \in \mathbb{N}$ and $x \in \Gamma_n$, the function $h_{n,x} \colon \mathbb{T}^d \to \mathbb{R}$ by

$$h_{n,x}(y) = h(n^{1/d}(y - x)).$$

Let now n and $\xi_n \in \mathbb{R}^{\Gamma_n}$ be fixed and define

$$g := \sum_{x \in \Gamma_n} (f(x) + \xi_n(x)) h_{n,x}.$$

Since the functions $h_{n,x}$, $x \in \Gamma_n$, have pairwise disjoint supports, it follows that

$$\begin{aligned}\|D^k g\|_{L^2} &= \sum_{x \in \Gamma_n} |f(x) + \xi_n(x)| \|D^k h_{n,x}\|_{L^2} = \sum_{x \in \Gamma_n} |f(x) + \xi_n(x)| n^{\frac{2k-d}{2d}} \|D^k h\|_{L^2} \\ &= n^{\frac{2k-d}{2d}} \|S_n f + \xi_n\|_{\ell^1} \|D^k h\|_{L^2} \leq n^{\frac{2k+d}{2d}} \big(\|f\|_{L^\infty} + \|\xi_n\|_{L^\infty}\big) \|D^k h\|_{L^2}.\end{aligned}$$

From the inequality $\|\xi_n\|_{L^\infty} \leq \|\xi_n\|_{\mathcal{B}}$ we thus obtain that, for some constant C only depending on q, d, and k,

$$\sup\Big\{\|\hat{f}_{\gamma_n} - f\|_{L^q} : \|\xi_n\|_{\mathcal{B}} = t\Big\} \leq C n^{\frac{2k+d}{2d}} (\|f\|_{L^\infty} + t).$$

As a consequence, we can estimate the last term in (33) by

$$\begin{aligned}\int_{\gamma_n}^{\infty} \sup\Big\{\|\hat{f}_{\gamma_n} - f\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t\Big\}\, dp(t) &\leq \int_{\gamma_n}^{\infty} C n^{\frac{2k+d}{2d}} (\|f\|_{L^\infty} + t)\, dp(t) \\ &= C n^{\frac{2k+d}{2d}} (\|f\|_{L^\infty} + \gamma_n)(1 - p(\gamma_n)) - C n^{\frac{2k+d}{2d}} \int_{\gamma_n}^{\infty} (p(t) - 1)\, dt.\end{aligned}$$

From Proposition 3.5 we obtain that

$$(1 - p(t)) \leq 2n^2 e^{-\frac{t^2}{2\sigma^2}}$$

for sufficiently large n. Thus we see that

$$\begin{aligned}&\int_{\gamma_n}^{\infty} \sup\Big\{\|\hat{f}_{\gamma_n} - f\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t\Big\}\, dp(t) \\ \leq\, &2C n^{\frac{2k+5d}{2d}} (\|f\|_{L^\infty} + \gamma_n) e^{-\frac{\gamma_n^2}{2\sigma^2}} + 2C n^{\frac{2k+5d}{2d}} \int_{\gamma_n}^{\infty} e^{-\frac{t^2}{2\sigma^2}}\, dt \leq C' n^{\frac{2k+5d}{2d}} \gamma_n e^{-\frac{\gamma_n^2}{2\sigma^2}}\end{aligned}$$

for sufficiently large n. Now the choice of γ_n implies that

$$n^{\frac{2k+5d}{2d}} \gamma_n e^{-\frac{\gamma_n^2}{2\sigma^2}} = \mathcal{O}(n^{-\varepsilon})$$

as $n \to \infty$ for some $\varepsilon > 0$. This shows that the second term in (33) tends to zero faster as $n \to \infty$, which concludes the proof of Theorem 4.4. □

C Proof of Proposition 5.2

The main idea of the proof of Proposition 5.2 is to approximate the function f with splines, the coefficients of which are (relatively) small with respect to the dual multiresolution norm. As a preparation, we will need several results concerning approximation properties of splines, most of which are well known in approximation theory, and a result that allows us to bound the dual multiresolution norm of a spline function.

C.1 Approximation properties of splines

Proposition C.1 (Condition number of B-splines). *Assume $\{Q_i^m(x), i = 0, \dots, n-1\}$ is the family of normalized B-splines in $\mathcal{S}_m(\Gamma_n; \mathbb{T})$. Then for any $c_i \in \mathbb{R}, i = 0, \dots, n-1$,*

$$\|(c)_{i=0}^{n-1}\|_p \le m2^m n^{1/p} \|\sum_{i=0}^{n-1} c_i Q_i^m\|_{L^p} \quad \text{for } 1 \le p \le \infty. \tag{34}$$

Proof. Let us first consider $1 \le p < \infty$. By $\{\tilde{Q}_i^m\}_{i=-m+1}^{ln-1}$ we denote the normalized B-splines on the real line with equally spaced knots $\{(-m+1)/n, (-m+2)/n, \dots, (ln+m-1)/n\}$. Let $\tilde{c}_i := c_{i \bmod n}$ for $i = -m+1, \dots, ln-1$. It is known from (Scherer and Shadrin, 1999, Theorem 1) that

$$\|(\tilde{c}_i)_{i=-m+1}^{ln-1}\|_p \le m2^m n^{1/p} \|\sum_{i=-m+1}^{ln-1} \tilde{c}_i \tilde{Q}_i^m\|_{L^p} \quad \text{for any } l \in \mathbb{N}.$$

It implies

$$l\|(c_i)_{i=0}^{n-1}\|_p^p + \|(c_i)_{i=n-m+1}^{n-1}\|_p^p \le n(m2^m)^p \left(l\|\sum_{i=0}^{n-1} c_i Q_i^m\|_{L^p}^p + \|\sum_{i=n-m+1}^{n-1} Q_i^m \mathbf{1}_{[0,\frac{m-1}{n})\cup[\frac{n-m+1}{n},1)}\|_{L^p}^p \right)$$

or

$$\|(c_i)_{i=0}^{n-1}\|_p^p + \frac{1}{l}\|(c_i)_{i=n-m+1}^{n-1}\|_p^p \le n(m2^m)^p \left(\|\sum_{i=0}^{n-1} c_i Q_i^m\|_{L^p}^p + \frac{1}{l}\|\sum_{i=n-m+1}^{n-1} Q_i^m \mathbf{1}_{[0,\frac{m-1}{n})\cup[\frac{n-m+1}{n},1)}\|_{L^p}^p \right).$$

By letting $l \to \infty$, we obtain (34) for $1 \le p < \infty$. Then, the case $p = \infty$ follows by taking $p \to \infty$. □

Remark C.2. This result is a generalization of a known result for splines on $\mathbb{R}$ (see Scherer and Shadrin, 1999) to periodic splines.

Proposition C.3 (Boundedness of L^2-projector). *Let $P_{\mathcal{S}}$ be the orthogonal projector onto $\mathcal{S}_m(\Gamma_n; \mathbb{T})$ in the topology of $L^2(\mathbb{T})$. Then there is a constant C depending only on m such that*

$$\|P_{\mathcal{S}} u\|_{L^p} \le C\|u\|_{L^p} \quad \text{for any } u \in L^p(\mathbb{T}) \text{ and } 1 \le p \le \infty.$$

Proof. By (Adams and Fournier, 2003, Theorem 2.11) and (Schumaker, 2007, Theorem 6.30), it is sufficient to prove this assertion only for $p = 1$ and $p = \infty$.

Consider first the case $p = \infty$. Let $Q_i^m \in \mathcal{S}_m(\Gamma_n; \mathbb{T})$ be the normalized B-splines, and $R_i^m := nQ_i^m$ implying that $\|R_i^m\|_{L^1} = 1$. If $P_{\mathcal{S}} f = \sum_{i=0}^{n-1} a_i Q_i^m$, then

$$\sum_{j=0}^{n-1} a_j \langle Q_j^m, R_i^m \rangle = \langle f, R_i^m \rangle,$$

that is, we have an equation of the form $Ga = b$ with $a := (a_i)_{i=0}^{n-1}$, $b := (\langle f, R_i^m \rangle)_{i=0}^{n-1}$ and $G := (\langle R_i^m, Q_j^m \rangle)_{i,j}$. Note that

$$\|b\|_\infty = \max_i |\langle f, R_i^m \rangle| \le \max_i \|f\|_{L^\infty} \|R_i^m\|_{L^1} = \|f\|_{L^\infty}.$$

This implies that

$$\|P_{\mathcal{S}} f\|_{L^\infty} = \Big\| \sum_{i=0}^{n-1} a_i Q_i^m \Big\|_{L^\infty} \le \|a\|_\infty \le \|G^{-1}\|_\infty \|b\|_\infty \le \|G^{-1}\|_\infty \|f\|_{L^\infty}.$$

It follows from (de Boor, 2012) that

$$\|G^{-1}\|_\infty \le C_m$$

for some constant C_m depending only on m. Thus, $\|P_{\mathcal{S}} f\|_{L^\infty} \le C_m \|f\|_{L^\infty}$.

Next consider $p = 1$. Let $P_{\mathcal{S}} f = \sum_{i=0}^{n-1} \tilde{a}_i R_i^m$, then $\sum_{j=0}^{n-1} \tilde{a}_j \langle R_j^m, Q_i^m \rangle = \langle f, Q_i^m \rangle$, i.e., $G^t \tilde{a} = \tilde{b}$, where $(\cdot)^t$ denotes transpose, $\tilde{a} := (\tilde{a}_i)_{i=0}^{n-1}$ and $\tilde{b} := (\langle f, Q_i^m \rangle)_{i=0}^{n-1}$. It follows from $\sum_i Q_i^m = 1$ and $Q_i^m \ge 0$ that

$$\|\tilde{b}\|_1 = \sum_i |\langle f, Q_i^m \rangle| \le \sum_i \big\langle |f|, Q_i^m \big\rangle = \Big\langle |f|, \sum_i Q_i^m \Big\rangle = \|f\|_{L^1}.$$

Then

$$\begin{aligned} \|P_{\mathcal{S}} f\|_{L^1} =& \Big\| \sum_{i=0}^{n-1} \tilde{a}_i R_i^m \Big\|_{L^1} \le \|\tilde{a}\|_1 \le \|G^{-t}\|_1 \|\tilde{b}\|_1 \\ =& \|G^{-1}\|_\infty \|\tilde{b}\|_1 \le \|G^{-1}\|_\infty \|f\|_{L^1} \le C_m \|f\|_{L^1}. \end{aligned}$$

That is, we obtain the assertion for $p = 1$. □

Remark C.4. The above result (of periodic splines with equally spaced knots) is probably proven in 1970s, but we are not aware of the reference. The proof we give here also shows the result for periodic splines with non-equally spaced knots, since de Boor (2012) proved the boundedness of the inverse Gram matrix of B-splines for any knots. Similar results for non-periodic splines with arbitrary knots are originally proven in (Shadrin, 2001), and recently shortened in (Golitschek, 2014).

Proposition C.5 (Approximation property). *Let $1 \le p, p', q \le \infty$. There exists a linear operator $A : L_0^1(\mathbb{T}) \to \mathcal{S}_m(\Gamma_n; \mathbb{T})$ such that for every $u \in L_0^1(\mathbb{T})$*

$$\|u - Au\|_{W_0^{r,q}} \le C_1 \frac{\|u\|_{B_{p,0}^{s,p'}}}{n^{s-r-(1/p-1/q)_+}} \quad \text{with } 1 \le s \le m,\ 0 \le r \le \lfloor s-1 \rfloor,$$

$$\|Au\|_{W_0^{r,q}} \le C_2 \frac{\|u\|_{B_{p,0}^{s,p'}}}{n^{s-r-(1/p-1/q)_+}} \quad \text{with } 1 \le s \le \lceil s \rceil \le r \le m-1,$$

where C_1, C_2 depend only on m, p. Moreover, both inequalities also hold for the Sobolev norm $\|\cdot\|_{W_0^{s,p}}$ when $p = p'$ and $s \in \mathbb{N}$.

Remark C.6. In the case of Sobolev norm $\|\cdot\|_{W_0^{s,p}}$, $s \in \mathbb{N}$, the assertions follow from (Schumaker, 2007, Theorem 8.12). Following the idea of the proof of (Schumaker, 2007, Theorem 6.31), such results can be extended to Besov norms using (Schumaker, 2007, Theorem 6.30).

Proposition C.7 (Finite differences and $W^{1,p}(\mathbb{T})$). *Let $h > 0$ and $1 \leq p \leq \infty$. Then*

$$\|D_{h,+}f\|_{L^p} = \|D_{h,-}f\|_{L^p} \leq h\|Df\|_{L^p} \quad \text{for } f \in W^{1,p}(\mathbb{T}). \tag{35}$$

Proof. The case of $p = \infty$ is obviously true. Now consider $1 \leq p < \infty$. Since $\|D_{h,+}f\|_{L^p} = \|D_{h,-}f\|_{L^p}$, it is sufficient to prove (35) only for $D_{h,+}$. Note that for each $f \in W^{1,p}(\mathbb{T})$ there is a sequence of smooth functions f_n, such that $\|D_{h,+}f_n\|_{L^p} \to \|D_{h,+}f\|_{L^p}$ and $\|Df_n\|_{L^p} \to \|Df\|_{L^p}$ as $n \to \infty$. Therefore, we assume without loss of generality that f is a smooth function. It follows from the equation $f(x+h) - f(x) = h\int_0^1 f'(x+th)dt$ that

$$\begin{aligned}
\int_0^1 |f(x+h) - f(x)|^p dx &\leq h^p \int_0^1 \Big(\int_0^1 |f'(x+th)|dt \Big)^p dx \\
&\leq h^p \int_0^1 \int_0^1 |f'(x+th)|^p dt\, dx \\
&= h^p \int_0^1 \int_0^1 |f'(x+th)|^p dx\, dt \\
&= h^p \|f'\|_{L^p}^p.
\end{aligned}$$

That is $\|D_{h,+}f\|_{L^p} \leq h\|Df\|_{L^p}$. □

C.2 Regular Systems

Next we state two technical lemmas, which allow us to estimate the dual multiresolution norm of piecewise constant vectors in case the system $\mathcal{B}$ is m-regular. These piecewise constant vectors will appear as spline coefficients for certain approximation splines needed for the proof of Proposition 5.2.

Lemma C.8. *Assume that $m \in \mathbb{N}$, $m \geq 2$, and that $n \in \mathbb{N}$ is written as*

$$n = \sum_{j=0}^{r} d_j m^j \qquad \text{with } d_j \in \{0, \ldots, m-1\}$$

and $r = \lfloor \log_m n \rfloor$. Then

$$\sum_{j=0}^{r} d_j m^{j/2} \leq (\sqrt{m} + 1)\sqrt{n}.$$

Proof. We prove this claim by induction over r. For $r = 0$ it is trivial.

Now assume that the claim holds for r and let n be such that $\lfloor \log_m n \rfloor = r+1$. Then

$$\begin{aligned}
&\Big(\sum_{j=0}^{r+1} d_j m^{j/2}\Big)^2 \\
=&\Big(\sum_{j=0}^{r} d_j m^{j/2}\Big)^2 + d_{r+1}^2 m^{r+1} + 2d_{r+1} m^{(r+1)/2} \sum_{j=0}^{r} d_j m^{j/2} \\
\le&(\sqrt{m}+1)^2 \sum_{j=0}^{r} d_j m^j + d_{r+1}^2 m^{r+1} + 2d_{r+1} m^{(r+1)/2} (\sqrt{m}+1) \Big(\sum_{j=0}^{r} d_j m^j\Big)^{1/2} \\
\le&(\sqrt{m}+1)^2 \sum_{j=0}^{r} d_j m^j + d_{r+1}^2 m^{r+1} + 2d_{r+1} m^{(r+1)/2} (\sqrt{m}+1) m^{(r+1)/2} \\
=&(\sqrt{m}+1)^2 \sum_{j=0}^{r} d_j m^j + \big(d_{r+1} + 2(\sqrt{m}+1)\big)\, d_{r+1} m^{r+1}.
\end{aligned}$$

Since $d_{r+1} \le m-1$ and $m - 1 + 2(\sqrt{m}+1) = (\sqrt{m}+1)^2$, this proves the assertion. □

Lemma C.9. *Assume that the family $\mathcal{B}$ is m-regular for some fixed $m \ge 2$. Let now $I = \{i_0, i_0+1, \dots, i_0+p-1\}/n \subset \Gamma_n$ and define $c \in \mathbb{R}^{\Gamma_n}$ by $c_i = 1$ if $i \in I$ and $c_i = 0$ if $i \notin I$. Then*

$$\|c\|_{\mathcal{B}^*} \le (\sqrt{m}+1)\sqrt{2mp}.$$

Proof. Let $r = \lceil \log_m n \rceil$. Let $\ell_- \in \mathbb{N}$ be maximal such that $\ell_- m^{-r} \le i_0/n$, and let $\ell_+ \in \mathbb{N}$ be minimal such that $\ell_+ m^{-r} > (i_0+p-1)/n$. Then

$$\ell_+ - \ell_- < \frac{m^r}{n}(p-1) + 2 < mp.$$

Now write

$$\ell_- = \sum_{j=0}^{r} d_j^- m^j \qquad \text{and} \qquad \ell_+ = \sum_{j=0}^{r} d_j^+ m^j.$$

Let moreover $0 \le s \le r-1$ be maximal such that $d_s^- < d_s^+$ and denote by $\hat{\ell}$ the minimal number of the form

$$\hat{\ell} = \hat{d}_s m^s + \sum_{j=s+1}^{r} d_j^+ m^j$$

such that $\ell_- \le \hat{\ell}$.

Next we denote by $\mathcal{B}_1$ the collection of intervals of the form

$$[\ell m^{k-r}, (\ell+1)m^{k-r}) \quad \text{where } 0 \le k \le s-1, \text{ and } \ell = \sum_{j=k+1}^{r} d_j^+ m^j + d m^k \text{ with } 0 \le d < d_k^+.$$

Similarly, we denote by $\mathcal{B}_2$ the collection of intervals of the form

$$[\ell m^{s-r}, (\ell+1)m^{s-r}) \qquad \text{where } \ell = \hat{\ell} + d m^s \text{ with } 0 \le d < d_s^+ - \hat{d}_s.$$

Then the intervals contained in $\mathcal{B}_1 \cup \mathcal{B}_2$ form a disjoint cover of $[\hat{\ell} m^{-r}, \ell_+ m^{-r})$.

Next we write

$$\hat{\ell} - \ell_- = \sum_{j=0}^{s-1} \hat{d}_j^- m^j$$

and denote by $\mathcal{B}_3$ the collection of intervals of the form

$$\hat{\ell} m^{-r} - (\ell m^{k-r}, (\ell+1) m^{k-r}] \quad \text{where } 0 \le k \le s-1, \text{ and } \ell = \sum_{j=k+1}^{s-1} \hat{d}_j^- m^j + d m^k \text{ with } 0 \le d < \hat{d}_k^-.$$

Then the intervals contained in $\mathcal{B}_3$ form a disjoint cover of $[\ell_- m^{-r}, \hat{\ell} m^{-r})$.

Note in addition that by construction all of these intervals are also contained in $\mathcal{B}$. Now denote $\hat{\mathcal{B}} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \mathcal{B}_3$ and define $c_B := 1$ if $B \in \hat{\mathcal{B}}$ and $B \cap I \neq \emptyset$ and $c_B := 0$ if $B \in \mathcal{B} \setminus \hat{\mathcal{B}}$ or $B \in \hat{\mathcal{B}}$ and $B \cap I = \emptyset$. Then $c_i = \sum_{B \ni i} c_B$ for all i and therefore

$$\|c\|_{\mathcal{B}^*} \le \sum_{B \in \mathcal{B}} |c_B| \sqrt{n(B)} \le \sum_{B \in \hat{\mathcal{B}}} \sqrt{n(B)}.$$

Now note that

$$\sqrt{n([\ell m^{k-r}, (\ell+1) m^{k-r}), n)} \le m^{k/2} \qquad \text{for all } 0 \le k \le r.$$

Therefore Lemma C.8 and the definition of $\hat{\mathcal{B}}$ imply that

$$\begin{aligned}
\|c\|_{\mathcal{B}^*} &\le \sum_{k=0}^{s-1} d_k^+ m^{k/2} + (d_s^+ - \hat{d}_s) m^{s/2} + \sum_{k=0}^{s-1} \hat{d}_k^- m^{k/2} \\
&\le (\sqrt{m}+1) \Big(\Big((d_s^+ - \hat{d}_s) m^s + \sum_{k=0}^{s-1} d_k^+ m^k \Big)^{1/2} + \Big(\sum_{k=0}^{s-1} \hat{d}_k^- m^k \Big)^{1/2} \Big) \\
&= (\sqrt{m}+1) \Big(\sqrt{\ell_+ - \hat{\ell}} + \sqrt{\hat{\ell} - \ell_-} \Big) \\
&\le (\sqrt{m}+1) \sqrt{2(\ell_+ - \ell_-)} \\
&\le (\sqrt{m}+1) \sqrt{2mp}.
\end{aligned}$$

□

Remark C.10. Note that the estimate in the previous Lemma can be improved to $\|c\|_{\mathcal{B}^*} \le (\sqrt{m}+1)\sqrt{2p}$ if n is some power of m, because in this case, with the notation of the Lemma, we have $\ell^+ - \ell^- = p$. Also we have the obvious estimate $\|c\|_{\mathcal{B}^*} \le \sqrt{p}$ in case the family $\mathcal{B}$ contains all intervals.

C.3 Main Part of the Proof

In order to estimate the multiscale distance function d_n, we need to approximate $D^k f$ by a function of the form $D^k S_n^* \omega$, where $\omega \in \mathbb{R}^{\Gamma_n}$ is small with respect to the dual multiresolution norm. We will perform this approximation in two steps: First, we will show that a spline of order $k+1$ defined on a coarser grid than Γ_n can be approximated well by a function of the form $D^k S_n^* \omega$ in such a way that the dual multiresolution norm of ω increases sufficiently slowly with the decreasing grid size (see Lemma C.11). In the second step, we then approximate $D^k f$ by a spline g of order $k+1$. Balancing the grid on which g is defined with n, then gives us the behavior of d_n claimed in Proposition 5.2.

Lemma C.11. *Let $\Gamma \subset \mathbb{T}$ be a finite set and $g \in \mathcal{S}_{k+1}(\Gamma;\mathbb{T})$ with $\int_{\mathbb{T}} g\,dx = 0$. Assume moreover that*

$$\tau_{\min} := \min\{\operatorname{dist}(x,y) : x \neq y \in \Gamma\} > \frac{2k+2}{n}$$

and that $\mathcal{B}$ is regular, denote

$$\tau_{\max} := \max\{\operatorname{dist}(x,y) : (x,y) \subset \mathbb{T} \setminus \Gamma\}$$

and let $1 \leq q \leq \infty$. Then there exists $c \in \mathbb{R}^{\Gamma_n}$ and constants C_1, $C_2 > 0$ only depending on k, q, and $\mathcal{B}$ such that

$$\begin{aligned}\|g - (S_n^* c)^{(k)}\|_{L^2} &\leq C_1 \frac{\|g^{(k)}\|_{L^2}}{n^k} \\ \|c\|_{\mathcal{B}^*} &\leq C_2 \|g^{(k)}\|_{L^q} |\Gamma|^{1-1/q} n^{1/q-1} (n\tau_{\max})^{(1/2-1/q)_+}.\end{aligned}$$

Proof. Let h be the best approximation of g in $\operatorname{span}\{\psi_{i,n}^k : i \in \Gamma_n\}$ in the L^2 sense, see (23). Then we can write

$$h = \sum_{i=0}^{n-1} \tilde{c}_i \psi_{i,n}^k$$

for some coefficients $\tilde{c}_i \in \mathbb{R}$. Because the functions $\psi_{i,n}^k$ are not linearly independent, the coefficients $\tilde{c}_i$ are not unique. It is, however, possible to choose them in such a way that

$$\sum_{i=0}^{n-1} \tilde{c}_i = 0.$$

Then

$$h = \sum_{i=0}^{n-1} \tilde{c}_i \psi_{i,n}^k = \sum_{i=0}^{n-1} \tilde{c}_i Q_i^k. \tag{36}$$

Now note that the fact that $\int_{\mathbb{T}} g\,dx = 0$ implies that h is at the same time the best approximation of g in $\mathcal{S}_k(\Gamma_n;\mathbb{T})$. Thus (36) shows that, actually, the coefficients $\tilde{c}_i$ are the coefficients of the k-th order spline that approximates g best in the L^2-sense. Thus it follows from Proposition C.5 that

$$\|h - g\|_{L^2} \leq C_1 \frac{\|g^{(k)}\|_{L^2}}{n^k}. \tag{37}$$

Now let

$$c_i := (-1)^k n^{k-1} (D_-^k \tilde{c})_i, \quad \text{for } i = 0, \dots, n-1,$$

which implies that

$$h = \sum_{i=0}^{n-1} c_i \varphi_{i,n}^{(k)} = (S_n^* c)^{(k)}.$$

We will next derive an upper bound for $\|c\|_{\mathcal{B}^*}$.

Since h is the best approximation of g within $\mathcal{S}_k(\Gamma_n;\mathbb{T})$, it follows that

$$\langle h, Q_j^k \rangle_{L^2} = \langle g, Q_j^k \rangle_{L^2}$$

for all j. Applying r-th order finite differences to these vectors, we obtain that

$$D_-^r\big((\langle h, Q_j^k\rangle_{L^2})_j\big) = D_-^r\big((\langle g, Q_j^k\rangle_{L^2})_j\big)$$

for all r. From this, we obtain that

$$\langle h, D^r_{\frac{1}{n},+} Q_j^k\rangle_{L^2} = \langle g, D^r_{\frac{1}{n},+} Q_j^k\rangle_{L^2}$$

for all j. Since $(D^r_{\frac{1}{n},+})^* = (-1)^r D^r_{\frac{1}{n},-}$, this further implies that

$$\langle D^r_{\frac{1}{n},-} h, Q_j^k\rangle_{L^2} = \langle D^r_{\frac{1}{n},-} g, Q_j^k\rangle_{L^2} \tag{38}$$

for all j and all r. Next we note that

$$D^{k+1}_{\frac{1}{n},-} h = \sum_{i=0}^{n-1} (D_-^{k+1}\tilde{c})_i Q_i^k = (-1)^k n^{1-k} \sum_{i=0}^{n-1} (D_- c)_i Q_i^k. \tag{39}$$

Now let $j/n \in \Gamma_n$ be such that $j/n \notin \Gamma + (-k/n, (k+1)/n)$ and let $x \in \operatorname{supp}(Q_j^k) = [j/n, (j+k)/n]$. Then the fact that g is a polynomial of degree k outside of Γ implies that

$$(D^{k+1}_{\frac{1}{n},-} g)(x) = 0.$$

As a consequence, we obtain from (38) with $r = k+1$ and (39) that

$$0 = \langle D^{k+1}_{\frac{1}{n},-} g, Q_j^k\rangle_{L^2} = (-1)^k n^{1-k} \sum_{i=0}^{n-1} (D_- c)_i \langle Q_i^k, Q_j^k\rangle.$$

Since this holds for every $j \in \Gamma_n$ with $j/n \notin \Gamma + (-k/n, (k+1)/n)$, it follows from the properties of B-splines that

$$(D_- c)_j = 0$$

for all j such that $j/n \notin \Gamma + (-k/n, (k+1)/n)$.

Now denote by $I \subset \Gamma_n$ the set of all points i/n for which $i/n \notin \Gamma + (-k/n, (k+1)/n)$. Then the set I consists of $|\Gamma|$ disjoint sets $I_j \subset \mathbb{T}$, $j = 1, \dots, |\Gamma|$, of subsequent grid points. The considerations above imply that for each of these sets I_j there exists $\omega_j \in \mathbb{R}$ such that $c_i = \omega_j$ for $i/n \in I_j$. Therefore Lemma C.9 implies that

$$\|c\|_{\mathcal{B}^*} \le \sum_{j=1}^{|\Gamma|} C|\omega_j| \sqrt{n(I_j, n)} + \sum_{i \notin I} |c_i| \tag{40}$$

for some constant $C > 0$ only depending on $\mathcal{B}$. Now define

$$t_i := \begin{cases} 1 & \text{if } i/n \notin I, \\ C \dfrac{1}{\sqrt{n(I_j, n)}} & \text{if } i/n \in I_j \text{ for some } j. \end{cases}$$

Then the right hand side term in (40) can also be written as a sum over all products $t_i|c_i|$, $i = 0, \dots, n-1$. Therefore

$$\|c\|_{\mathcal{B}^*} \le \sum_{i=0}^{n-1} t_i |c_i|.$$

Applying Hölder's inequality gives

$$\|c\|_{\mathcal{B}^*} \le \|c\|_q \|t\|_{q_*} = \|c\|_q \Big(|\Gamma_n \setminus I| + C^{q_*} \sum_{j=1}^{|\Gamma|} n(I_j,n)^{1-q_*/2} \Big)^{1/q_*}$$
$$\le \|c\|_q \Big(2k|\Gamma| + C^{q_*} \sum_{j=1}^{|\Gamma|} n(I_j,n)^{1-q_*/2} \Big)^{1/q_*}$$

for any $1 \le q \le \infty$ and $q_* = q/(q-1)$. Since $1 \le n(I_j,n) \le n\tau_{\max}$ for all j, this further implies that

$$\|c\|_{\mathcal{B}^*} \le \|c\|_q \big(2k|\Gamma| + C^{q_*}|\Gamma|(n\tau_{\max})^{(1-q_*/2)_+} \big)^{1/q_*}$$
$$\le C\|c\|_q |\Gamma|^{1/q_*} (n\tau_{\max})^{(1/q_*-1/2)_+} = C\|c\|_q |\Gamma|^{1-1/q} (n\tau_{\max})^{(1/2-1/q)_+}. \quad (41)$$

Now note that (38) implies that $D^k_{1/n,-}h$ is the best approximating spline in the L^2-sense of $D^k_{1/n,-}g$. Thus the definition of c and Propositions C.1, C.3 and C.7 imply that

$$\|c\|_q \;=\; n^{k-1}\|D^k_- \tilde{c}\|_q \;\le\; Cn^{k-1+1/q}\|D^k_{1/n,-}h\|_{L^q} \;\le\; Cn^{k-1+1/q}\|D^k_{1/n,-}g\|_{L^q} \;\le\; Cn^{-1+1/q}\|g^{(k)}\|_{L^q}.$$

Together with (41) this shows that

$$\|c\|_{\mathcal{B}^*} \le C\|g^{(k)}\|_{L^q} |\Gamma|^{1-1/q} n^{1/q-1} (n\tau_{\max})^{(1/2-1/q)_+}$$

for some constant $C > 0$. □

Proof (of Proposition 5.2). Assume first that $p \ge 2$. Proposition C.5 applied with $u = f^{(k)} \in B^{s,p'}_{p,0}(\mathbb{T})$, $m = k+1$, and $q = p$ implies for every $\lambda \in \mathbb{N}$ the existence of a spline $g \in \mathcal{S}_{k+1}(\Gamma_\lambda; \mathbb{T})$ such that

$$\|f^{(k)} - g\|_{L^2} \le C \frac{\|f\|_{B^{k+s,p'}_{p,0}}}{\lambda^s},$$
$$\|g^{(k)}\|_{L^p} \le C \frac{\|f\|_{B^{k+s,p'}_{p,0}}}{\lambda^{s-k}}.$$

Next we obtain from Lemma C.11 the existence of a vector $c \in \mathbb{R}^{\Gamma_n}$ such that

$$\|g - (S^*_n c)^{(k)}\|_{L^2} \le C \frac{\|g^{(k)}\|_{L^2}}{n^k},$$
$$\|c\|_{\mathcal{B}^*} \le C\|g^{(k)}\|_{L^p} \lambda^{1/2} n^{-1/2},$$

provided that λ is sufficiently large (here we use that, in the notation of the Lemma, $|\Gamma| = \lambda$ and $\tau_{\max} = 1/\lambda$). Combining these estimates, it follows that, for

$$t \ge C\|f\|_{B^{k+s,p'}_{p,0}} n^{-1/2} \lambda^{1/2-s+k}$$

we have

$$d_n(t) \le C\|f\|_{B^{k+s,p'}_{p,0}} \lambda^{-s}(1 + \lambda^k n^{-k}).$$

Choosing

$$\lambda \sim n^{1/(2s+2k+1)}(\log n)^{-2r/(2k+2s+1)},$$

we obtain that

$$\min_{t\geq 0}\big(d_n(t) + (\log n)^{r/2}t^{1/2}\big) = \mathcal{O}\big(n^{-\mu}(\log n)^{2r\mu}\big) \qquad \text{with} \qquad \mu = \frac{s}{2s+2k+1}$$

as $n \to \infty$.

Now let $p \leq 2$. Again applying Proposition C.5 with $u = f^{(k)} \in B^{s,p'}_{p,0}(\mathbb{T})$ and $m = k+1$, but now with $q = 2$ yields $g \in \mathcal{S}_{k+1}(\Gamma_\lambda;\mathbb{T})$ such that

$$\begin{aligned}\|f^{(k)} - g\|_{L^2} &\leq C\frac{\|f\|_{B^{k+s,p'}_{p,0}}}{\lambda^{s-1/p+1/2}},\\ \|g^{(k)}\|_{L^2} &\leq C\frac{\|f\|_{B^{k+s,p'}_{p,0}}}{\lambda^{s-k-1/p+1/2}},\end{aligned}$$

and we obtain, for λ sufficiently large, from Lemma C.11 the existence of $c \in \mathbb{R}^{\Gamma_n}$ with

$$\begin{aligned}\|g - (S_n^* c)^{(k)}\|_{L^2} &\leq C\frac{\|g^{(k)}\|_{L^2}}{n^k},\\ \|c\|_{\mathcal{B}^*} &\leq C\|g^{(k)}\|_{L^2}\lambda^{1/2}n^{-1/2}.\end{aligned}$$

This shows that, for

$$t \geq C\|f\|_{B^{k+s,p'}_{p,0}} n^{-1/2}\lambda^{1/p-s+k},$$

we have

$$d_n(t) \leq C\|f\|_{B^{k+s,p'}_{p,0}}\lambda^{1/p-1/2-s}(1+\lambda^k n^{-k}).$$

Choosing

$$\lambda \sim n^{1/(2k+2s+2-2/p)}(\log n)^{-2r/(2k+2k+s-2/p)},$$

we obtain that

$$\min_{t\geq 0}\big(d_n(t) + (\log n)^{r/2}t^{1/2}\big) = \mathcal{O}\big(n^{-\mu}(\log n)^{2r\mu}\big) \qquad \text{with} \qquad \mu = \frac{s-1/p+1/2}{2(s+k+1-1/p)},$$

which proves the assertion.

The above argument holds also for $f \in W_0^{k+s,p}(\mathbb{T})$ if we replace $\|\cdot\|_{B^{k+s,p'}_{p,0}}$ by $\|\cdot\|_{W_0^{k+s,p}}$. □

D Proof of Theorem 5.5

Lemma D.1. *Let $\mathcal{B}$ be a family of intervals, $f \in \mathcal{C}^1(\mathbb{T})$, and $F(t) := \int_0^t f(x)dx$. Then,*

$$\frac{\|S_n f\|_{\mathcal{B}}}{\sqrt{n}} \leq \|F\|_{W^{1/2,\infty}} + \frac{\|Df\|_{L^\infty}}{2n}.$$

Proof. Denote in the following

$$\mathcal{B}_n := \Big\{ [i/n, j/n] : \text{ there exists } B \in \mathcal{B} \text{ such that } B \cap \Gamma_n = \{i/n, \dots, j/n\} \Big\}.$$

Then

$$\begin{aligned}
\frac{\|S_n f\|_{\mathcal{B}}}{\sqrt{n}} =& \max_{[i/n,j/n]\in\mathcal{B}_n} \frac{\sqrt{n}}{\sqrt{j-i+1}} \Big| \frac{1}{n} \sum_{k=i}^{j} f\big(\frac{k}{n}\big) \Big| \\
\leq& \max_{[i/n,j/n]\in\mathcal{B}_n} \Big\{ \frac{\sqrt{n}}{\sqrt{j-i+1}} \Big| \int_{i/n}^{(j+1)/n} f(x)dx \Big| \\
&\qquad + \frac{\sqrt{n}}{\sqrt{j-i+1}} \sum_{k=i}^{j} \Big| \frac{1}{n} f\big(\frac{k}{n}\big) - \int_{k/n}^{(k+1)/n} f(x)dx \Big| \Big\} \\
\leq& \sup_{s\neq t\in\mathbb{T}} \frac{|F(t)-F(s)|}{\sqrt{t-s}} \\
&\qquad + \max_{[i/n,j/n]\in\mathcal{B}_n} \frac{\sqrt{n}}{\sqrt{j-i+1}} \sum_{k=i}^{j} \int_{k/n}^{(k+1)/n} \Big| f\big(\frac{k}{n}\big) - f(x) \Big| dx \\
\leq& \|F\|_{W^{1/2,\infty}} + \max_{[i/n,j/n]\in\mathcal{B}_n} \frac{\sqrt{n}}{\sqrt{j-i+1}} \frac{j-i+1}{2n^2} \|Df\|_{L^\infty} \\
\leq& \|F\|_{W^{1/2,\infty}} + \frac{\|Df\|_{L^\infty}}{2n}.
\end{aligned}$$

□

Proof (of Theorem 5.5). Let $\tilde{\epsilon} > 0$ be fixed and set

$$\lambda := \left\lfloor \left(\frac{n}{\log n} \right)^{\frac{1}{2s+1}} (\log n)^{\tilde{\epsilon}} \right\rfloor.$$

Let $G_\lambda(t) \in \mathcal{S}_{k+2}(\Gamma_\lambda; \mathbb{T})$ be the approximation spline of $F(t) := \int_0^t f(x)dx$ as in Proposition C.5, and $g_\lambda(t) := G'_\lambda(t)$. It follows that $g_\lambda \in H_0^k(\mathbb{T})$, and

$$\begin{aligned}
\|f - g_\lambda\|_{L^q} = \|F' - G'_\lambda\|_{L^q} &\leq C \frac{\|F\|_{W^{s+1,\infty}}}{\lambda^s} \\
&\leq C \left(\frac{\log n}{n} \right)^{\frac{s}{2s+1}} (\log n)^{-s\tilde{\epsilon}} \|f\|_{W^{s,\infty}}, \\
\|D^k g_\lambda\|_{L^2} = \|D^{k+1} G_\lambda\|_{L^2} &\leq C \lambda^{k-s} \|F\|_{W^{s+1,\infty}} \\
&\leq C \left(\frac{n}{\log n} \right)^{\frac{k-s}{2s+1}} (\log n)^{(k-s)\tilde{\epsilon}} \|f\|_{W^{s,\infty}}.
\end{aligned}$$

The second relation implies that

$$\tilde{d}_n(t) := \min_{\|w\|_{\mathcal{B}^*}\leq t} \|D^k S_n^* w - 2D^k g_\lambda\|_{L^2} \leq \tilde{d}_n(0) = 2\|D^k g_\lambda\|_{L^2}$$

$$\leq 2C\left(\frac{n}{\log n}\right)^{\frac{k-s}{2s+1}}(\log n)^{(k-s)\tilde{\epsilon}}\|f\|_{W^{s,\infty}} \quad (42)$$

for every $t \geq 0$. By Proposition C.5 and Lemma D.1, we have

$$\begin{aligned}\|S_n f - S_n g_\lambda\|_{\mathcal{B}} &\leq \sqrt{n}\|F - G_\lambda\|_{W^{1/2,\infty}} + o(1)\\ &\leq C\sqrt{n}\frac{\|f\|_{W^{s,\infty}}}{\lambda^{s+1/2}} \leq C(\log n)^{\frac{1}{2}-(s+\frac{1}{2})\tilde{\epsilon}}\|f\|_{W^{s,\infty}}.\end{aligned}$$

for sufficiently large n. Consequently

$$\begin{aligned}\|S_n g_\lambda - y_n\|_{\mathcal{B}} &\leq \|S_n g_\lambda - S_n f\|_{\mathcal{B}} + \|S_n f - y_n\|_{\mathcal{B}}\\ &\leq C(\log n)^{\frac{1}{2}-(s+\frac{1}{2})\tilde{\epsilon}}\|f\|_{W^{s,\infty}} + \|\xi_n\|_{\mathcal{B}}.\end{aligned} \quad (43)$$

Set $\tilde{\gamma}_n := C_0\sqrt{\log n} < \gamma_n$ with some $C_0 > \sigma\sqrt{5 + 2k/d}$. Then $\|\xi_n\|_{\mathcal{B}} \leq \tilde{\gamma}_n$, together with (43), implies that $\|S_n g_\lambda - y_n\|_{\mathcal{B}} \leq \gamma_n$ for large enough n. In such case we can apply Theorem 4.2, but with f replaced by its approximation g_λ, and obtain the estimate

$$\|\hat{f}_{\gamma_n} - g_\lambda\|_{L^q} \leq C\max\left\{\frac{\gamma_n^{2\vartheta}}{n^\vartheta}\min_{t\geq 0}\left(\tilde{d}_n(t) + (\gamma_n t)^{1/2}\right)^{1-2\vartheta}, \frac{\gamma_n}{n^{1/2}}, \frac{1}{n^{\vartheta'}}\min_{t\geq 0}\left(\tilde{d}_n(t) + (\gamma_n t)^{1/2}\right)\right\},$$

with $\vartheta = k/(2k+1)$, $\vartheta' = 2k^2/(2k+1)$. By (42), this further implies that, for sufficiently large n, the estimate

$$\begin{aligned}&\|\hat{f}_{\gamma_n} - g_\lambda\|_{L^q}\\ \leq& C\max\left\{\frac{(\log n)^{\frac{s}{2s+1}+\epsilon}}{n^{\frac{s}{2s+1}}}\|f\|_{W^{s,\infty}}^{\frac{1}{2k+1}}, \frac{(\log n)^r}{n^{1/2}}, \frac{(\log n)^{(k-s)\tilde{\epsilon}-\frac{k-s}{2s+1}}}{n^{\frac{s}{2s+1}+\vartheta\frac{4ks-1}{2s+1}}}\|f\|_{W^{s,\infty}}\right\}\\ \leq& C(\log n)^{\frac{s}{2s+1}+\epsilon}n^{-\frac{s}{2s+1}}\|f\|_{W^{s,\infty}}^{\frac{1}{2k+1}},\end{aligned}$$

with $\epsilon = \frac{(k-s)\tilde{\epsilon}+(2r-1)k}{2k+1} > \frac{(2r-1)k}{2k+1}$. Note that $\lim_{n\to\infty}\mathbb{P}\{\|\xi_n\|_{\mathcal{B}} > \tilde{\gamma}_n\} = 0$ by Proposition 3.5. Thus we obtain that

$$\begin{aligned}\|\hat{f}_{\gamma_n} - f\|_{L^q} &\leq \|\hat{f}_{\gamma_n} - g_\lambda\|_{L^q} + \|g_\lambda - f\|_{L^q}\\ &\leq C(\log n)^{\frac{s}{2s+1}+\epsilon}n^{-\frac{s}{2s+1}}\|f\|_{W^{s,\infty}}^{\frac{1}{2k+1}} + C(\log n)^{\frac{s}{2s+1}-s\tilde{\epsilon}}n^{-\frac{s}{2s+1}}\|f\|_{W^{s,\infty}}\\ &\leq C(\log n)^{\frac{s}{2s+1}+\epsilon}n^{-\frac{s}{2s+1}}\|f\|_{W^{s,\infty}}^{\frac{1}{2k+1}}\end{aligned}$$

almost surely as $n \to \infty$.

If $p(t) := \mathbb{P}\{\|\xi_n\|_{\mathcal{B}} \leq t\}$, it follows that for n sufficiently large,

$$\begin{aligned}&\mathbb{E}\left[\|\hat{f}_{\gamma_n} - f\|_{L^q}\right]\\ \leq&\mathbb{P}\{\|\xi_n\|_{\mathcal{B}} \leq \tilde{\gamma}_n\}\, C(\log n)^{\frac{s}{2s+1}+\epsilon}n^{-\frac{s}{2s+1}}\|f\|_{W^{s,\infty}}^{\frac{1}{2k+1}}\\ &+\int_{\tilde{\gamma}_n}^{\infty}\sup\left\{\|\hat{f}_{\gamma_n} - f\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t\right\}dp(t)\end{aligned}$$

$$\leq C(\log n)^{\frac{s}{2s+1}+\epsilon} n^{-\frac{s}{2s+1}} \|f\|_{W^{s,\infty}}^{\frac{1}{2k+1}} + \int_{\tilde{\gamma}_n}^{\infty} \sup\left\{ \|\hat{f}_{\gamma_n} - f\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t \right\} dp(t).$$

As in the proof of Theorem 4.4, we see that

$$\int_{\tilde{\gamma}_n}^{\infty} \sup\left\{ \|\hat{f}_{\gamma_n} - f\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t \right\} dp(t) \leq C n^{\frac{2k+5d}{2d}} \tilde{\gamma}_n \exp\left(-\frac{\tilde{\gamma}_n^2}{2\sigma^2} \right) \to 0,$$

as $n \to \infty$.
It is easy to see that the above argument also holds for $f \in B_{\infty,0}^{s,p'}(\mathbb{T})$ with $1 \leq p' \leq \infty$. This completes the proof. □

References

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202.

Besicovitch, A. S. (1945). A general form of the covering principle and relative differentiation of additive functions. *Proc. Cambridge Philos. Soc.*, 41:103–110.

Besicovitch, A. S. (1946). A general form of the covering principle and relative differentiation of additive functions. II. *Proc. Cambridge Philos. Soc.*, 42:1–10.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636.

Boyle, J. P. and Dykstra, R. L. (1986). A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer.

Brenner, S. C. and Scott, L. R. (2008). *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition.

Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924.

Cai, T. (2002). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sinica*, 12(4):1241–1273.

Cai, T., Wang, L., and Xu, G. (2010). Stable recovery of sparse signals and an oracle inequality. *IEEE Trans. Inform. Theory*, 56(7):3516–3522.

Cai, T. and Zhou, H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.*, 37(2):569–595.

Candès, E. J. and Guo, F. (2002). New multiscale transforms, minimum total variation synthesis: applications to edge-preserving image reconstruction. *Signal Processing*, 82:1519–1543.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874. Dedicated to the memory of Lucien Le Cam.

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

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statist. Sinica*, 23(1):409–428.

Chesneau, C., Fadili, J., and Starck, J.-L.-L. (2010). Stein block thresholding for wavelet-based image deconvolution. *Electron. J. Stat.*, 4:415–435.

Cohen, A., Hoffmann, M., and Reiß, M. (2004). Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42(4):1479–1501 (electronic).

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29(1):1–65. With discussion and rejoinder by the authors.

Davies, P. L., Kovac, A., and Meise, M. (2009). Nonparametric regression, confidence regions and regularization. *Ann. Statist.*, 37:2597–2625.

Davies, P. L. and Meise, M. (2008). Approximating data with weighted smoothing splines. *J. Nonparametr. Stat.*, 20(3):207–228.

de Boor, C. (2012). On the (Bi)infinite case of Shadrins theorem concerning the L_∞-boundedness of the L_2-spline projector. *Proc. Steklov Inst. Math.*, 277:73–78.

Deng, W. and Yin, W. (2015). On the global and linear convergence of the generalized alternating direction method of multipliers. *J. Sci. Comput.* In press.

Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression-what is a reasonable choice? *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 60(4):751–764.

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

Donoho, D. L. (1995a). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627.

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

Donoho, D. L., Elad, M., and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 57(2):301–369. With discussion and rejoinder by the authors.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Universal near minimaxity of wavelet shrinkage. In Pollard, D. and Yang, G., editors, *Festschrift for Lucien Le Cam*, pages 183–218. Springer, New York.

Dümbgen, L. and Kovac, A. (2009). Extensions of smoothing via taut strings. *Electron. J. Stat.*, 3:41–75.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152.

Dyn, N., Narcowich, F. J., and Ward, J. D. (1999). Variational principles and Sobolev-type estimates for generalized interpolation on a Riemannian manifold. *Constr. Approx.*, 15(2):175–208.

Eggermont, P. P. B. and LaRiccia, V. N. (2009). *Maximum Penalized Likelihood Estimation. Volume II.* Springer Series in Statistics. Springer, Dordrecht. Regression.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Flemming, J. (2012). Solution smoothness of ill-posed equations in Hilbert spaces: four concepts and their cross connections. *Appl. Anal.*, 91(5):1029–1044.

Flemming, J. and Hofmann, B. (2010). A new approach to source conditions in regularization with general residual term. *Numer. Funct. Anal. Optim.*, 31(2):254–284.

Frick, K., Marnitz, P., and Munk, A. (2012). Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electron. J. Stat.*, 6:231–268.

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

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580. With discussion and rejoinder by the authors.

Glaz, J. and Balakrishnan, N., editors (1999). *Scan Statistics and Applications*. Statistics for Industry and Technology. Birkhäuser, Boston.

Goldenshluger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2):135–170.

Golitschek, M. v. (2014). On the L_∞-norm of the orthogonal projector onto splines. A short proof of A. Shadrin's theorem. *J. Approx. Theory*, 181:30–42.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Groetsch, C. W. (1984). *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, Boston.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer-Verlag, New York.

Hall, P., Kay, J. W., and Titterinton, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77:521–528.

Hall, P., Penev, S., Kerkyacharian, G., and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statistics and Computing*, 7:115–124.

Haltmeier, M. and Munk, A. (2013). Extreme value analysis of empirical frame coefficients and implications for denoising by soft-thresholding. *Appl. Comput. Harmon. Anal.*, page in press.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Hein, T. (2008). Convergence rates for regularization of ill-posed problems in Banach spaces by approximate source conditions. *Inverse Probl.*, 24(4):045007, 10.

Hoffmann, M. and Reiss, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.*, 36(1):310–336.

Hofmann, B. (2006). Approximate source conditions in Tikhonov-Phillips regularization and consequences for inverse problems with multiplication operators. *Math. Methods Appl. Sci.*, 29(3):351–371.

Hofmann, B., Kaltenbacher, B., Pöschl, C., and Scherzer, O. (2007). A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Probl.*, 23(3):987–1010.

Hofmann, B. and Yamamoto, M. (2005). Convergence rates for Tikhonov regularization based on range inclusions. *Inverse Probl.*, 21(3):805–820.

Ivanov, V. K., Vasin, V. V., and Tanana, V. P. (2002). *Theory of Linear Ill-posed Problems and Its Applications*, volume 36. Walter de Gruyter, second edition. Translated and revised from the 1978 Russian original version.

Kabluchko, Z. (2011). Extremes of the standardized Gaussian noise. *Stochastic Process. Appl.*, 121(3):515–533.

Korostelev, A. and Korosteleva, O. (2011). *Mathematical Statistics*, volume 119 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI. Asymptotic minimax theory.

Kress, R. (1998). *Numerical Analysis*, volume 181 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

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(3):929–947.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.

Mallat, S. (2009). *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier/Academic Press, Amsterdam, third edition.

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

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.*, 9(1):141–142.

Narcowich, F. J., Schaback, R., and Ward, J. D. (2002). Approximations in Sobolev spaces by kernel expansions. *J. Approx. Theory*, 114(1):70–83.

Narcowich, F. J., Ward, J. D., and Wendland, H. (2003). Refined error estimates for radial basis function interpolation. *Constr. Approx.*, 19(4):541–564.

Nemirovski, A. (1985). Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet. (in Russian)*, 3:50–60. *J. Comput. System Sci.*, 23:1–11, 1986 (in English).

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin.

Nesterov, Y., Nemirovskii, A., and Ye, Y. (1994). *Interior-point Polynomial Algorithms in Convex Programming*, volume 13. SIAM.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, 12(4):1215–1230.

Rivera, C. and Walther, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat.*, 40:752–769.

Scherer, K. and Shadrin, A. (1999). New upper bound for the B-spline basis condition number. II. A proof of de Boor's 2^k-conjecture. *J. Approx. Theory*, 99(2):217–229.

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. (2009). *Variational Methods in Imaging*, volume 167. Springer, New York.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, Cambridge, third edition.

Shadrin, A. Y. (2001). The L_∞-norm of the L_2-spline projector is bounded independently of the knot sequence: A proof of de Boor's conjecture. *Acta Math.*, 187(1):59–137.

Sharpnack, J. and Arias-Castro, E. (2014). Exact asymptotics for the scan statistic and fast alternatives. *arXiv:1409.7127*.

Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4):1285–1297.

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

Triebel, H. (1983). *Theory of Function Spaces*. Modern Birkhäuser Classics. Birkhäuser Verlag, Basel, Basel.

Triebel, H. (1992). *Theory of Function Spaces. II*, volume 84 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel.

Triebel, H. (1995). *Interpolation Theory, Function Spaces, Differential Operators*. Johann Ambrosius Barth, Heidelberg, second edition.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van de Geer, S. A. (1988). *Regression analysis and empirical processes*, volume 45 of *CWI Tract*. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam.

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

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.

Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033.

Ziemer, W. P. (1989). *Weakly Differentiable Functions. Sobolev Spaces and Functions of Bounded Variation*, volume 120 of *Graduate Texts in Mathematics*. Springer Verlag, Berlin etc.