Variational Estimators in Statistical Multiscale Analysis

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

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

vorgelegt von

Housen Li

aus Liaoning, China

Göttingen, 2016

Betreuungsausschuss:

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

Prof. Dr. Markus Haltmeier,
Institut für Mathematik – Angewandte Mathematik, Universität Innsbruck

Mitglieder der Prüfungskommission:

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

Korreferent:
Prof. Dr. Markus Haltmeier,
Institut für Mathematik – Angewandte Mathematik, Universität Innsbruck

Weitere Mitglieder der Prüfungskommission:

PD Dr. Timo Aspelmeier,
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Dorothea Bahns,
Mathematisches Institut, Universität Göttingen

Prof. Dr. Tatyana Krivobokova,
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Max Wardetzki,
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Tag der mündlichen Püfung: 17.02.2016

Dedicated to the memory of my grandfather, 1924 – 2015

谨以此书献给我的祖父（1924 - 2015）

Summary

In recent years, a novel type of multiscale variational statistical approaches, based on so-called multiscale statistics, have received increasing popularity in various applications, such as signal recovery, imaging and image processing, mainly because they in general perform uniformly well over a range of different scales (i.e. sizes of features). By contrast, the underlying statistical theory for these methods is still lacking, in particular with regard to the asymptotic convergence behavior. For the sake of narrowing such gap, we propose and analyze a constrained variational approach, which we call MultIscale Nemirovski-Dantzig (MIND) estimator, for recovering smooth functions in the settings of nonparametric regression and statistical inverse problems. It 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). To be precise, 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 work is the derivation of convergence rates of MIND both almost surely and in expectation for nonparametric regression and linear statistical inverse problems. To this end, we generalize the Nemirovski's interpolation inequality for the multiresolution norm and Sobolev norms, and introduce the method of approximate source conditions to our statistical setting. Based on these tools, we are able to obtain certain convergence rates under abstract smoothness assumptions about the truth. For a one-dimensional signal, such assumptions can be translated into classical smoothness classes and source sets by means of the approximation properties of B-splines. As a consequence, MIND attains almost minimax optimal rates simultaneously over a large range of Sobolev and Besov classes, for nonparametric regression of functions and their derivatives. Analogous results have been also obtained for certain linear statistical inverse problems, such as deconvolution if the Fourier coefficients of the convolution kernel is of polynomial decay. Put differently, these results reveal that MIND possesses certain adaptation to the smoothness of the underlying true signal. In parallel, we have presented a similar analysis for a penalized version of MIND, and its parameter choice via the Lepskiĭ balancing principle. Finally, complimentary to the asymptotic analysis, we examine the finite sample performance of MIND by various numerical simulations.

Acknowledgement

First of all, I would like to express my sincere gratitude to my principal advisor Prof. Axel Munk for introducing me into the field of mathematical statistics, and providing this stimulating and challenging topic of this work. I benefit a lot from his deep and penetrating views on so many areas of mathematical sciences, and feel particularly indebted to him for his thoughts on what are really interesting problems. Besides his guidance, encouragement, and contributions regarding this project, he is also a great mentor for my life. Further, I would like to thank my second advisor Prof. Markus Haltmeier for his support and guidance throughout the first stage of my study and the first project, for hosting a nice stay at the University of Innsbruck, and for proofreading most parts of this work.

Special thanks are owed to Prof. Markus Grasmair for his extraordinary assistance with this work, as well as for his patience, ideas and immense knowledge. This work would be impossible without various vivid discussions that we had when we together spent several weekends in Göttingen, and half a month at the Catholic University of Eichsttt-Ingolstadt.

I am indebted to Prof. Jens Frahm and his group for cooperation on the project of dynamic magnetic resonance imaging, and to Dr. Hannes Sieling for the joint work on the FDR-control in change-points estimation. Moreover, I want to thank Prof. Tatiana Krivobokova and Prof. Robert Schaback for discussion on splines, and Dr. Timo Aspelmeier for helpful comments and computational assistance.

I wish to express my gratitude to all the members at the IMS, and to many colleagues form Prof. Erwin Neher's group, as well as Dr. Michael Habeck and his students at the MPIbpc. Special thanks go to my officemate Dr. Frank Werner for his companionship and extensive discussions, for providing some enlightenment on the over-smoothing topic, and for proofreading part of this work. In addition, I want to thank Prof. Lizhi Cheng, the Chinese community at the MPIbpc, and many other friends for their supportiveness.

The financial support by the China Scholarship Council (CSC), the SFB 755 "Nanoscale Photonic Imaging", the Felix Bernstein Institute for Mathematical Statistics in the Bioscience (FBMS), and the RTG 2088 "Discovering structure in complex data: Statistics meets Optimization and Inverse Problems" is gratefully acknowledged.

Finally, I greatly appreciate the constant support and understanding from my family and my fiancé Qian Liu. In particular, the encouragement, surprises, and love from Qian made my whole PhD study a pure enjoyment.

Contents

1. Introduction

In this work, we will consider the estimation of a smooth function $f\colon [0,1]^d \to \mathbb{R}$ from n measurements

$$y_n(x) = (Tf)(x) + \xi_n(x) \qquad \text{for } x \in \Gamma_n, \tag{1.1}$$

where Γ_n is the regular grid on $[0,1]^d$ containing n equidistant points, $\{\xi_n(x); x \in \Gamma_n\}$ a set of independent, identically distributed (i.i.d.) centered sub-Gaussian random variables, and T a bounded linear operator. In particular, we are interested in the nonparametric regression, i.e. T is the identity operator, and the statistical inverse problems, i.e. T does not have a bounded inverse, as well. The model (1.1) is typical for a considerable number of practical applications, see e.g. (Korostelëv and Tsybakov, 1993; Chan and Shen, 2005; Mallat, 2009). 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 is known.

1.1. Methodology

1.1.1. Variational statistical estimation

Since the fundamental work of (Nadaraya, 1964; Stone, 1984) and many others, the literature on nonparametric 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. As an extension, statistical inverse problems (due to Sudakov and Khalfin, 1964) deal with indirect data, and cast inverse problems as statistical estimation and inference problems. This research topic has been expanded and developed along with the nonparametric regression, see (O'Sullivan, 1986; Plaskota, 1996; Tenorio, 2001; Evans and Stark, 2002; Kaipio and Somersalo, 2005; Cavalier, 2008) for surveys. A prodigious amount of these estimation methods in both nonparametric regression and statistical inverse problems can be cast 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 1.1.

Penalized estimation is a solution of the Lagrangian variational problem (also known as

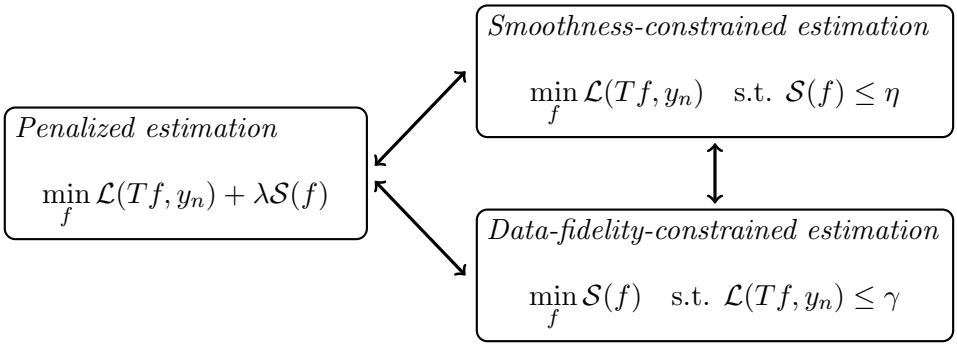

Figure 1.1.: Variational statistical estimation.

generalized Tikhonov(-Phillips) regularization, Phillips, 1962; Tikhonov, 1963a,b)

$$\min_f \mathcal{L}(Tf, y_n) + \lambda\mathcal{S}(f). \tag{1.2}$$

The *regularization term* $\mathcal{S}(f)$ accounts for a-priori assumptions of the truth f, such as smoothness, sparsity, etc. The *data fidelity term* $\mathcal{L}(Tf, y_n)$ 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 estimation (see e.g. van de Geer, 1988; Mair and Ruymgaart, 1996; Bissantz et al., 2007; Eggermont and LaRiccia, 2009; Bühlmann and van de Geer, 2011, for general exposition), and maximum a posteriori estimation (see e.g. Kaipio and Somersalo, 2005; Stuart, 2010) from Bayesian perspective. Prominent examples include smoothing splines (Wahba, 1990), local polynomial estimators (Fan and Gijbels, 1996), locally adaptive splines (Mammen and van de Geer, 1997), and non-concave penalized methods (Antoniadis and Fan, 2001; Fan and Li, 2001). 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 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; Goldenshluger and Pereverzev, 2000; Mathé and Pereverzev, 2006).

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

$$\min_f \mathcal{L}(Tf, y_n) \qquad \text{subject to } \mathcal{S}(f) \leq \eta. \tag{1.3}$$

It includes the well-known lasso (Tibshirani, 1996) for $\mathcal{L}(\cdot, y_n) = \|\cdot - y_n\|_2$ and $\mathcal{S} = \|\cdot\|_1$ as a special case. Another example is Nemirovski's (1985) regression estimator $\hat{f}_{p,\eta}$ defined as a solution to

$$\min_f \|S_n f - y_n\|_{\mathcal{B}} \qquad \text{subject to } \|D^k f\|_{L^p} \leq \eta, \tag{1.4}$$

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 on cubes specified by $\mathcal{B}$ (see Section 2.2 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 (1.3), given by

$$\min_f \mathcal{S}(f) \qquad \text{subject to } \mathcal{L}(Tf, y_n) \leq \gamma. \tag{1.5}$$

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 the regression of 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 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, which has the form

$$\min_{f\in\mathbb{R}^p} \|f\|_1 \qquad \text{subject to } \|T^*(Tf - y_n)\|_\infty \leq \gamma, \qquad \text{with matrix } T \in \mathbb{R}^{n\times p}. \tag{1.6}$$

Many other ℓ^1-minimization approaches for recovering sparse signals also take the form of (1.5), see (Donoho et al., 2006; Cai et al., 2010) for example.

In the most common case that $\mathcal{L}(\cdot, y_n)$ and $\mathcal{R}(\cdot)$ are convex functionals, all three estimation methods in Figure 1.1 can be regarded, from a convex analysis point of view, as equivalent, as under rather weak assumptions each estimator in (1.2), (1.3), (1.5) can be obtained as a solution of the other optimization problems (cf. Bickel et al., 2009, for this in the case of the lasso and the Dantzig selector). More precisely, if $\hat{f}$ is a solution of (1.2), then it is also a solution of (1.5) for $\gamma := \mathcal{L}(T\hat{f}, y_n)$. Conversely, if $\hat{f}$ is a solution of (1.5), and $\hat{f} \notin \arg\min \mathcal{S}$, and if the Slater's condition

$$\mathcal{L}(Tf_0, y_n) < \gamma \qquad \text{for some } f_0 \text{ in the domain of } \mathcal{S}$$

holds, then there exists some $\lambda \geq 0$ such that $\hat{f}$ also solves (1.2). Similar relation holds between (1.2) and (1.3) as well. These equivalences essentially follow from duality (cf. Ekeland and Témam, 1999, Proposition 3.1, Chapter III) in convex optimization, see (Ivanov et al., 2002; Teuber et al., 2013; Haltmeier and Munk, 2015) for a detailed argument. However,

we emphasize that the correspondence between the parameters λ, η, γ for the equivalence relations 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 (1.5) 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 right hand side of (1.5) with probability at least $1 - \alpha$, which immediately leads to the so called *smoothness guarantee* of the estimate $\hat{f}$ in (1.5),

$$\inf_f \mathbb{P}\left\{\mathcal{S}(\hat{f}) \leq \mathcal{S}(f)\right\} \geq 1 - \alpha. \tag{1.7}$$

1.1.2. 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 (1.4) as well, which is given by

$$\min_f \|D^k f\|_{L^p} \qquad \text{subject to } \|S_n f - y_n\|_{\mathcal{B}} \leq \gamma. \tag{1.8}$$

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 right hand side of (1.8), Nemirovski (1985) derived the convergence rate of (1.8) (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 for nonparametric regression and statistical inverse problems has not been provided so far, to the best of our knowledge. Intending to fill such gap, we will consider a particular multiscale fidelity-constrained estimation method $\hat{f}_{\gamma_n}$ defined by

$$\hat{f}_{\gamma_n} = \arg\min_f \frac{1}{2}\|D^k f\|_{L^2}^2 \qquad \text{subject to } \|S_n T f - y_n\|_{\mathcal{B}} \leq \gamma_n, \tag{1.9}$$

which we call the *MultIscale Nemirovski-Dantzig estimator* (MIND). The choice of the name credits the fact that it generalizes a particular "reverse" Nemirovski's estimator (1.8) with $p = 2$ to statistical inverse problems, and the right hand side is a (multiscale) extension of the Dantzig estimator (1.6).

1.2. A heuristic explanation

For simplicity, we assume throughout this section that the random errors $\xi_n(x)$ in (1.1) are i.i.d. standard Gaussian distributed.

1.2.1. Separation between signal and noise

Concerning the rationale for the chosen methodology, we illustrate the intuition behind MIND's ability to recover features of the truth in a multiscale fashion by a *toy example* in the setting of nonparametric regression.

Example 1. Let us consider the estimation of a smooth function $f\colon [0,1]^d \to \mathbb{R}$ from measurements in (1.1) with T being the identity operator. Assume now that we have an estimator $\hat{f} \equiv \hat{f}_{s,t,a}$, such that

$$\hat{f}_{s,t,a}(x) := f(x) + s\varphi_a(x) + t\xi_n(x) \qquad \text{for every } x \in \Gamma_n,$$

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

$$\varphi(x) := \begin{cases} Ce^{\frac{1}{|x|^2-1}} & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1 \end{cases} \qquad \text{for } x \in \mathbb{R}^d,$$

with the constant C such that $\|\varphi\|_{L^2} = 1$. 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 $t\xi_n$. By elementary computations one can show that

$$\left|t - \frac{s}{a^{d/2}}\right| n \lesssim \|\hat{f} - f\|_1 \lesssim \left(t + \frac{s}{a^{d/2}}\right) n,$$

$$\|\hat{f} - f\|_2 \sim (t+s)\sqrt{n},$$
$$\|\hat{f} - f\|_\infty \sim t\sqrt{\log n} + sa^{d/2},$$

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 for the cube $B := [1/2 - 1/a, 1/2 + 1/a)^d$,

$$\frac{1}{\sqrt{\#B \cap \Gamma_n}} \Big| \sum_{x \in B \cap \Gamma_n} \hat{f}(x) - f(x) \Big| \sim t\sqrt{\log n} + s\sqrt{n}, \tag{1.10}$$

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 (1.10). 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 (1.10). Still, it is possible to use the local averages in (1.10) 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 (1.10), cf. Figure 2.1. Therefore, if we choose $\gamma_n \sim \log n$, the multiscale constraint of MIND in (1.9) will guarantee that every feasible candidate contains no deterministic distortion, and the smoothness-enforcing regularization term will then select the one without random distortion. The combination of both ensures that MIND avoids both deterministic and random distortions, thus being close to the truth.

1.2.2. Multiscale testing

In addition, we give an interpretation of the multiscale constraint in MIND from a hypothesis testing perspective. Under the model (1.1), given a cube $B \subset [0,1]^d$, and a function $\tilde{f}$, we have by simple calculation that the normalized local average on B

$$\frac{1}{\sqrt{\#B \cap \Gamma_n}} \Big| \sum_{x \in B \cap \Gamma_n} y_n - T\tilde{f}(x) \Big|$$

is the likelihood ratio testing statistic for the multiple hypotheses

$$H_0 : (T\tilde{f})_B = (Tf)_B \quad \text{vs.} \quad H_1 : (T\tilde{f})_B \neq (Tf)_B \tag{1.11}$$

where $g_B := \sum_{x \in B \cap \Gamma_n} g(x)/\sqrt{\#B \cap \Gamma_n}$ for any function g. As a direct implication, the multiresolution norm $\|y_n - S_n T\tilde{f}\|_{\mathcal{B}}$ is a statistic for testing the hypotheses (1.11) simultaneously over $B \in \mathcal{B}$ (cf. Definition 2.2.1). Thus, if we calibrate the threshold in such a way that the family-wise error is controlled, it holds with the prescribed probability that for every candidate function $\tilde{f}$ lying in the constraint of MIND (the right hand side of (1.9)), $(T\tilde{f})_B$ is close to $(Tf)_B$ uniformly over $B \in \mathcal{B}$, which in turn confirms that $T\tilde{f}$ is close to Tf over various scales and locations specified by $\mathcal{B}$. This is again a merit of the multiscale data fidelity term, which is not shared by the data fidelity term with respect to classic ℓ^p-norms, i.e. $\|y_n - T\tilde{f}\|_p$, with $1 \leq p \leq \infty$.

Importantly, we note that for mildly ill-posed problems (which are studied in Chapter 3), every minimax optimal test for $H_0 : T\tilde{f} = Tf$ is necessarily minimax optimal for $H_0' : \tilde{f} = f$, see (Laurent et al., 2011). This indicates that the multiscale test in the form of (1.11) would also be a reasonable test for $H_0' : \tilde{f} = f$. It further suggests that $\|y_n - S_n T\tilde{f}\|_{\mathcal{B}}$ is a plausible measure on the closeness between $\tilde{f}$ and f, which we are mainly interested in. For more details on testing problems in inverse problems, we refer to (Holzmann et al., 2007; Butucea et al., 2009; Laurent et al., 2012; Ingster et al., 2012).

1.3. Main results

We mainly focus on the bounds for the L^q-risk ($1 \leq q \leq \infty$) of the MIND estimator (1.9) for nonparametric regression and statistical inverse problems. The main contributions are summarized as follows. First, we derive two interpolation inequalities of the multiresolution norm and Sobolev norms, as extensions of the original one by (Nemirovski, 1985). These inequalities provide a crucial link between the loss, the regularization, and the multiscale data fidelity functional, which is fundamental for the theoretical analysis of MIND.

Second, 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 and statistical inverse problems. By combining them with the interpolation inequalities mentioned above, 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).

Third, we present both the L^q-risk convergence rate ($1 \leq q \leq \infty$) and the almost sure convergence rate of MIND for nonparametric regression and statistical inverse problems, 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.

Fourth, we show a *partial adaptation* property of MIND in one dimension. More precisely, for nonparametric regression of functions and derivatives and for a fixed k, it attains minimax optimality (up to a log-factor) simultaneously over Sobolev ellipsoids in $W^{s,p}$ and Besov ellipsoids in $B_p^{s,p'}$ for all $(s,p) \in [1,k) \times \{\infty\} \cup \{k\} \times [2,\infty] \cup [k+1,2k] \times [2,\infty]$ and $p' \in [1,\infty]$. In case of statistical inverse problems, if the operator T and its adjoint T^* are β-smoothing (see Definition 3.1.1) for some $\beta \geq 0$, MIND with a fixed k-th order regularization adapts to the truth smoothness, and is almost minimax optimal, over functions f that satisfy Hölder-type source conditions

$$f = T^* g \qquad \text{with } g \in W^{s,2}$$

for any fixed $s \in \{k-\beta\} \cup [k-\beta+1, 2k]$. These results explain to some extent the remarkably good multiscale reconstruction properties of MIND empirically found in various signal recovery and imaging applications, see Sections 2.7 and 3.4, and (Candès and Guo, 2002; Davies et al., 2009; Frick et al., 2013).

Finally, we note that a penalized version of MIND given by

$$\min_f \|S_n T f - y_n\|_{\mathcal{B}} + \frac{\alpha}{2} \|D^k f\|_{L^2}^2, \tag{1.12}$$

if combined with the Lepskiĭ balancing principle, performs nearly the same as MIND in asymptotic sense, e.g. possessing the aforementioned partial adaptation property. In particular, we give an exemplary analysis for this variant of MIND in the setting of nonparametric regression. For a fixed sample size, a certain constant involved in Lepskiĭ balancing principle turns out to be quite pessimistic, and may deteriorate the performance of the penalized variant in practice. Thus, as far as the finite sample behavior is concerned, we recommend the original MIND, which allows for a universal threshold, and meanwhile provides statistical inference on the smoothness of the truth (cf. the smoothness guarantee in (1.7)). In contrast, we emphasize again that the Nemirovski's estimator $\hat{f}_{p,\eta}$ in (1.4) for nonparametric regression is only known to be (nearly) minimax optimal if the parameters k and p match the actual smoothness of the truth perfectly. Such a strict requirement makes it practically difficult to select the proper values for k, p, and η.

The rest of the work is organized as follows. In Chapter 2, we focus on the nonparametric regression of functions and the derivatives. After some necessary notation in Section 2.1, we present the multiresolution norm together with its deterministic and stochastic properties in Section 2.2. Section 2.3 is devoted to approximate source conditions and so called distance functions, which provide methods for analyzing the L^q-loss ($1 \leq q \leq \infty$) 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 2.4. In parallel, Section 2.5 provides an asymptotical analysis of the penalized version of MIND, as well as the Lepskiĭ balancing principle. In addition to the asymptotic results, based on the algorithms in Section 2.6, the finite sample behavior, as well as choices of the tuning parameter, of MIND is examined empirically on simulated examples in Section 2.7.

In Chapter 3, we extend the previous analyses to statistical inverse problems. By means of an extended interpolation inequality, we derive the convergence rates for MIND in terms of approximate source conditions in Section 3.2. Section 3.3 considers the case $d = 1$. By the estimate of distance function in the previous chapter, the abstract smoothness assumptions are translated into classical Hölder-type source conditions, from which we again derive the explicit convergence rates and the partial adaptation property for MIND. Moreover, some numerical studies are collected in Section 3.4.

The last part of this work is contained in Chapter 4, where we present some discussions and open questions. Technical proofs are given in the appendix.

2. Nonparametric Regression

In this chapter, we consider the asymptotic properties of MIND for regression of smooth functions and their derivatives. We start with a general framework of convergence analysis by means of approximate source conditions. On the basis of this framework, in one dimensional setting, we derive convergence rates for Sobolev and Besov smooth classes, and examine the minimax optimality and adaptation behaviors. Complimentary to the asymptotic findings, we study the finite sample performance of MIND by numerical simulations as well. In parallel, we also consider a penalized formulation of MIND, and investigate its properties in the same setting. Most of the results have appeared in (Grasmair et al., 2015), while the extensions mainly include a proof of an interpolation inequality between multiresolution norms and Sobolev norms, asymptotics for estimation of derivatives, an analysis of the penalized version of MIND and Lepskiĭ's balancing principle, and some additional numerical studies.

2.1. Model and notation

The nonparametric regression problem is to estimate a function $f : [0,1]^d \to \mathbb{R}$ from n measurements

$$y_n(x) = f(x) + \xi_n(x) \qquad \text{for } x \in \Gamma_n. \tag{2.1}$$

Here, the *regular grid* Γ_n on $[0,1]^d$ is given by

$$\Gamma_n := \left\{ (\frac{\tau_1}{n^{1/d}}, \dots, \frac{\tau_d}{n^{1/d}}); \tau_i = 0, \dots, n^{1/d} - 1, \text{ for } i = 1, \dots, d \right\}, \tag{2.2}$$

and the error $\{\xi_n(x); x \in \Gamma_n\}$ is a set of i.i.d. centered sub-Gaussian random variables with scale parameter σ, i.e., each $\xi_n(x)$ satisfies

$$\mathbb{E}\left[e^{\tau \xi_n(x)}\right] \leq e^{(\tau\sigma)^2/2} \qquad \text{for every } \tau \in \mathbb{R}. \tag{2.3}$$

The sub-Gaussian random variable in (2.3) includes centered Gaussian random variable, and bounded and centered random variable on $[-\sigma/2, \sigma/2]$ as special cases. We now introduce the point evaluation S_n on the grid Γ_n as the mapping

$$f \mapsto S_n f = \big(f(x)\big)_{x \in \Gamma_n} \in \mathbb{R}^{\Gamma_n},$$

for every continuous function $f\colon [0,1]^d \to \mathbb{R}$. The nonparametric regression (2.1) can then be rewritten as

$$y_n = S_n f + \xi_n,$$

where $y_n := \big(y_n(x)\big)_{x\in\Gamma_n}$, and $\xi_n := \big(\xi_n(x)\big)_{x\in\Gamma_n}$.

For technical simplicity, we make the following assumption throughout this chapter.

Assumption 1. (a) The truth f is *periodic*, in the sense that it can be regarded as a (continuous) function defined on the d-dimensional torus $\mathbb{T}^d \sim \mathbb{R}^d/\mathbb{Z}^d$.

(b) The truth f has *mean zero*, i.e., $\int_{\mathbb{T}^d} f(x)\,dx = 0$.

(c) The scale parameter σ in (2.3) of the random error is *known*.

Remark 2.1.1. The main reason for assumption (a) is that this avoids having to deal with boundary conditions that would have to be taken into account in non-periodic cases. The assumption (b) is to simplify norms of Sobolev spaces, and we will see that dropping it is actually possible (see Proposition 2.3.6). The last assumption is reasonable, since the level σ of random error can be easily pre-estimated with $\sqrt{n}$-rate, see e.g. (Rice, 1984; Hall et al., 1990; Dette et al., 1998) among other references.

In this setting, the MIND estimator $\hat{f}_{\gamma_n}$ (i.e. $T = I$ in (1.9)) is defined by

$$\hat{f}_{\gamma_n} = \operatorname*{arg\,min}_{f\in H_0^k(\mathbb{T}^d)} \frac{1}{2}\|D^k f\|_{L^2}^2 \qquad \text{subject to } \|S_n f - y_n\|_{\mathcal{B}} \le \gamma_n, \tag{2.4}$$

where $\|\cdot\|_{\mathcal{B}}$ is the multiresolution norm with respect to the system $\mathcal{B}$ of cubes (the definition will be given in Section 2.2), and

$$\gamma_n = C(\log n)^r \quad \text{for some } r \ge \frac{1}{2} \text{ and } C > \begin{cases} 0 & \text{if } r > \frac{1}{2}, \\ \sigma\sqrt{6 + \frac{2k}{d}} & \text{if } r = \frac{1}{2}. \end{cases} \tag{2.5}$$

We stress that such choice of threshold 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 the sample size n only.

Note that the MIND estimator $\hat{f}_{\gamma_n}$ has derivatives up to order k, so its derivative can be used as a natural estimator for that of the truth f, that is, $D^\alpha \hat{f}_{\gamma_n} \approx D^\alpha f$ for each $\alpha \in \mathbb{N}_0^d$ and $|\alpha| \in [0,k)$. In what follows this derivative estimator will also be analyzed asymptotically.

2.1.1. Smooth functions on $\mathbb{T}^d$

As it is required, we will give a brief introduction to Sobolev and Besov spaces on $\mathbb{T}^d \sim \mathbb{R}^d/\mathbb{Z}^d$. These spaces are defined in a similar way as those on $\mathbb{R}^d$ or $[0,1]^d$, see (Triebel, 1983, 1992, 1995; Adams and Fournier, 2003) for further details.

Let us first introduce the *multi-index* notation for partial derivatives. A multi-index α, is a d-tuple of nonnegative integers α_i, i.e. $\alpha := (\alpha_i)_{i=1}^d \in \mathbb{N}_0^d$. The length of α is defined as

$$|\alpha| := \sum_{i=1}^{d} \alpha_i.$$

For a sufficiently smooth function f, we denote partial (weak) derivatives by

$$D^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}, \qquad \text{and } D^l f := \big(D^\alpha f\big)_{|\alpha|=l,\, \alpha\in\mathbb{N}_0^d} \quad \text{for } l \in \mathbb{N}_0.$$

Given $1 \le p \le \infty$ and $k \in \mathbb{N}_0$, we define the *Sobolev space* $W^{k,p}(\mathbb{T}^d)$ by

$$W^{k,p}(\mathbb{T}^d) := \big\{ f \in L^p(\mathbb{T}^d) \colon D^\alpha f \in L^p(\mathbb{T}^d) \text{ for every } \alpha \in \mathbb{N}_0^d \text{ and } 0 \le |\alpha| \le k \big\}.$$

The norm on $W^{k,p}(\mathbb{T}^d)$ is defined by

$$\|f\|_{W^{k,p}} := \begin{cases} \Big(\sum_{0\le|\alpha|\le k,\, \alpha\in\mathbb{N}_0^d} \|D^\alpha f\|_{L^p}^p \Big)^{1/p} & \text{if } 1 \le p < \infty \\ \max_{0\le|\alpha|\le k,\, \alpha\in\mathbb{N}_0^d} \|D^\alpha f\|_{L^\infty} & \text{if } p = \infty \end{cases}$$

for every $f \in W^{k,p}(\mathbb{T}^d)$. It is known that $\big(W^{k,p}(\mathbb{T}^d), \|\cdot\|_{W^{k,p}}\big)$ is actually a Banach space.

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

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

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

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

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

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

Given $1 \le p \le \infty$, $t \ge 0$ and $r \in \mathbb{N}$, the r-th *modulus of smoothness* of $f \in L^p(\mathbb{T}^d)$ is defined 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* $\|\cdot\|_{B_p^{s,p'}}$, with $s > 0, 1 \le p, p' \le \infty$, as

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

where $r > s$, $r \in \mathbb{N}$ is arbitrary, and

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

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

$$B_p^{s,p'}(\mathbb{T}^d) := \{ f \in L^p(T^d); \|f\|_{B_p^{s,p'}} < \infty \}.$$

For a non-integer $s \in (0, \infty)$, the *fractional order Sobolev space* $W^{s,p}(\mathbb{T}^d)$ (a.k.a. Sobolev-Slobodeckij space) is defined by $W^{s,p}(\mathbb{T}^d) := B_p^{s,p}(\mathbb{T}^d)$. One should, however, be aware of the fact that $W^{k,p}(\mathbb{T}^d) \neq B_p^{k,p}(\mathbb{T}^d)$ for all $k \in \mathbb{N}$ and $p \neq 2$. In the case of $k \in \mathbb{N}$, it actually holds that (cf. Adams and Fournier, 2003, Paragraph 7.33)

$$\begin{aligned} B_1^{k,p}(\mathbb{T}^d) \subset W^{k,p}(\mathbb{T}^d) \subset B_\infty^{k,p}(\mathbb{T}^d) \qquad & \text{for } 1 \leq p < \infty, \\ B_p^{k,p}(\mathbb{T}^d) \subset W^{k,p}(\mathbb{T}^d) \subset B_2^{k,p}(\mathbb{T}^d) \qquad & \text{for } 1 < p \leq 2, \\ \text{and } B_2^{k,p}(\mathbb{T}^d) \subset W^{k,p}(\mathbb{T}^d) \subset B_p^{k,p}(\mathbb{T}^d) \qquad & \text{for } 2 \leq p < \infty. \end{aligned}$$

Equivalently, Besov spaces can be introduced by means of the (real) interpolation theory of Banach spaces. We in particular recall the K-method. Let $(X_0, \|\cdot\|_0)$, $(X_1, \|\cdot\|_1)$ be closed subspaces of a common Banach space. If $0 < t < \infty$ and $f \in X_0 + X_1$, then Peetre (1963a,b)'s celebrated *K-functional* is given by

$$\mathcal{K}(t, f) := \inf\{ \|f_0\|_0 + t\|f_1\|_1; f = f_0 + f_1, f_0 \in X_0, f_1 \in X_1 \}.$$

For $1 \leq p \leq \infty$ and $0 < \theta < 1$, the *interpolation space* $(X_0, X_1)_{\theta,p}$ is defined as

$$(X_0, X_1)_{\theta,p} := \{ f \in X_0 + X_1; \|f\|_{\theta,p} < \infty \},$$

where

$$\|f\|_{\theta,p} := \begin{cases} \left(\int_0^1 [t^{-\theta} \mathcal{K}(t,f)]^p \frac{dt}{t} \right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \operatorname*{ess\,sup}_{0<t<1} t^{-\theta} \mathcal{K}(t,f) & \text{if } p = \infty. \end{cases}$$

One key result from interpolation space theory is given below (Triebel, 1983, Section 2.4).

Proposition 2.1.2 (Convexity theorem). *Let $1 \leq p \leq \infty$ and $0 < \theta < 1$. If A is a bounded linear operator mapping X_i into Y_i with norm $\|A\|_i$, for $i = 0, 1$, then it is also a bounded linear operator mapping $(X_0, X_1)_{\theta,p}$ into $(Y_0, Y_1)_{\theta,p}$ with norm $\|A\|_{\theta,p} \leq \|A\|_0^{1-\theta} \|A\|_1^\theta$.*

Now, for $0 < s < \infty$ and $1 \leq p, p' \leq \infty$, the Besov space $B_p^{s,p'}(\mathbb{T}^d)$ can be also defined as

$$B_p^{s,p'}(\mathbb{T}^d) := (W^{k_0,p}(\mathbb{T}^d), W^{k_1,p}(\mathbb{T}^d))_{\theta,p'}$$

where $s = (1-\theta)k_0 + \theta k_1$, $k_0 \neq k_1$ and $\theta \in (0,1)$.

In addition, it is worth noting that $\mathcal{C}^\infty(\mathbb{T}^d)$ is dense in $W^{s,p}(\mathbb{T}^d)$ and $B_p^{s,p'}(\mathbb{T}^d)$ for all $s \in (0,\infty)$ and $1 \leq p, p' \leq \infty$.

2.1.2. Functions with zero mean

Let us consider the particular closed subspaces of Sobolev and Besov spaces, which consist of functions with zero mean. We denote these spaces by

$$W_0^{s,p}(\mathbb{T}^d) := \{f \in W^{s,p}(\mathbb{T}^d); \int_{\mathbb{T}^d} f(x)dx = 0\},$$
$$\text{and } B_{p,0}^{s,p'}(\mathbb{T}^d) := \{f \in B_p^{s,p'}(\mathbb{T}^d); \int_{\mathbb{T}^d} f(x)dx = 0\},$$

where $0 < s < \infty$, and $1 \leq p, p' \leq \infty$. It is possible to introduce equivalent norms of a simpler form for them. For instance, for $W_0^{k,p}(\mathbb{T}^d)$ with $k \in \mathbb{N}$ and $1 \leq p \leq \infty$, the semi-norm given by

$$\|f\|_{W_0^{k,p}} := \Big(\sum_{|\alpha|=k,\, \alpha \in \mathbb{N}_0^d} \|D^\alpha f\|_{L^p}^p \Big)^{1/p} \tag{2.6}$$

turns out to be indeed a norm, see the following proposition.

Proposition 2.1.3 (Ziemer (1989), Theorem 4.4.2). *Let $k \in \mathbb{N}$ and $1 \leq p \leq \infty$. There exists constant C depending only on k, p such that*

$$\|f\|_{W_0^{k,p}} \leq \|f\|_{W^{k,p}} \leq C\|f\|_{W_0^{k,p}} \qquad \textit{for every } f \in W_0^{k,p}(\mathbb{T}^d).$$

From now on, when referring to $W_0^{k,p}(\mathbb{T}^d)$, we will always assume the norm $\|\cdot\|_{W_0^{k,p}}$ in (2.6), which we call the *homogeneous Sobolev norm*. If $p = 2$, we also denote $W_0^{k,2}(\mathbb{T}^d)$ by $H_0^k(\mathbb{T}^d)$, and the corresponding norm $\|\cdot\|_{W_0^{k,2}}$ by $\|\cdot\|_{H_0^k}$. In this case, the fact of Proposition 2.1.3 can be easily seen by means of Fourier series expansions, that is,

$$\|f\|_{H_0^k}^2 = (2\pi)^{2k} \sum_{\lambda \in \mathbb{Z}^d \setminus 0} |\lambda|^{2k} |c_\lambda|^2.$$

Here, the Fourier coefficient $c_\lambda := \langle f, e^{-2\pi i \langle \lambda, \cdot \rangle} \rangle_{L^2}$; in particular, $c_0 = \int_{\mathbb{T}^d} f(x)dx = 0$. Similarly, for $H_0^s(\mathbb{T}^d) := W_0^{s,2}(\mathbb{T}^d)$ with $s \in \mathbb{R}$, we introduce the equivalent norm

$$\|f\|_{H_0^s} := (2\pi)^s \Big(\sum_{\lambda \in \mathbb{Z}^d \setminus 0} |\lambda|^{2s} |c_\lambda|^2 \Big)^{1/2}. \tag{2.7}$$

2.2. Multiresolution norm

We now consider the multiresolution norm, which is one of the main tools we are working with, and its properties as well. For the sake of generality, all the results are given for functions on $[0,1]^d$ in this section. In particular, they apply to functions on $\mathbb{T}^d$ as well.

First, we define a *cube* B as a subset of $[0,1]^d$ of the form $B = \prod_{i=1}^d [a_i, a_i + h)$, where $0 \le a_i < 1$, $i = 1, \dots, d$, and $0 < h \le 1$. By $|B|$ we denote its d-dimensional volume h^d, i.e. the Lebesgue measure of B.

Definition 2.2.1 (Nemirovski (1985)). Given a non-empty system of cubes $\mathcal{B}$, the *multiresolution norm* $\|\cdot\|_{\mathcal{B}}$ on $\mathbb{R}^{\Gamma_n}$ is defined by

$$\|y\|_{\mathcal{B}} := \sup_{B \in \mathcal{B},\, B \cap \Gamma_n \neq \emptyset} \frac{1}{\sqrt{\#\Gamma_n \cap B}} \Big| \sum_{x \in \Gamma_n \cap B} y(x) \Big| \qquad \text{for } y = \big(y(x)\big)_{x \in \Gamma_n} \in \mathbb{R}^{\Gamma_n}. \tag{2.8}$$

The multiresolution norm simultaneously screens a signal on cubes of various scales and locations. With regard to multiresolution, the system of cubes $\mathcal{B}$ should be sufficiently rich. For this purpose, we introduce the *normality* and the *regularity* of a system to characterize its richness.

Definition 2.2.2. A system of cubes $\mathcal{B}$ is called *normal* (or *c-normal*), if there is a constant $c > 1$, such that for every cube $B \subseteq [0,1]^d$ there is a cube $\tilde{B} \in \mathcal{B}$ satisfying

$$\tilde{B} \subseteq B, \text{ and } |\tilde{B}| \ge |B|/c.$$

The above concept is a generalization of normality in (Nemirovski, 1985), where it was defined with $c = 6$.

Definition 2.2.3. A system of cubes $\mathcal{B}$ is called *regular* (or *m-regular*) for some $m \in \mathbb{N}$, $m \ge 2$, if it contains at least the *m-partition system*, which is defined as all sets of the form

$$[\ell m^{-j}, (\ell + 1) m^{-j}) \qquad \text{for all } \ell \in \mathbb{N}_0^d,\ j \in \mathbb{N}_0.$$

From the definition, it is clear that every m-regular system of cubes is necessarily normal (or precisely m-normal). The converse, however, does not hold in general. That is, there exist normal systems of cubes that are not m-regular for any $m \in \mathbb{N}$ (see Example 2 (c)).

Formally, the normality and regularity of a system $\mathcal{B}$ are independent of the grid Γ_n. For a given 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 number of distinct cubes of $\mathcal{B}$ on Γ_n, namely, $\#\{B \cap \Gamma_n; B \in \mathcal{B}\}$, which we call the *effective cardinality* of $\mathcal{B}$, that determines the computational complexity of evaluating the multiresolution norm, and of solving optimization problems with the multiresolution norm $\|\cdot\|_{\mathcal{B}}$ (e.g. MIND). In order to obtain

numerically feasible algorithms, one therefore would like to choose this effective cardinality of $\mathcal{B}$ as small as possible while still retaining multiresolution nature. Some examples of $\mathcal{B}$ are given below.

Example 2. (a) **The system of all cubes.** It is clearly normal and regular. The corresponding multiresolution norm also appears as a particular scan statistics, which is the maximum likelihood ratio statistic in the Gaussian setting. The scan statistics is a standard tool for detecting a deterministic signal with unknown spatial extent against a noisy background, see e.g. (Kulldorff, 1997; Glaz and Balakrishnan, 1999; Siegmund and Yakir, 2000; Dümbgen and Spokoiny, 2001; Glaz et al., 2009). However, the effective cardinality of all the cubes on Γ_n is $\mathcal{O}\left(n^2\right)$, making it computationally impractical for large scale problems.

(b) **The system of cubes with dyadic edge lengths.** It consists of cubes of the form

$$\prod_{i=1}^{d}[a_i, a_i + 2^{-l}), \qquad \text{for } a_i \in [0,1),\ l = 0, 1, \ldots.$$

It is easy to see that this system is normal, 2-regular, and of effective cardinality $\mathcal{O}(n \log n)$ on Γ_n. This system has been employed in (Frick et al., 2014) to accelerate the computation of a multiscale inference procedure for multiple change-points detection.

(c) **The sparse systems with optimal detection power.** In one dimension, one can construct a normal system of effective cardinality $\mathcal{O}(n)$ by combining the one introduced in (Rivera and Walther, 2013), and some intervals of small scales (namely lengths $\leq \log(n)/n$)

$$\bigcup_{l=0}^{\lfloor \log_2(\log n) \rfloor} \left\{ \left[\frac{(2i-1)2^l}{n}, \frac{i2^{l+1}}{n} \right) : i = 1, \ldots, n2^{-l-1} \right\}.$$

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 (see Rivera and Walther, 2013), 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.** It has linear effective cardinality in terms of the number of measurements, i.e., $\mathcal{O}(n)$ on Γ_n, while it is normal and obviously m-regular. As we will see in Section 2.4, this system is rich enough to guarantee that MIND recovers smooth functions in a nearly optimal way (cf. Section 2.7 for the practical performance). In particular, for $m = 2$, it corresponds to the support sets of the wavelet

multiresolution scheme. The 2-partition system has been used in (Davies and Kovac, 2001; Davies et al., 2009; Pein et al., 2015) for inference of one dimensional signals.

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. Clearly this is the case if $B \cap \Gamma_n = \emptyset$ for every $B \in \mathcal{B}$, in which case $\|\cdot\|_{\mathcal{B}}$ is identically zero. However, if the system $\mathcal{B}$ is normal, 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 always assume the following.

Assumption 2. The system of cubes $\mathcal{B}$ is rich enough such that $\|\cdot\|_{\mathcal{B}}$ is a *norm*.

This allows us later to define the dual norm of $\|\cdot\|_{\mathcal{B}}$. Moreover, in the case of every system in Example 2, it is easy to see that $\|\cdot\|_{\mathcal{B}}$ is indeed a norm, and that it can be bounded from below by the maximum norm $\|\cdot\|_\infty$ on $\mathbb{R}^{\Gamma_n}$.

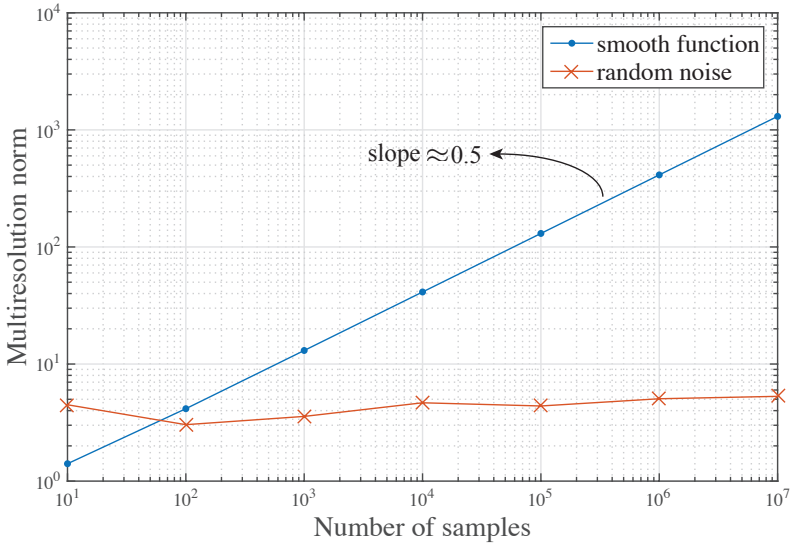

Figure 2.1.: Illustration of the growth of multiresolution norm $\|\cdot\|_{\mathcal{B}}$ with respect to 2-partition systems. The multiresolution norm of a realization of standard Gaussian random variables and that of smooth function $f(x) = x^2$ are plotted against the number of samples.

The main property of the multiresolution norm is that it allows to distinguish between

random noise and smooth functions, see Figure 2.1. 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 2.2.4. *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 (2.3) with scale parameter $\sigma > 0$. Then there exists a constant C depending only on θ such that*

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

Proof. Let $\xi_B := \left(\sum_{x\in\Gamma_n\cap B}\xi_n(x)\right)/\sqrt{\#\Gamma_n \cap B}$. Then

$$\begin{aligned}\mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \geq t\right\} &\leq \sum_{B\in\mathcal{B}} \mathbb{P}\left\{|\xi_B| \geq t\right\} \leq \sum_{B\in\mathcal{B}} \min_{\tau\geq 0} e^{-\tau t}\mathbb{E}\left[e^{\tau|\xi_B|}\right] \\ &\leq \sum_{B\in\mathcal{B}} \min_{\tau\geq 0} e^{-\tau t}\mathbb{E}\left[e^{\tau\xi_B} + e^{-\tau\xi_B}\right] \\ &\leq n^2 \min_{\tau\geq 0} 2e^{\frac{(\tau\sigma)^2}{2}-\tau t} = 2n^2 e^{-\frac{t^2}{2\sigma^2}}.\end{aligned}$$

The second result follows from the first using

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

□

Remark 2.2.5. If for every $n \in \mathbb{N}$ and $x \in \Gamma_n$ there exists a cube $B \in \mathcal{B}$ such that $x \in B$ and $\#\Gamma_n \cap B = 1$, then it is known from (Kabluchko and Munk, 2009) that

$$\limsup_{n\to\infty} \frac{\|\xi_n\|_{\mathcal{B}}}{\sqrt{2\log n}} = \mathrm{sd}(\xi_n) \quad \text{a.s.}$$

where $\mathrm{sd}(\xi_n)$ is the common standard deviation of $\xi_n(x)$ for $x \in \Gamma_n$. It follows directly that under such a condition the upper bound for the expectation given above is actually optimal in order.

Proposition 2.2.6. *Given any function $f \in \mathcal{C}([0,1]^d)$, it holds that*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}}\|S_n f\|_{\mathcal{B}} = \sup_{B\in\mathcal{B}} \frac{1}{\sqrt{|B|}}\Big|\int_B f(x)dx\Big|.$$

Proof. Let $m := n^{1/d}$, and $B_n := B \cap \Gamma_n + [-\frac{1}{2m}, \frac{1}{2m})^d$ for every $B \in \mathcal{B}$. It can be easily shown that

$$\lim_{n\to\infty} \sup_{B\in\mathcal{B},\, B\cap\Gamma_n\neq\emptyset} \frac{1}{\sqrt{|B_n|}}\Big|\int_{B_n} f(x)dx\Big| = \sup_{B\in\mathcal{B}} \frac{1}{\sqrt{|B|}}\Big|\int_{B} f(x)dx\Big|. \tag{2.9}$$

Given any $\varepsilon > 0$, the uniform continuity of f implies that for sufficiently large n

$$\Big|\int_{x+[-\frac{1}{2m},\frac{1}{2m})^d} f(z)dz - \frac{1}{n}f(x)\Big| \le \frac{\varepsilon}{n} \qquad \text{for every } x \in \Gamma_n.$$

It follows that for large enough n

$$\begin{aligned}
&\Big|\sup_{B\in\mathcal{B},\, B\cap\Gamma_n\neq\emptyset} \frac{1}{\sqrt{|B_n|}}\Big|\int_{B_n} f(z)dz\Big| - \frac{1}{\sqrt{n}} \sup_{B\in\mathcal{B},\, B\cap\Gamma_n\neq\emptyset} \frac{1}{\sqrt{\#B\cap\Gamma_n}}\Big|\sum_{x\in B\cap\Gamma_n} f(x)\Big|\Big| \\
&\le \sup_{B\in\mathcal{B},\, B\cap\Gamma_n\neq\emptyset} \frac{1}{\sqrt{|B_n|}}\Big|\int_{B_n} f(z)dz - \frac{1}{n}\sum_{x\in B\cap\Gamma_n} f(x)\Big| \\
&\le \sup_{B\in\mathcal{B},\, B\cap\Gamma_n\neq\emptyset} \frac{1}{\sqrt{|B_n|}} \sum_{x\in B\cap\Gamma_n} \Big|\int_{x+[-\frac{1}{2m},\frac{1}{2m})^d} f(z)dz - \frac{1}{n}f(x)\Big| \le \varepsilon.
\end{aligned}$$

This together with (2.9) completes the proof. □

The next result provides an interpolation inequality for the L^q-norm of a function in terms of its multiresolution norm and the L^p-norm of its k-th order derivative. For k, l, $d \in \mathbb{N}$ and $1 \le p, q \le \infty$, we introduce

$$\vartheta_l \equiv \vartheta_l(k,d,p,q) := \begin{cases} \frac{k-l}{2k+d} & \text{if } \frac{q}{p} \le \frac{2k+d}{2l+d}, \\ \frac{k-l-d/p+d/q}{2k+d-2d/p} & \text{if } \frac{q}{p} \ge \frac{2k+d}{2l+d}; \end{cases} \tag{2.10}$$

and

$$\vartheta_l' \equiv \vartheta_l'(k,d,p,q) := \Big(\frac{2k}{d} + 1 - \frac{2}{p}\Big)\vartheta_l(k,d,p,q). \tag{2.11}$$

Theorem 2.2.7. *Let $\mathcal{B}$ be a c-normal system of cubes, and assume that $1 \le p, q \le \infty$, $l \in \{0, \dots, k-1\}$ and $k, d \in \mathbb{N}$ satisfy either $k > d/p$ or $k = d$ and $p = 1$. Then there are constants C and n_0, both depending on c, k, d and p only, such that for every $f \in W^{k,p}([0,1]^d)$ and for $n \ge n_0$,*

$$\|D^l f\|_{L^q} \le C \max\Big\{\frac{\|S_n f\|_{\mathcal{B}}^{2\vartheta_l}}{n^{\vartheta_l}} \|D^k f\|_{L^p}^{1-2\vartheta_l}, \frac{\|S_n f\|_{\mathcal{B}}}{n^{1/2}}, \frac{\|D^k f\|_{L^p}}{n^{\vartheta_l'}}\Big\}, \tag{2.12}$$

where $\vartheta_l = \vartheta_l(k,d,p,q)$ is given by (2.10) and $\vartheta_l' = \vartheta_l'(k,d,p,q)$ by (2.11).
Proof. See Appendix A.1. □

Remark 2.2.8. This is an extension of the result in (Nemirovski, 1985), where (2.12) was shown to hold for c-normal systems with $c = 6$, and $p > d$ or $p = d = 1$. It is known that $k > d/p$ or $k = d$ and $p = 1$ is the weakest condition to guarantee the continuity of $f \in W^{k,p}([0,1]^d)$ (cf. Adams and Fournier, 2003, Theorem 4.12), which is required for the definitions of the evaluation S_n and the multiresolution norm $\|\cdot\|_{\mathcal{B}}$. From this perspective, this result is already in its most general form.

2.3. Asymptotics under abstract smoothness assumptions

For the asymptotic analysis 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 regard the point evaluation 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 with equidistant knots Γ_n.

Therefore, we use a different, but related, approach based on *approximate source conditions* (see Hofmann and Yamamoto, 2005; Hofmann, 2006). The idea here 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.

(i) 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.

(ii) 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.

2.3.1. Multiscale distance functions

Recall that the multiresolution norm $\|\cdot\|_{\mathcal{B}}$ is indeed a norm (cf. Assumption 2). Thus, we can consider its dual norm $\|\cdot\|_{\mathcal{B}^*}$ 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 (2.8) it readily follows that for proper real numbers $(c_B)_{B \in \mathcal{B}}$

$$\|\omega\|_{\mathcal{B}^*} = \min\Big\{ \sum_{B \in \mathcal{B}} |c_B| \sqrt{\#\Gamma_n \cap B}; \omega(x) = \sum_{B \ni x} c_B \text{ for all } x \in \Gamma_n \Big\}.$$

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 2.3.1. Given $n \in \mathbb{N}$ and $t \geq 0$, the *multiscale distance function* $d_n(t)$ for f is defined as

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

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 what follows we will provide some description of the mapping S_n^*. We first 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

$$\varphi_x := S_n^* e_x \in H_0^k(\mathbb{T}^d).$$

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

$$(-1)^k \Delta^k \varphi_x = \delta_x - 1. \tag{2.13}$$

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

$$d_n(t) = \min\Big\{ \|f - \sum_{x \in \Gamma_n} c_x \varphi_x\|_{H_0^k}; \|(c_x)_{x \in \Gamma_n}\|_{\mathcal{B}^*} \le t \Big\}. \tag{2.14}$$

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

$$d_n(t) = \min\Big\{ \|f - \sum_{B \in \mathcal{B}} c_B \varphi_B\|_{H_0^k}; \sum_{B \in \mathcal{B}} |c_B| \sqrt{\#\Gamma_n \cap B} \le t \Big\}. \tag{2.15}$$

Remark 2.3.2. By means of Fourier series expansions, one can derive a solution in series form to the equation (2.13), which is given by

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

In addition, it is worth noting that

$$\varphi_x(z) = R(x,z) \qquad \text{for every } x \in \Gamma_n \text{ and } z \in \mathbb{T}^d,$$

where $R(\cdot,\cdot)$ is the *reproducing kernel* of the Hilbert space $H_0^k(\mathbb{T}^d)$. Finally, we point out that equations (2.14) and (2.15) translate the behavior of multiscale distance functions $d_n(t)$ into the approximation property of bases $\big(\varphi_x\big)_{x\in\Gamma_n}$ and frames $\big(\varphi_B\big)_{B\in\mathcal{B}}$, respectively.

2.3.2. Abstract convergence rates

We will derive the convergence rates of the MIND estimator $\hat{f}_{\gamma_n}$, which is defined as the solution of the optimization problem given in (2.4), in terms of multiscale distance functions $d_n(t)$ (see Definition 2.3.1). 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 2.3.3. *Let $l \in \{0,\dots,k-1\}$, k, $d \in \mathbb{N}$, $k > d/2$ and $1 \le q \le \infty$. Assume that $\mathcal{B}$ is c-normal and the inequality*

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

is satisfied, and denote by $\hat{f}_{\gamma_n}$ the MIND estimator (2.4). In addition, define

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

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

$$\|D^l \hat{f}_{\gamma_n} - D^l f\|_{L^q} \le C \max\Big\{\frac{\gamma_n^{2\vartheta_l} c_n^{1-2\vartheta_l}}{n^{\vartheta_l}},\, \frac{\gamma_n}{n^{1/2}},\, \frac{c_n}{n^{\vartheta_l'}}\Big\} \quad \text{for } n \ge n_0, \tag{2.16}$$

where $\vartheta_l = \vartheta_l(k,d,2,q)$ is given by (2.10) and $\vartheta_l' = \vartheta_l'(k,d,2,q)$ by (2.11).

Proof. See Appendix A.2.1. □

Note that the estimate (2.16) provides error bounds for estimating the function f and its derivatives $D^\alpha f$ for $\alpha \in \mathbb{N}_0^d$ and $0 < |\alpha| < k$. 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 (see Proposition 2.2.4), we obtain an asymptotic convergence rate almost surely for the MIND estimator.

Corollary 2.3.4. *Let $l \in \{0, \dots, k-1\}$, k, $d \in \mathbb{N}$, $k > d/2$ and $1 \leq q \leq \infty$. Assume that $\mathcal{B}$ is normal, that γ_n is chosen as in (2.5), and that*

$$\min_{t \geq 0} \left(d_n(t) + (\log n)^{r/2} t^{1/2} \right) = \mathcal{O}\left(n^{-\mu}\right) \qquad \text{as } n \to \infty \tag{2.17}$$

for some $0 \leq \mu < 1/2$. Then there exists a constant C such that the MIND estimator $\hat{f}_{\gamma_n}$ in (2.4) satisfies the estimate

$$\limsup_{n \to \infty} \left(n^{\mu(1-2\vartheta_l)+\vartheta_l} (\log n)^{-2r\vartheta_l} \| D^l \hat{f}_{\gamma_n} - D^l f \|_{L^q} \right) \leq C \qquad \text{a.s.} \tag{2.18}$$

with $\vartheta_l = \vartheta_l(k, d, 2, q)$ in (2.10).

Proof. With the given choice of γ_n, Proposition 2.2.4 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 2.3.3 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 (2.16) is always dominant, we obtain (2.18). □

The condition (2.17) is often referred to as the *approximate source condition*. It encodes the smoothness of the truth f essentially by how fast the function $d_n(t_n)$ decays, for some proper choice of t_n, as the number of samples n tends to infinity. In other words, the smoothness of the truth f is measured by the asymptotic closeness between f and functions in $\{S_n^* w; \|w\|_{\mathcal{B}^*} \leq t_n\}$ with respect to the homogeneous Sobolev norm.

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 2.3.3 does not apply. Thus it is relegated to the appendix.

Theorem 2.3.5. *Assume the same setting as Corollary 2.3.4. Then the MIND estimator $\hat{f}_{\gamma_n}$ in (2.4) satisfies*

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

as $n \to \infty$, with $\vartheta_l = \vartheta_l(k, d, 2, q)$ given in (2.10).

Proof. See Appendix A.2.2. □

The convergence rates in Corollary 2.3.4 and Theorem 2.3.5 are somewhat "abstract" in the sense that they rely on the approximate source condition (2.17). The merit here is, however, to transform the statistical convergence analysis of MIND into a deterministic approximation problem in terms of $d_n(t)$. As a consequence, the following example provides concrete convergence rates based on a simple upper bound of the multiscale distance functions $d_n(t)$.

Example 3 (Proper smoothing). Assume that

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

It readily follows that

$$\min_{t\geq 0}\big(d_n(t) + (\log n)^{r/2} t^{1/2}\big) \leq d_n(0) = \|D^k f\|_{L^2} \leq \|f\|_{W_0^{k,p}}.$$

Therefore, by Corollary 2.3.4 and Theorem 2.3.5, we obtain for MIND

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

almost surely and in expectation, where $\vartheta_l = \vartheta_l(k,d,2,q)$ is given in (2.10). In particular, in the case of $2 < p \leq \infty$, $1 \leq q \leq \frac{2k+d}{2l+d}p$, and of $p = 2$, $1 \leq q \leq \infty$, this rate

$$\|D^l \hat{f}_{\gamma_n} - D^l f\|_{L^q} = \mathcal{O}\left(n^{-\frac{k-l}{2k+d}}(\log n)^{\frac{2r(k-l)}{2k+d}}\right)$$

actually coincides with the minimax optimal rate up to a log-factor over Sobolev ellipsoids in $W_0^{k,p}(\mathbb{T}^d)$ for estimating the l-th order derivatives with $l \in \{0,\ldots,k-1\}$ (see Nemirovski, 1985, for the minimax optimal rates).

Example 3 essentially shows that the MIND estimator is nearly minimax optimal for those functions of the same smoothness as required by the regularization term $\|D^k \cdot\|_{L^2}$. One natural question arises whether it is possible for MIND with a fixed k to achieve faster or even nearly optimal rates for functions of higher order smoothness. To answer it, we need better estimates of the multiscale distance functions $d_n(t)$. As mentioned before, this relates to the approximation property of the bases $\{\varphi_x; x \in \Gamma_n\}$, or the frames $\{\varphi_B; B \in \mathcal{B}\}$, with the size of coefficients controlled in certain sense, see (2.14) and (2.15). In one dimension, we are able to derive sharp error bounds for such approximation problem, using the well-developed theory of B-splines, and give affirmative answer to the previous question, see the next section. However, in higher dimensions, the approximation problems (2.14) and (2.15) remain still *open*. Note that there exist some results on the approximation error of bases $\{\varphi_x; x \in \Gamma_n\}$ (see Dyn et al., 1999; Narcowich et al., 2002, 2003), but we are not aware of any results about the size of the coefficients.

We conclude this section with some discussion on the possibility of dropping the zero mean requirement, i.e. Assumption 1(b). The idea is to first compute the MIND estimator with zero mean from centered data; and then we adjust it by the mean of the data. To be precise, we consider the *modified* MIND estimator $\hat{f}_{\gamma_n}$ given by

$$\begin{aligned} \hat{f}^0_{\gamma_n} &:= \operatorname*{arg\,min}_{f \in H_0^k(\mathbb{T}^d)} \frac{1}{2}\|D^k f\|_{L^2}^2 \quad \text{subject to } \|S_n f - (y_n - \bar{y}_n)\|_{\mathcal{B}} \leq \gamma_n, \\ \hat{f}_{\gamma_n} &:= \hat{f}^0_{\gamma_n} + \bar{y}_n, \end{aligned} \tag{2.19}$$

where $\bar{y}_n := \sum_{x\in\Gamma_n} y_n(x)/n$. Under certain conditions, we are able to obtain the same results as above for this modified estimator.

Proposition 2.3.6. *Under Assumptions 1(a), 1(c), and 2, let $d = 1$ or 2, $k \in \mathbb{N}$, $k > d/2$, $l \in \{0,\dots,k-1\}$, and $1 \le q \le \infty$. Assume that the truth $f \in \mathcal{C}^1(\mathbb{T}^d) \cap H^k(\mathbb{T}^d)$, that $\mathcal{B}$ is normal, and that $\hat{f}_{\gamma_n}$ is the modified MIND estimator in (2.19). In addition, define $\vartheta_l := \vartheta_l(k,d,2,q)$ in (2.10), $\vartheta_l' := \vartheta_l'(k,d,2,q)$ in (2.11), and*

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

(i) If the inequality $\|\xi_n\|_{\mathcal{B}} = \|S_n f - y_n\|_{\mathcal{B}} \le \gamma_n$ holds, we have

$$\|D^l \hat{f}_{\gamma_n} - D^l f\|_{L^q} = \mathcal{O}\Big(\max\Big\{\frac{\gamma_n^{2\vartheta_l} c_n^{1-2\vartheta_l}}{n^{\vartheta_l}}, \frac{\gamma_n}{n^{1/2}}, \frac{c_n}{n^{\vartheta_l'}}\Big\}\Big) \qquad \text{as } n \to \infty.$$

(ii) If γ_n is chosen as in (2.5), and $c_n = \mathcal{O}(n^{-\mu})$ for some $0 \le \mu < 1/2$, it holds almost surely and in expectation that

$$\mathbb{E}\left[\|D^l \hat{f}_{\gamma_n} - D^l f\|_{L^q}\right] = \mathcal{O}\left(n^{-\mu(1-2\vartheta_l)-\vartheta_l}(\log n)^{2r\vartheta_l}\right) \qquad \text{as } n \to \infty.$$

Proof. See Appendix A.2.3. □

Remark 2.3.7. We note from the proof that the requirement $f \in \mathcal{C}^1(\mathbb{T}^d)$ is only needed for the control of approximation error

$$\Big|\int_{\mathbb{T}^d} f(z)dz - \frac{1}{n}\sum_{x\in\Gamma_n} f(x)\Big| = \mathcal{O}\Big(\frac{1}{n^{1/d}}\Big). \tag{2.20}$$

This even holds under the weaker assumption that f has bounded variation in the sense of Hardy and Krause (see Kuipers and Niederreiter, 1974, Chapter 2, Theorem 5.5). Moreover, we have to restrict ourselves to the cases of $d = 1, 2$ because we want the approximation error (2.20) to be no slower than the parametric estimation rate $1/\sqrt{n}$. In higher dimensions $d \ge 3$, this is possible for some grid on $[0,1]^d$ other than Γ_n. For instance, the left hand side of (2.20) with instead the average on the *Hammersley* grid is of order $n^{-1}\log^{d-1} n$, see e.g. (Davis and Rabinowitz, 1984, Section 5.5) for details.

2.4. Examples in one dimension

In this section we will apply the general results of the previous section to the particular setting of nonparametric regression of one-dimensional (periodic) functions. Here it is possible to translate the approximate source conditions introduced previously into conditions concerning the Besov or Sobolev smoothness of the function f to be estimated.

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 rich. 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.

2.4.1. Dual operator, reproducing kernel and splines

We start with some notation. Given $m \in \mathbb{N}$, by $\mathcal{P}_m$ we denote the space of polynomials of order m (or equivalently, of degree $\leq m-1$), 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

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

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 about splines can be found in (Wahba, 1990; Schumaker, 2007) for example.

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

$$(-1)^k \varphi_{i,n}^{(2k)} = \delta\Big(\frac{i}{n} - \cdot \Big) - 1,\ \varphi_{i,n} \in H_0^k(\mathbb{T}). \tag{2.21}$$

As demonstrated in Section 2.3.1, it follows that $S_n^* e_{i/n} = \varphi_{i,n}$. We recall from Remark 2.3.2 that the reproducing kernel $R(\cdot,\cdot)$ for $H_0^k(\mathbb{T})$ (see also Wahba, 1990, Section 2.1) is

$$R(x,z) = \sum_{\lambda=1}^{\infty} \frac{2}{(2\pi\lambda)^{2k}} \cos\big(2\pi\lambda(x-z)\big).$$

It relates to the *periodic Bernoulli polynomial* B_{2k} of degree $2k$ (see e.g. Kress, 1998, Section 9.4) via

$$R(x,z) = (-1)^{k-1} B_{2k}(x-z),$$

where the periodic Bernoulli polynomial B_m of degree m is defined recursively by

$$B_0(x) = 1 \quad \text{and } B_m'(x) = B_{m-1}(x) \quad \text{for } x \in \mathbb{T} \text{ and } m \in \mathbb{N}$$

with the normalization condition $\int_{\mathbb{T}} B_m(x)dx = 0$, $m \in \mathbb{N}$. It readily implies that

$$\varphi_{i,n}(x) = R\Big(x, \frac{i}{n}\Big) = (-1)^{k-1} B_{2k}\Big(x - \frac{i}{n}\Big) \qquad \text{for } x \in \mathbb{T}.$$

We will further 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

$$\chi_n(z) := \begin{cases} 1, & \text{if } 0 \le z < \frac{1}{n}, \\ 0, & \text{if } \frac{1}{n} \le z < 1. \end{cases}$$

By integrating both sides of (2.21) 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

$$\begin{aligned} \psi_{i,n}^m(z) &:= (-1)^k n^{m-1} D_{\frac{1}{n},-}^m \varphi_{i,n}^{(2k-m)}(z) \\ &= n^{m-1} (\chi_n *^{m-1} \chi_n)\Big(z - \frac{i}{n}\Big) - \frac{1}{n} = Q_i^m(z) - \frac{1}{n} \end{aligned} \tag{2.22}$$

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

$$D_{1/n,-}\varphi_{i,n}^{(l)}(z) = \varphi_{i,n}^{(l)}(z) - \varphi_{i,n}^{(l)}\Big(z - \frac{1}{n}\Big) = \varphi_{i,n}^{(l)}(z) - \varphi_{i+1,n}^{(l)}(z) \qquad \text{for } l \in \mathbb{N}_0,$$

we see that

$$\begin{aligned} h =& (-1)^k n^{m-1} \sum_{i=0}^{n-1} \tilde{c}_i D_{\frac{1}{n},-}^m \varphi_{i,n}^{(2k-m)} \\ =& (-1)^k n^{m-1} \sum_{i=0}^{n-1} \tilde{c}_i \Big(D_{\frac{1}{n},-}^{m-1} \varphi_{i,n}^{(2k-m)} - D_{\frac{1}{n},-}^{m-1} \varphi_{i+1,n}^{(2k-m)} \Big) \\ =& (-1)^k n^{m-1} \sum_{i=0}^{n-1} (D_- \tilde{c})_i D_{\frac{1}{n},-}^{m-1} \varphi_{i,n}^{(2k-m)}. \end{aligned}$$

Repeating this argumentation m times, we obtain

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

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

2.4.2. Convergence rates for Sobolev/Besov classes

We now derive the main results of this chapter, where we prove convergence rates in the one-dimensional case for f contained in various Sobolev and Besov spaces (cf. Section 2.1.1).

Under-smoothing

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 $\|D^k \hat{f}\|_{L^2}^2$. In this case, it turns out that indeed a higher order convergence

rate is obtained than the one discussed in Example 3. For this to hold, however, we have to assume that the system of intervals $\mathcal{B}$ is regular (see Definition 2.2.3), 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 2.4.1. *Assume that $d = 1$, $r \geq 1/2$, $k \in \mathbb{N}$, that $\mathcal{B}$ is regular, and that*

$$f \in B^{s,p'}_{p,0}(\mathbb{T}) \qquad \text{for some } s \in [k+1, 2k] \text{ and } p, p' \in [1, \infty].$$

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

$$\mu = \frac{s - k - \left(\frac{1}{p} - \frac{1}{2}\right)_+}{2s + 1 - 2\left(\frac{1}{p} - \frac{1}{2}\right)_+}. \tag{2.23}$$

The same result holds for $f \in W^{s,p}_0(\mathbb{T})$ with $k+1 \leq s \leq 2k$ and $1 \leq p \leq \infty$.

Proof. See Appendix A.3.3. □

Theorem 2.4.2. *Assume that $d = 1$, $l \in \{0, \dots, k-1\}$, $k \in \mathbb{N}$, that $\mathcal{B}$ is regular, and that*

$$f \in B^{s,p'}_{p,0}(\mathbb{T}) \qquad \text{for some } s \in [k+1, 2k] \text{ and } p, p' \in [1, \infty].$$

Then the MIND estimator $\hat{f}_{\gamma_n}$ satisfies, with a parameter choice γ_n given by (2.5),

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

almost surely and in expectation, with $\vartheta_l = \vartheta_l(k, 1, 2, q)$ given in (2.10) and μ in (2.23). The same result holds for $f \in W^{s,p}_0(\mathbb{T})$ with $k+1 \leq s \leq 2k$ and $1 \leq p \leq \infty$.

Proof. This is a direct consequence of Proposition 2.4.1, Corollary 2.3.4, and Theorem 2.3.5□

Remark 2.4.3. Note that the rate obtained in the previous result greatly simplifies in the case where $p \geq 2$ and $q \leq \frac{4k+2}{2l+1}$. Then, a short computation shows that it can be written as

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

Over-smoothing

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 $f \in H^k_0(\mathbb{T})$. 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 2.4.4. *Let $\mathcal{B}$ be normal, $d = 1$, $k \in \mathbb{N}$, 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 (2.4) with the homogeneous Sobolev norm $\|D^k \cdot\|_{L^2}$, and the threshold γ_n in (2.5). Then it holds almost surely and in expectation that

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

for every $\epsilon > \frac{(k-l)(2r-1)}{2k+1}$ with r in (2.5), and for every $l \in \{0, \dots, \lfloor s \rfloor - 1\}$ and $q \in \left[1, \frac{4k+2}{2l+1}\right]$.

Proof. See Appendix A.3.4. □

Remark 2.4.5. For simplicity, the convergence rates results of Theorems 2.4.2 and 2.4.4 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. For instance, it can be shown that the constant in Theorem 2.4.2 can, in the Besov space case, be written in the form

$$C \|f\|_{B_{p,0}^{k+s,p'}}^{1-2\vartheta_l} \qquad \text{with } C > 0 \text{ only depending on } k,\ s,\ p, \text{ and } \mathcal{B},$$

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.

2.4.3. Minimax optimality and partial adaptation

Given a class $\mathcal{F}$ of functions, we define the minimax L^q-risk of nonparametric regression (2.1) of the l-th order derivative over $\mathcal{F}$ by

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

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 particular, when $l = 0$, it is the minimax L^q-risk of nonparametric regression (2.1) of the function f itself.

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 $\mathcal{E}_{q,l}$. To that end, we introduce, for $s \geq 0$, $1 \leq p \leq \infty$, the Sobolev ball of radius $L > 0$ by

$$S_L^{s,p} := \left\{ f \in W_0^{s,p}(\mathbb{T}) : \|f\|_{W_0^{s,p}} \leq L \right\}, \tag{2.24}$$

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

$$B_L^{s,p,p'} := \left\{ f \in B_{p,0}^{s,p'}(\mathbb{T}) : \|f\|_{B_{p,0}^{s,p'}} \leq L \right\}. \tag{2.25}$$

In (Nemirovski, 1985) it has been shown that, for $s \in \mathbb{N}$, $l \in \{0, \dots, s-1\}$, and n sufficiently large, there exists a constant $C > 0$ depending only on s such that

$$\mathcal{E}_{q,l}(n; S_L^{s,p}) \geq C \begin{cases} \left(\frac{\sigma^2}{n}\right)^{\beta} L^{1-2\beta} & \text{if } q < \frac{2s+1}{2l+1}p \quad \text{(regular zone)}, \\ \left(\frac{\sigma^2 \log n}{n}\right)^{\beta} L^{1-2\beta} & \text{if } q \geq \frac{2s+1}{2l+1}p \quad \text{(logarithmic zone)}, \end{cases} \tag{2.26}$$

where $\beta = \beta(k, l, p, q) := \vartheta_l(k, 1, p, q)$ given in (2.10).

Similar to the proof of lower bounds in (Nemirovski, 1985), one can show that this result (2.26) still holds for non-integer $s > 1/p$ or $s = p = 1$, and also for all the Besov balls $B_L^{s,p,p'}$ with $s > 1/p$ or $s = p = 1$, whenever $0 \leq l \leq \lfloor s \rfloor - 1$. Even more, in the case of $q = \frac{2s+1}{2l+1}p$ (critical zone), the lower bound can be tightened to include the logarithmic factor

$$(\log n)^{\frac{1}{q}\left(1 - \frac{p}{\min\{p, p'\}}\right)_+}$$

see (Donoho et al., 1996, Theorem 1) for details.

Partial adaptation

Comparing these minimax L^q-risks with the convergence rates of MIND in Theorems 2.4.2 and 2.4.4, and Example 3, we see that, for $l \in \left\{0, \dots, \min\{k, \lfloor s \rfloor\} - 1\right\}$ and $1 \leq q \leq \frac{4k+2}{2l+1}$, 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 2.2). 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 proposed method MIND automatically adapts to the smoothness of the function f independent of our prior guess. Note further that the adaptation range of MIND scales with the smoothness order of regularization k.

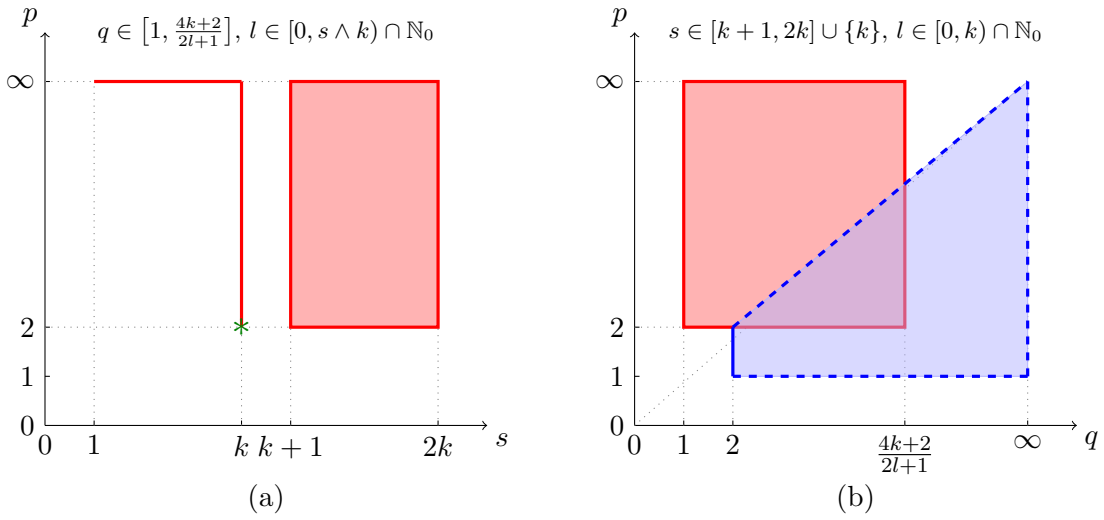

Figure 2.2.: Adaptive minimax optimality of the MIND estimator in (2.4) with the homogeneous Sobolev norm of order k, over balls in $W_0^{s,p}(\mathbb{T})$ or $B_{p,0}^{s,p'}(\mathbb{T})$ in terms of $\|D^l \cdot\|_{L^q}$-loss. (a) For $l \in \{0, \dots, \min\{\lfloor s \rfloor, k\} - 1\}$ and $q \in [1, \frac{4k+2}{2l+1}]$, MIND attains minimax optimal rates up to a log-factor, simultaneously in all classes $S_L^{s,p}$ or $B_L^{s,p,p'}$ with smoothness parameters s, p within the red region ("partial adaptation"). By contrast, the Nemirovski's estimator $\hat{f}_{2,\eta}$ in (1.4) is minimax optimal up to a log-factor only for $S_L^{k,2}$, marked by a green asterisk. (b) For $l \in \{0, \dots, k-1\}$ and $s \in [k+1, 2k] \cup \{k\}$, MIND is minimax optimal up to a log-factor over $S_L^{s,p}$ or $B_L^{s,p,p'}$ with parameters q, p within the red region. Note that no linear estimator is optimal for parameters q, p in the blue region.

Remark 2.4.6. For the estimation of function $f \in W_0^{s,p}(\mathbb{T})$ or $B_{p,0}^{s,p'}(\mathbb{T})$, we have evaluated the performance of MIND in (2.4) with respect to $\|D^l \cdot\|_{L^q}$-loss for $l \in [0, \min\{s, k\}) \cap \mathbb{N}_0$ and $q \in [1, \infty]$ so far. We do note that it is possible to extend it to general Sobolev $\|\cdot\|_{W_0^{r,q}}$ or Besov $\|\cdot\|_{B_{q,0}^{r,q'}}$ losses, with $0 \leq r < s - (1/p - 1/q)_+$ and $1 \leq q, q' \leq \infty$, by means of interpolation relations between Sobolev and Besov spaces, see e.g. Chapter 3. We conjecture that the corresponding convergence rates will still be minimax optimal up to a log-factor, under the same smoothness assumption as above, i.e. $(s, p) \in [1, k] \times \{\infty\} \cup$

$\{k\} \times [2, \infty] \cup [k+1, 2k] \times [2, \infty]$.

2.5. Penalized MIND and Lepskiĭ principle

As an alternative, we now consider a penalized version of the MIND estimator for the nonparametric regression problem (2.1). More precisely, we study the estimator $\hat{f}_\alpha$ given by

$$\hat{f}_\alpha := \operatorname*{arg\,min}_{f \in H_0^k(\mathbb{T}^d)} \|S_n f - y_n\|_{\mathcal{B}} + \frac{\alpha}{2} \|D^k f\|_{L^2}^2, \tag{2.27}$$

which we call *penMIND* for abbreviation. This is a special case of (1.12) with $T = \mathrm{I}$, and also known as *Tikhonov regularization* as mentioned in Chapter 1. We note that the strict convexity and coercivity of the regularization functional implies that $\hat{f}_\alpha$ exists and is unique for every data $y_n \in \mathbb{R}^{\Gamma_n}$. In this section, we will derive nearly the same results for penMIND as those for MIND in Sections 2.3 and 2.4.

We start by a general upper bound on the L^q-loss of $\hat{f}_\alpha$ for estimating f and its derivatives in terms of the noise ξ_n and the multiscale distance function $d_n(t)$ (cf. Definition 2.3.1).

Theorem 2.5.1. *Assume that $l \in \{0, \dots, k-1\}$, $k, d \in \mathbb{N}$, $k > d/2$ and $1 \le q \le \infty$, and that $\mathcal{B}$ is c-normal. Then for the penMIND estimator $\hat{f}_\alpha$ in (2.27) there exist constants $C > 0$, and $n_0 \in \mathbb{N}$, both depending only on c, k and d, such that for $n \ge n_0$*

$$\|D^l \hat{f}_\alpha - D^l f\|_{L^q} \le C \max\Bigg\{ \frac{\alpha^{2\vartheta_l} d_n(\frac{1}{2\alpha})^{1+2\vartheta_l}}{n^{\vartheta_l}} + \frac{\|\xi_n\|_{\mathcal{B}}^{1/2+\vartheta_l}}{\alpha^{1/2-\vartheta_l} n^{\vartheta_l}}; \\ \frac{\alpha d_n(\frac{1}{2\alpha})^2}{\sqrt{n}} + \frac{\|\xi_n\|_{\mathcal{B}}}{\sqrt{n}}; \frac{d_n(\frac{1}{2\alpha})}{n^{\vartheta_l'}} + \frac{(\|\xi_n\|_{\mathcal{B}})^{1/2}}{n^{\vartheta_l'} \sqrt{\alpha}} \Bigg\},$$

where $\vartheta_l = \vartheta_l(k, d, 2, q)$ is given by (2.10) and $\vartheta_l' = \vartheta_l'(k, d, 2, q)$ by (2.11).
Proof. See Appendix A.4 □

We note that the previous error estimate is deterministic in the sense that the error estimate takes into account the actually realized noise level $\|\xi_n\|_{\mathcal{B}}$. Since the estimate holds independent of the size of the error, it is, however, easy to obtain statistical estimates and also convergence rates if one additionally postulates some behavior of the multiscale distance function d_n.

Corollary 2.5.2. *Let $l \in \{0, \dots, k-1\}$, $k, d \in \mathbb{N}$, $k > d/2$ and $1 \le q \le \infty$. Assume that $\mathcal{B}$ is normal, and that*

$$\min_{t \ge 0} \big(d_n(t) + (\log n)^{1/4} t^{1/2} \big) = \mathcal{O}\left(n^{-\mu}\right) \tag{2.28}$$

for some $0 \leq \mu < 1/2$. *Then for a parameter choice*

$$\alpha \sim n^{2\mu}\sqrt{\log n}$$

the penMIND estimator $\hat{f}_\alpha$ *in* (2.27) *satisfies that*

$$\|D^l \hat{f}_\alpha - D^l f\|_{L^q} = \mathcal{O}\left(n^{-\mu(1-2\vartheta_l)-\vartheta_l}(\log n)^{\vartheta_l}\right) \qquad \text{as } n \to \infty,$$

both almost surely and in expectation, with $\vartheta_l = \vartheta_l(k, d, 2, q)$ *given in* (2.10).

Proof. This follows directly from Theorem 2.5.1 and the fact in Proposition 2.2.4 that

$$\mathbb{E}\left[\|\xi_n\|_{\mathcal{B}}^\theta\right] = \mathcal{O}\left((\log n)^{\frac{\theta}{2}}\right) \qquad \text{and } \lim_{n\to\infty} \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} > (2+\epsilon)\sigma\sqrt{\log n}\right\} = 0,$$

for every $\theta > 0$ and every $\epsilon > 0$. □

We stress that the choice of α above is the one that gives the minimal error bound for the L^q-loss as n goes to infinity. The assumption (2.28) on the multiscale distance function, as mentioned already in Section 2.3.2, is known as the approximate source condition, which encodes the smoothness assumption of the truth. And it is actually a special case of (2.17) there with $r = 1/2$. Note further that this corollary provides exactly the same rate for penMIND as Corollary 2.3.4 and Theorem 2.3.5 for MIND. In particular, in the case considered by Example 3 (i.e. proper smoothing), the penMIND estimator $\hat{f}_\alpha$ in (2.27) with $\alpha \sim \sqrt{\log n}$ also attains the nearly minimax rate for estimating the l-th order derivative of $f \in W_0^{k,p}(\mathbb{T}^d)$ in terms of L^q-loss with either $2 < p \leq \infty$, $1 \leq q \leq \frac{2k+d}{2l+d}p$ or $p = 2$, $1 \leq q \leq \infty$. This follows readily from Corollary 2.5.2 with $\mu = 0$.

2.5.1. Lepskiĭ balancing principle

One problem of penMIND is that of parameter selection. If one knows the smoothness class of the truth f (which is encoded in the multiscale distance function d_n) in advance, then one can define the penalization parameter α in such a way that the estimates become optimal in rate (cf. Corollary 2.5.2). In general, however, the precise smoothness class is unknown but has to be estimated and therefore the best convergence rates are difficult to obtain in practice. Now we will discuss an a-posteriori parameter choice rule, which does not require such a priori knowledge.

Let us consider the case that noise ξ_n satisfies

$$\|\xi_n\|_{\mathcal{B}} \leq \theta\sigma\sqrt{\log n} \qquad \text{with some } \theta > \sqrt{6 + \frac{2k}{d}}. \tag{2.29}$$

From Theorem 2.5.1, we further have that for $n \geq n_0$

$$\|D^l\hat{f}_\alpha - D^l f\|_{L^q} \le C_0 \max\Big\{ \frac{\alpha^{2\vartheta_l} d_n(\frac{1}{2\alpha})^{1+2\vartheta_l}}{n^{\vartheta_l}} + \frac{(\sigma^2\log n)^{(1+2\vartheta_l)/4}}{\alpha^{1/2-\vartheta_l} n^{\vartheta_l}}; \\ \frac{\alpha d_n(\frac{1}{2\alpha})^2}{\sqrt{n}} + \frac{\sqrt{\sigma^2\log n}}{\sqrt{n}};\ \frac{d_n(\frac{1}{2\alpha})}{n^{\vartheta'_l}} + \frac{(\sigma^2\log n)^{1/4}}{n^{\vartheta'_l}\sqrt{\alpha}} \Big\}, \quad (2.30)$$

where $C_0 = C_0(k, d, \theta, \mathcal{B})$ and $n_0 = n_0(k, d, \mathcal{B})$ are some constants. For abbreviation, we denote each term in the maximum by $\Phi_{i,n}(\alpha) + \Psi_{i,n}(\alpha)$, $i = 1, 2, 3$, with

$$\Phi_{1,n}(\alpha) := \frac{\alpha^{2\vartheta_l} d_n(\frac{1}{2\alpha})^{1+2\vartheta_l}}{n^{\vartheta_l}}, \qquad \Psi_{1,n}(\alpha) := \frac{(\sigma^2\log n)^{(1+2\vartheta_l)/4}}{\alpha^{1/2-\vartheta_l} n^{\vartheta_l}},$$
$$\Phi_{2,n}(\alpha) := \frac{\alpha d_n(\frac{1}{2\alpha})^2}{\sqrt{n}}, \qquad \Psi_{2,n}(\alpha) := \frac{\sqrt{\sigma^2\log n}}{\sqrt{n}},$$
$$\Phi_{3,n}(\alpha) := \frac{d_n(\frac{1}{2\alpha})}{n^{\vartheta'_l}}, \qquad \Psi_{3,n}(\alpha) := \frac{(\sigma^2\log n)^{1/4}}{n^{\vartheta'_l}\sqrt{\alpha}}.$$

For the optimal choice of α, we want to minimize the estimate in the right hand side of (2.30) over a sequence of *bounded* sets $\mathcal{C}_n \subset H_0^k(\mathbb{T}^d)$, which the truth belongs to. That is, the optimal α is given by

$$\max_{i\in\{1,2,3,\}} \sup_{f\in\mathcal{C}_n} \big\{\Phi_{i,n}(\alpha) + \Psi_{i,n}(\alpha)\big\} \to \min_\alpha .$$

Note that it is reasonable to consider the loss uniformly over some set $\mathcal{C}_n$ instead of a fixed function f in the asymptotic analysis, because the richness of $\mathcal{C}_n$ essentially characterizes the complexity of the problem, see (Tsybakov, 2009, Section 1.2.4) for a detailed argument. Unfortunately, as mentioned early, set $\mathcal{C}_n$ is often unknown, so it is *unrealistic* to find such optimal α in general. We claim, however, that it is possible to find some α that is almost as good as the optimal one, namely, the α that balances $\Phi_{i,n}$ and $\Psi_{i,n}$. A heuristic explanation goes as follows (see also Mathé, 2006): Note that $\Phi_{i,n}$'s are non-decreasing in terms of α while $\Psi_{i,n}$'s are non-increasing, so it amounts to find

$$\max \alpha \qquad \text{subject to } \Phi_{i,n}(\alpha) \le \Psi_{i,n}(\alpha) \quad \text{ for } i = 1, 2, 3.$$

Because of (2.30), it is further "equivalent" to select

$$\max \alpha \qquad \text{subject to } \|D^l\hat{f}_\alpha - D^l f\|_{L^q} \le 2C_0 \max_{i\in\{1,2,3\}} \Psi_{i,n}(\alpha).$$

For every $\tilde{\alpha} \le \alpha$ satisfying the constraint above, it always holds that

$$\|D^l\hat{f}_{\tilde{\alpha}} - D^l f_\alpha\|_{L^q} \le \|D^l\hat{f}_{\tilde{\alpha}} - D^l f\|_{L^q} + \|D^l\hat{f}_\alpha - D^l f\|_{L^q} \le 4C_0 \max_{i\in\{1,2,3\}} \Psi_{i,n}(\tilde{\alpha}).$$

This suggests that we should choose

$$\max \alpha \qquad \text{subject to } \|D^l\hat{f}_{\tilde{\alpha}} - D^l f_\alpha\|_{L^q} \le 4C_0 \max_{i\in\{1,2,3\}} \Psi_{i,n}(\tilde{\alpha}) \quad \text{ for every } \tilde{\alpha} \le \alpha,$$

which is now "computable", and is almost the α that we are searching for (cf. (2.33)).

To make the above argument precise, we first select α_0 (which might depend on n) satisfying

$$\alpha_0 \sup_{f\in\mathcal{C}_n} d_n(1/2\alpha_0)^2 \le \sigma\sqrt{\log n}. \tag{2.31}$$

Such α_0 always exists since

$$\alpha \sup_{f\in\mathcal{C}_n} d_n(1/2\alpha) \le \alpha \sup_{f\in\mathcal{C}_n} d_n(0) = \alpha \sup_{f\in\mathcal{C}_n} \|D^k f\|_{L^2}^2 \to 0 \quad \text{as } \alpha \to 0.$$

If sets $\mathcal{C}_n$'s are further known to be uniformly bounded, the condition (2.31) will asymptotically hold for any fixed α_0 as n goes to infinity.

For an arbitrary $\kappa > 1$, we next consider a discrete set $\mathcal{A}_\kappa$ of candidate parameters by

$$\mathcal{A}_\kappa := \{\alpha_0\kappa^i \colon i = 0,\, 1, \ldots\}. \tag{2.32}$$

We now define an empirical rule for the selection of parameter α, which is known as the *Lepskiĭ (balancing) principle* (Lepskiĭ, 1990), by

$$\alpha_L := \max\Big\{\alpha \in \mathcal{A}_\kappa \colon \|D^l \hat{f}_{\tilde{\alpha}} - D^l \hat{f}_\alpha\|_{L^q} \le 4C_0 \max_{i\in\{1,2,3\}} \Psi_{i,n}(\tilde{\alpha}) \\ \text{for each } \tilde{\alpha} \le \alpha \le n,\, \tilde{\alpha} \in \mathcal{A}_\kappa\Big\}. \tag{2.33}$$

Note that α_L can be computed if the constant C_0 is known. In fact, the explicit value of C_0 depends on the constant in the interpolation inequality in Theorem 2.2.7, which can be calculated by tracing down the proof in Appendix A.1. Alternatively, one can simply replace C_0 by a logarithmic in n factor, for which all the asymptotic analysis still holds at the expense of adding such a log-factor in the rate.

Theorem 2.5.3. *Let $l \in \{0, \ldots, k-1\}$, k, $d \in \mathbb{N}$, $k > d/2$ and $1 \le q \le \infty$. Assume that $\mathcal{B}$ is normal, that the inequality* (2.29) *holds, and that a sequence of bounded sets $\mathcal{C}_n$ in $H_0^k(\mathbb{T}^d)$ satisfy*

$$n \sup_{f\in\mathcal{C}_n} d_n(1/2n)^2 > \sigma\sqrt{\log n} \qquad \text{for } n \ge n_1. \tag{2.34}$$

Assume also that α_L is given by (2.33) *with $\mathcal{A}_\kappa$ in* (2.32) *and α_0 in* (2.31)*. Then the penMIND estimator $\hat{f}_{\alpha_L}$ satisfies that for $n \ge \max\{n_0, n_1\}$,*

$$\sup_{f\in\mathcal{C}_n} \|D^l \hat{f}_{\alpha_L} - D^l f\|_{L^q} \le 6\sqrt{\kappa} C_0 \min_\alpha \max_{i\in\{1,2,3\}} \sup_{f\in\mathcal{C}_n} \big\{\Phi_{i,n}(\alpha) + \Psi_{i,n}(\alpha)\big\},$$

where C_0 and n_0 are the constants in (2.30)*.*

Proof. See Appendix A.4. □

Remark 2.5.4. Note that the condition (2.34) essentially requires that the convergence rate over $\mathcal{C}_n$ is slower than $\sqrt{\log n/n}$, which is often the case for typical choices of $\mathcal{C}_n$, such as Sobolev/Besov balls (cf. Section 2.4.3). In fact, if the reverse of (2.34) holds for n large enough, it follows from (2.30) that the L^q-loss is of order $\sqrt{\log n/n}$ under the assumption (2.29). This leads to the same convergence rate for the L^q-risk, see the coming corollary.

Note also the influence of the design parameter κ onto the error bound. By choosing $\kappa > 1$ small we have little loss compared to the best possible error, but many comparisons have to be carried out in order to find α_L. Thus, in practice, one has to compromise between desired accuracy and computing time.

Corollary 2.5.5. *Let $l \in \{0, \dots, k-1\}$, k, $d \in \mathbb{N}$, $k > d/2$ and $1 \le q \le \infty$. Assume that $\mathcal{B}$ is normal, that a sequence of bounded sets $\mathcal{C}_n$ in $H_0^k(\mathbb{T}^d)$ satisfy (2.34), and that α_L is the same as in Theorem 2.5.3. Then for penMIND $\hat{f}_{\alpha_L}$ it holds almost surely that*

$$\sup_{f \in \mathcal{C}_n} \|D^l \hat{f}_{\alpha_L} - D^l f\|_{L^q} = \mathcal{O}\Big(\min_{\alpha} \max_{i \in \{1,2,3\}} \sup_{f \in \mathcal{C}_n} \big\{\Phi_{i,n}(\alpha) + \Psi_{i,n}(\alpha)\big\}\Big) \qquad \text{as } n \to \infty.$$

Moreover, if for any $\epsilon > 0$

$$\sup_{f \in \mathcal{C}_n} \|f\|_{L^\infty} = o(n^\epsilon) \qquad \text{as } n \to \infty, \tag{2.35}$$

then the assertion above also holds in expectation.

Proof. See Appendix A.4. □

Remark 2.5.6. The technical condition (2.35) says that sets $\mathcal{C}_n$ are *almost* uniformly bounded in $L^\infty(\mathbb{T}^d)$, in the sense that the size of $\mathcal{C}_n$ is allowed to increase at a logarithmic rate as $n \to \infty$. We stress that such requirement is mild, since the L^∞-norm is rather weaker than the norm of $H_0^k(\mathbb{T}^d)$.

2.5.2. Convergence rates for $d = 1$

Based on the results from Section 2.4, we can further derive the concrete convergence rates of penMIND for Sobolev/Besov functions in the one-dimensional case.

Proposition 2.5.7 (Under-smoothing). *Assume that $d = 1$, $l \in \{0, \dots, k-1\}$, $k \in \mathbb{N}$, that $\mathcal{B}$ is regular, and that*

$$f \in B_{p,0}^{s,p'}(\mathbb{T}) \qquad \text{for some } s \in [k+1, 2k] \text{ and } p, p' \in [1, \infty].$$

Assume also that μ is defined in (2.23). Then the penMIND estimator $\hat{f}_\alpha$ in (2.27) with

$$\alpha \sim n^{2\mu} \sqrt{\log n}$$

satisfies almost surely and in expectation that

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

with $\vartheta_l = \vartheta_l(k, 1, 2, q)$ *given in* (2.10).
The same result holds for $f \in W_0^{s,p}(\mathbb{T})$ *with* $k+1 \leq s \leq 2k$ *and* $1 \leq p \leq \infty$.
Proof. It follows readily from Proposition 2.4.1 and Corollary 2.5.2. □

Proposition 2.5.8 (Over-smoothing). *Let* $\mathcal{B}$ *be normal,* $d = 1$, $k \in \mathbb{N}$, *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 penMIND estimator by* (2.27) *with the homogenous Sobolev norm* $\|D^k \cdot\|_{L^2}$ *and the penalization parameter*

$$\alpha \sim n^{-\frac{2(k-s)}{2s+1}} (\log n)^{\frac{2(k-s)}{2s+1}+\frac{1}{2}}.$$

Then it holds almost surely and in expectation that

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

for every $l \in \{0, \ldots, \lfloor s \rfloor - 1\}$ *and every* $q \in \left[1, \frac{4k+2}{2l+1}\right]$.
Proof. See Appendix A.4. □

The above two propositions present the convergence rates for penMIND, which are of the same order as those for MIND in Theorems 2.4.2 and 2.4.4, and even with a slightly sharper logarithmic factor in the over-smoothing case. Note, however, that the results here hold only when we choose α properly according to the smoothness of the truth. Provided that such priori information is available, as a comparison to the discussion after Corollary 2.5.2, the results here make it possible to choose a smaller k for the regularization term of penMIND in (2.27) than the exact smoothness order of the truth. This will enhance the numerical stability for the computation of penMIND. In addition, recall that the penMIND estimator $\hat{f}_{\alpha_L}$ with Lepskiĭ principle performs as well as we know a-priori the best choice of α. Thus the above results also serve as theoretical tools for studying the adaptation property of $\hat{f}_{\alpha_L}$. As a consequence, we have the following result.

Theorem 2.5.9 (Partial adaptation). *Let* $d = 1$, $l \in \{0, \ldots, \min\{\lfloor s \rfloor, k\} - 1\}$, $s \in \mathbb{R}$, $k \in \mathbb{N}$, *and* $\mathcal{B}$ *be regular. Let also* $\mathcal{C}$ *be either the Sobolev ball* $S_L^{s,p}$ *in* (2.24) *or the Besov ball* $B_L^{s,p,p'}$ *in* (2.25), *with*

$$(s, p) \in [1, k) \times \{\infty\},\ \{k\} \times [2, \infty],\ \text{or } [k+1, 2k] \times [2, \infty],$$

and $1 \le p' \le \infty$, $0 < L < \infty$. *Then the penMIND estimator* $\hat{f}_{\alpha_L}$ *with Lepskiĭ principle* (2.33) *satisfies almost surely and in expectation that*

$$\sup_{f \in \mathcal{C}} \|D^l \hat{f}_{\alpha_L} - D^l f\|_{L^q} = \mathcal{O}\left(\left(\frac{\log n}{n}\right)^{\frac{s-l}{2s+1}}\right) \qquad \text{as } n \to \infty,$$

for every $1 \le q \le (4k+2)/(2l+1)$.

Proof. In the case of $s = k$ or $k+1 \le s \le 2k$, the assertion follows directly from Corollary 2.5.5 with $\mathcal{C}_n = \mathcal{C}$, Proposition 2.5.7, and the discussion below Corollary 2.5.2.
We next consider the case of $1 \le s < k$. Let us introduce $g_\lambda \equiv g_\lambda(f)$ for every $f \in \mathcal{C}$ and $\lambda \equiv \lambda_n = \lfloor (n/\log n)^{1/(2s+1)} \rfloor$ as in the proof of Proposition 2.5.8. We define

$$\mathcal{C}_n := \big\{ g_{\lambda_n} \equiv g_{\lambda_n}(f) \colon f \in \mathcal{C} \big\} \qquad \text{for } n \in \mathbb{N}.$$

It follows from Proposition A.3.4 that

$$\sup_{g_{\lambda_n} \in \mathcal{C}_n} \|g_{\lambda_n}\|_{L^\infty} \le C \frac{1}{\lambda_n^s} \sup_{f \in \mathcal{C}} \|f\|_{W^{s,\infty}} = \mathcal{O}\big((\log n/n)^{s/(2s+1)}\big),$$

which implies that $\mathcal{C}_n$ satisfies the condition (2.35). Thus, the assertion for $1 \le s < k$ comes from Corollary 2.5.5 and Proposition 2.5.7. □

We emphasize that the rate in the above theorem is minimax optimal over such choice of $\mathcal{C}$. It shows that the penMIND estimator possesses the same adaptation property as MIND, see Section 2.4.3, and in particular Figure 2.2.

2.6. Computation

We mainly discuss here some efficient algorithms for the computation of the MIND estimator. In what follows, we describe the details of the algorithms, and the corresponding computation complexity. Moreover, we briefly talk about the computation of the penMIND estimator. The implementation is provided in our MATLAB package "Multiscale OPtimization (MOP) ", which is available at `http://www.stochastik.math.uni-goettingen.de/mop`.

2.6.1. Discretization and algorithms

The MIND estimator defined by (2.4) is the solution to a high-dimensional non-smooth convex optimization problem, due to the multiresolution norm. It is clear that the solution

$\hat{f}_{\gamma_n}$ always *exists* and is *unique* since the norm $\|D^k \cdot\|_{L^2}$ is strictly convex and coercive on $H_0^k(\mathbb{T}^d)$. From the convex optimization theory, we know that $\hat{f}_{\gamma_n}$ is characterized by

$$-\hat{f}_{\gamma_n} \in \partial(\chi_{\mathcal{B}} \circ S_n),$$

where the *subgradient* on the right hand side is defined with respect to $\langle \cdot, \cdot \rangle_{H_0^k}$, and

$$\chi_{\mathcal{B}}(z) := \mathbf{1}_{\{\|z - y_n\|_{\mathcal{B}} \le \gamma_n\}} = \begin{cases} 0 & \text{if } \|z - y_n\|_{\mathcal{B}} \le \gamma_n \\ \infty & \text{otherwise} \end{cases} \qquad \text{for every } z \in \mathbb{R}^{\Gamma_n}. \tag{2.36}$$

Note that $\partial(\chi_{\mathcal{B}} \circ S_n) = S_n^* \partial \chi_{\mathcal{B}}$ by the chain rule for subdifferentials (see e.g. Ekeland and Témam, 1999, Proposition I.5.7). It readily follows that

$$\hat{f}_{\gamma_n} \in \operatorname{Ran}(S_n^*) = \operatorname{span}\{\varphi_x \colon x \in \Gamma_n\} \qquad \text{with } \varphi_x \text{ in (2.13)}.$$

In particular, for $d = 1$ the solution $\hat{f}_{\gamma_n}$ is a $2k$-order spline, see (2.22). Thus, it is sufficient to compute only the discretized values of $\hat{f}_{\gamma_n}$ on Γ_n, i.e. $S_n \hat{f}_{\gamma_n}$. The original solution $\hat{f}_{\gamma_n}$ can then be recovered by the interpolation in terms of $\{\varphi_x \colon x \in \Gamma_n\}$.

There are various ways of discretizing the homogeneous Sobolev norm $\|D^k \cdot\|_{L^2}$. For instance, it can be discretized by means of finite differences, with computation complexity $\mathcal{O}(kn)$. Since differentiation turns out to be a simple multiplication after Fourier series expansion, we can also discretize $\|D^k \cdot\|_{L^2}$ by means of discrete Fourier transforms. This can be efficiently computed by the *fast Fourier transform*, with complexity $\mathcal{O}(n \log n)$. Furthermore, we do note that in one dimension it is possible to compute $\|D^k \cdot\|_{L^2}$ a little more precisely, based on the fact that the derivative of a spline is again a spline (but of lower order), and the corresponding coefficients with respect to the B-spline bases are related via finite difference (cf. Schumaker, 2007, Theorem 5.9). In practice, we found that these different discretization schemes lead to almost the same solution of the optimization problem (2.4). Thus, for the sake of simplicity, we always choose the discretization by finite differences.

After discretization, it is easy to see that the optimization (2.4) turns out to be a quadratic program (i.e. quadratic objective under linear constraints). For small sample size (such as in one dimension), this can be efficiently solved, for instance, by *interior point methods* (see e.g. Nesterov and Nemirovskii, 1994). The closeness of the l-th iteration to the optimal solution is often measured by the duality gap μ_l, and it is known that for any $\varepsilon > 0$ it needs

$$l = \mathcal{O}\Big((\#\mathcal{B})^\tau \log \frac{\mu_0}{\varepsilon}\Big)$$

iterations to ensure the duality gap $\mu_l \le \varepsilon$, where $\tau = 1/2$, 1 or 2 depending on the algorithm, see (Potra and Wright, 2000; Bonnans et al., 2006). Implementations of interior point methods are widely available, such as the MATLAB built-in function quadprog.

For large sample sizes, the interior point methods, however, become infeasible (see Marnitz, 2010, for example). In such situation, as mentioned in Section 1.1.2, there are several efficient algorithms nowadays that are able to tackle the high dimensional optimization problems like (2.4). In particular, we have chosen the *alternating direction method of multipliers* (ADMM) algorithm (see Fortin and Glowinski, 1983; Boyd et al., 2011, for example), which is indeed the Douglas-Rachford splitting algorithm (Lions and Mercier, 1979) applied to the dual problem. We follow the approaches outlined in (Frick et al., 2012, 2013), and illustrate the algorithm for a general problem (involving a convex functional $\mathcal{R}$ and a linear operator T) of the form

$$\min_{f\in\mathbb{R}^m} \mathcal{R}(f) + \chi_{\mathcal{B}}(Tf), \tag{2.37}$$

where $\chi_{\mathcal{B}}$ is given by (2.36), and matrix $T \in \mathbb{R}^{n\times m}$. Note that the mean zero requirement of f can be incorporated into the functional $\mathcal{R}$. By introducing a slack variable $g \in \mathbb{R}^n$, we can rewrite the above problem into the equivalent problem

$$\min_{f,g} \mathcal{R}(f) + \chi_{\mathcal{B}}(g) \qquad \text{subject to } Tf - g = 0.$$

By the convex duality theory, it is equivalent to find the saddle point of the *augmented Lagrangian* $L_\lambda(f, g; h)$, that is,

$$\min_{f,g}\max_{h} L_\lambda(f, g; h) := \mathcal{R}(f) + \chi_{\mathcal{B}}(g) + \langle h, Tf - g\rangle_2 + \frac{\lambda}{2}\|Tf - g\|_2^2,$$

where $h \in \mathbb{R}^n$ is the Lagrangian multiplier, and $\lambda > 0$. As its name suggests, the ADMM algorithm solves such saddle point problem alternately over f, g and h. The details are given below.

Algorithm 1: Alternating direction method of multipliers (ADMM)

Input: data $y_n \in \mathbb{R}^n$, step size $\lambda > 0$, tolerance $\varepsilon > 0$, initial values $f_0 \in \mathbb{R}^m, g_0, h_0 \in \mathbb{R}^n$
Iterate for $l = 1, 2, ...$

$$f_l := \arg\min_f \frac{\lambda}{2}\|Tf - (g_{l-1} - \lambda^{-1}h_{l-1})\|_2^2 + \mathcal{R}(f) \tag{2.38}$$

$$g_l := \arg\min_g \frac{\lambda}{2}\|g - (Tf_l + \lambda^{-1}h_{l-1})\|_2^2 + \chi_{\mathcal{B}}(g) \tag{2.39}$$

$$h_l := h_{l-1} + \lambda(Tf_l - g_l) \tag{2.40}$$

until $\max\{\|Tf_l - g\|_2, \|T(f_l - f_{l-1})\|_2\} \le \varepsilon$

Note first that the update of dual variable h in (2.40) is simply a gradient assent step of maximizing $L_\lambda(f, g; h)$ over h. We next discuss how to solve the subproblems (2.38) and (2.39): The subproblem (2.38) is a typical regularization problem, the algorithm for

which depends on the choice of $\mathcal{R}$. For instance, it can be solved by the inversion of linear equations if $\mathcal{R}(f) = \|Af\|_2$ with some matrix A, which covers the case of MIND. Furthermore, one can even avoid the operator T by considering instead the following problem

$$f_l := \arg\min_f \frac{\lambda}{2}\|Tf - (g_{l-1} - \lambda^{-1}h_{l-1})\|_2 + \mathcal{R}(f) + \frac{\lambda}{2}(\tau\|f - f_{l-1}\|_2 - \|T(f - f_{l-1})\|_2)$$

with $\tau \geq \|T\|_2$. Such modification leads to the *inexact ADMM* (Frick et al., 2013), also known as *Chambolle-Pock* algorithm (Chambolle and Pock, 2011). The subproblem (2.39) is to find the projection onto the intersection of a finite number of convex sets (more precisely, closed half spaces). It can be computed by the *Dykstra's algorithm* (Dykstra, 1983; Boyle and Dykstra, 1986), which converges linearly (Deutsch and Hundal, 1994). See (Birgin and Raydan, 2005) for an efficient stopping rule.

For the optimization problem (2.4) of MIND, we have a linear convergence guarantee from (Deng and Yin, 2015, Corollary 3.1) for the ADMM algorithm

$$\lambda\|f_l - f_*\|_2^2 + \frac{1}{\lambda}\|h_l - h_*\|_2^2 \leq \left(\frac{1}{1+\delta}\right)^l \left(\lambda\|f_0 - f_*\|_2^2 + \frac{1}{\lambda}\|h_0 - h_*\|_2^2\right)$$

with

$$\delta = 2\left(\frac{\lambda}{n^k \sin^{2k}(2\pi/n^{1/d})} + \frac{2^{2k}n^k}{\lambda}\right)^{-1}.$$

In particular, the choice of $\lambda = \left(2n\sin(2\pi/n^{1/d})\right)^k$ yields the largest

$$\delta = 2^{-k}\sin^{2k}(2\pi/n^{1/d}) \sim \left(\pi/n^{1/d}\right)^k.$$

This also gives suggestion on the choice of parameter λ, although such bound turns out to be rather pessimistic in practice.

For small sample sizes, we found by simulation that the ADMM algorithm works as well as the interior point methods. Thus, for consistency, we always assume the ADMM algorithm when referring to the computation of MIND.

We now turn to the optimization problem (2.27) of penMIND, which is equivalent to

$$\min_{f,\delta} \frac{\alpha}{2}\|D^k f\|_{L^2}^2 + \delta \qquad \text{subject to } \|S_n f - y_n\|_{\mathcal{B}} \leq \delta,$$

This is similar to the problem (2.4) of MIND, so all the discussions above also apply to the computation of penMIND.

2.6.2. Software

As a companion to this work, the package "MOP" implements the algorithms discussed above for the numerical computing software MATLAB, The MathWorks, Inc. The code is designed for both one-dimensional signals and two-dimensional images. It includes procedures for solving the general optimization problem (2.37) in the case that $\mathcal{R}$ is either a total variation (TV) semi-norm, or a homogeneous H^k-norm (i.e. the MIND estimator), by means of Algorithm 1. Moreover, an implementation based on interior point methods is also provided for such problems in the case of $d = 1$. For instance, the procedure

$$\underbrace{\text{mind}}_{\text{estimator}}\overbrace{\text{Regression2d}}^{\text{problem}}\underbrace{\text{ADMM}}_{\text{algorithm}}(...)$$

computes the MIND estimator in (2.4) for two-dimensional nonparametric regression by the ADMM algorithm. Following this naming convention, one can easily find the correct piece of code for a specific purpose. In addition, the package also contains the implementation of the Nemirovski's estimator in (2.44), the penMIND estimator (2.27), and the other estimators considered in Sections 2.7 and 3.4.

2.7. Numerical experiments

In this section, we consider the finite sample behavior, that is, the practical performance of the MIND estimator for the recovery of signals living on one-dimensional domain, by means of numerical simulations.

2.7.1. Practical considerations

Note that the MIND estimator by (2.4) involves in total three parameters: the system of cubes $\mathcal{B}$, the threshold γ_n, and the smoothness parameter k. It follows from the asymptotic analysis in previous sections that both of them can be chosen automatically in a way that relies rather weakly on the underlying true signal. We next discuss some adjustment in order to improve the performance of MIND for fixed sample sizes.

The choice of the system of cubes $\mathcal{B}$. The abstract convergence rate results (cf. Corollary 2.3.4 and Theorem 2.3.4) require that the system $\mathcal{B}$ should be normal (see Definition 2.2.2). In particular, when $d = 1$, the result of concrete rates needs the same requirement for the over-smoothing case (cf. Theorem 2.4.4), while it imposes a slightly stronger condition, namely that the system $\mathcal{B}$ should be regular (see Definition 2.2.3) for the under-smoothing case (see Theorem 2.4.2). Thus, every regular system $\mathcal{B}$ is sufficient

to guarantee that all the theoretical analyses hold. For instance, the m-partition system is regular, which is also the sparsest system of m-regularity, see Example 2 for more examples. In some applications, if the true signal is known *a-priori* to have features of certain scales and locations, we can incorporate this prior information by choosing a regular system $\mathcal{B}$ that includes all cubes of those scales and locations. Concerning different regular systems, the richer ones tend to give better performance, while such difference is diminishing as the sample size n increases, and even visually indistinguishable when n is large. We know from Section 2.6 that the computational complexity of MIND increases as the system $\mathcal{B}$ becomes larger. As a compromise between computation and performance, we recommend to use the system of cubes with dyadic edge lengths for small and medium scale problems, and the 2-partition system for large scale ones.

The choice of the threshold γ_n. The asymptotic theory only requires that γ_n satisfies the condition (2.5), which is independent of the system of cubes $\mathcal{B}$, and the smoothness of the truth. In the finite sample situation, we recommend a refined choice, which has a direct statistical interpretation, see Chapter 1 and also (Donoho, 1995a; Dümbgen and Walther, 2008; Davies et al., 2009; Frick et al., 2014). It selects γ_n as the α-quantile of the multiscale statistic $\|\xi_n\|_{\mathcal{B}}$, i.e.,

$$\gamma_n(\alpha) := \inf\Big\{\gamma : \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} > \gamma\right\} \le \alpha\Big\}. \tag{2.41}$$

This ensures that the truth f lies in the confidence set defined by the multiscale constraint in the right hand side of (2.4) with probability at least $1-\alpha$. Thus we have

$$\mathbb{P}\left\{\|D^k \hat{f}_{\gamma_n}\|_{L^2} \le \|D^k f\|_{L^2}\right\} \ge 1-\alpha, \tag{2.42}$$

that is, the MIND estimator is smoother than the truth with probability at least $1-\alpha$. In this way, the choice of threshold γ_n amounts to select the significance level α. It is clear that the MIND estimator $\hat{f}_{\gamma_n(\alpha)}$ is smoother for smaller α. In addition, MIND $\hat{f}_{\gamma_n(\alpha)}$ is actually quite robust against the choice of α, see Section 2.7.2.

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 (2.5) when $d=1$, the asymptotic analysis in this chapter 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 $\gamma_n(\alpha)$ can be done by Monte-Carlo simulations using the distribution of noise ξ_n. If ξ_n is only known to be sub-Gaussian in (2.3), we then draw ξ_n from $\mathcal{N}(0,\sigma^2)$,

i.e. the relation (2.3) holds with equality. This gives an upper bound of $\gamma_n(\alpha)$, and makes the interpretation (2.41) still valid. As mentioned early, the noise level σ can be easily estimated when it is unknown (see Remark 2.1.1). In general, the computational complexity of multiscale statistic $\|\xi_n\|_{\mathcal{B}}$ depends on the effective cardinality of $\mathcal{B}$ (cf. discussion below Definition 2.2.3). In the case of $d = 1$, however, there are fast algorithms with linear complexity in terms of the sample size n, even though the effective cardinality of $\mathcal{B}$ can be $\mathcal{O}(n^2)$, see (Bernholt et al., 2007, 2009). It is worth noting that the computation of $\gamma_n(\alpha)$ is needed only once for a fixed size of measurements n and a fixed system of cubes $\mathcal{B}$.

The choice of smoothness parameter k. From Section 2.4.3 we see that the adaptation region of MIND enlarges as the smoothness order of regularization k increases (see in particular Figure 2.2). This suggests that we should choose k as large as possible. The minimization problem in (2.4) becomes, however, more numerically unstable as k increases, since it involves k-th order derivatives. Thus, the choice of k should balance the adaptivity and the numerical stability. In practice, we find that it works fine for $k = 1, 2, 3$ (see Section 2.7.2).

2.7.2. Simulation results

In the simulations, we always assume the noise level is known. The MIND estimator (2.4) is computed by an ADMM algorithm, see Section 2.6. If there is no explicit statement, the multiresolution norm is defined using 2-partition system, and for MIND the threshold $\gamma_n(\alpha)$ in (2.41) with the significance level $\alpha = 0.1$ is chosen, which is estimated by Monte-Carlo simulations with 10^5 repetitions. All the experiments can be reproduced by means of the MATLAB package "MOP" (see Section 2.6.2 for details).

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

$$\min_f \|S_n f - y_n\|_2 + \lambda \|D^k f\|_{L^2}^2, \tag{2.43}$$

and the Nemirovski's estimator (Nem) $\hat{f}_\eta$ as the solution of

$$\min_f \|S_n f - y_n\|_{\mathcal{B}} \qquad \text{subject to } \|D^k f\|_{L^2} \le \eta, \tag{2.44}$$

which is indeed a particular case of $\hat{f}_{p,\eta}$ in (1.4) 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 finite differences. As MIND, the Nemirovski's estimator is computed by an ADMM algorithm.

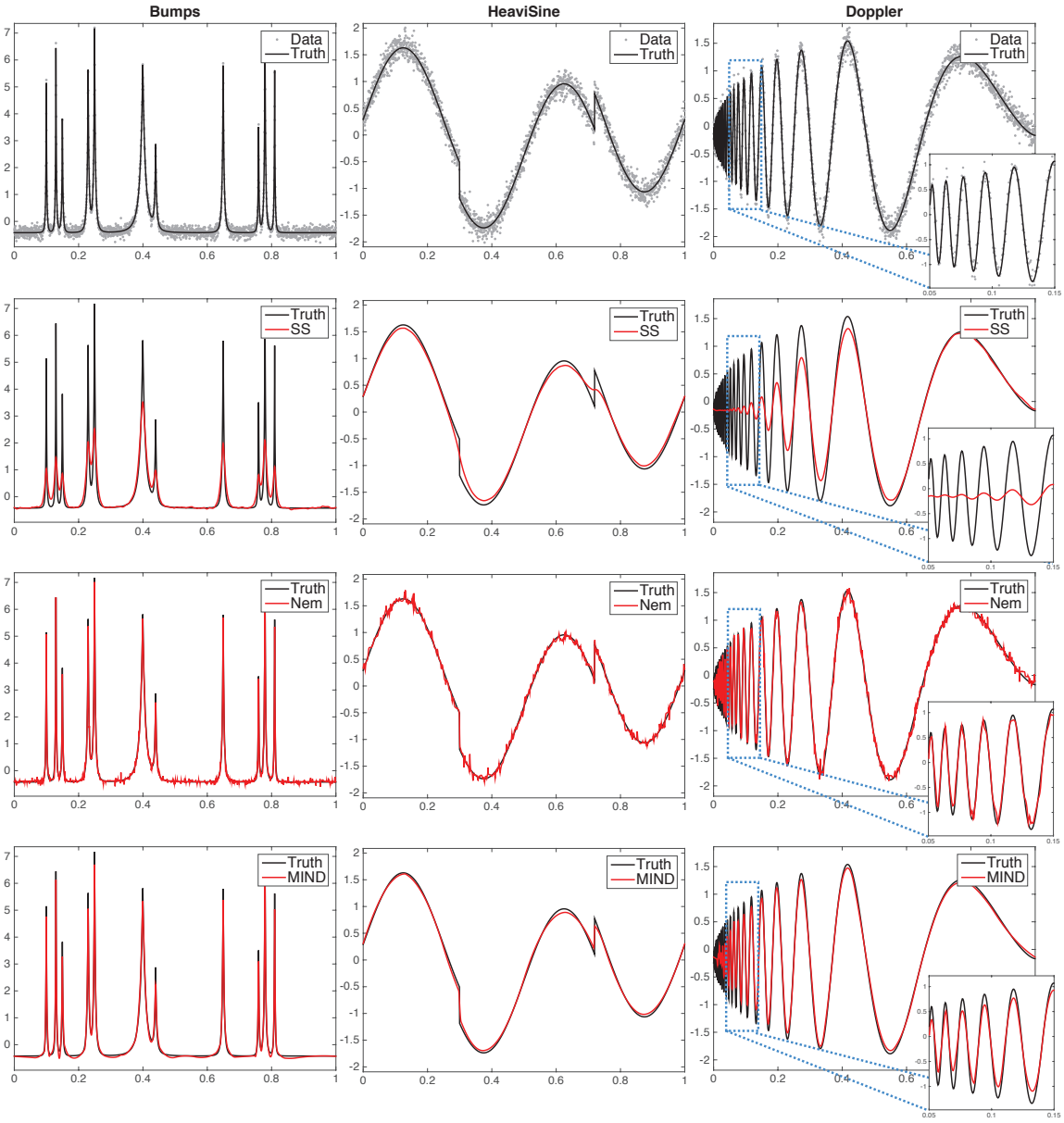

Figure 2.3.: Comparison of SS in (2.43), Nem in (2.44), and MIND in (2.4) (number of samples $n = 2^{11}$, noise level $\sigma = 0.12\|f\|_{L^2}$).

The simulation results are summarized in Figure 2.3. 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 1.1.1) 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.

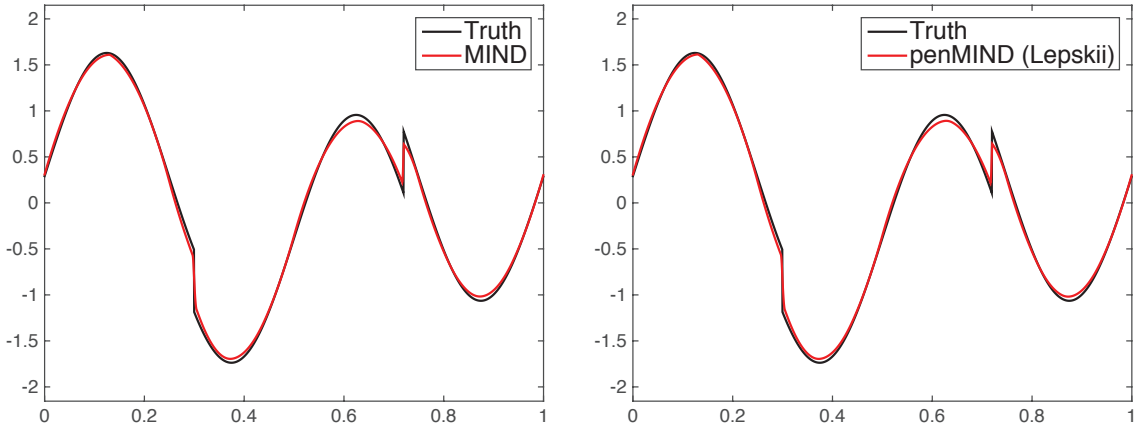

Figure 2.4.: Comparison of MIND in (2.4) and penMIND in (2.27) with Lepskiĭ principle in (2.33) on the HeaviSine signal (number of samples $n = 2^{11}$, noise level $\sigma = 0.12\|f\|_{L^2}$).

Moreover, we make a comparison between MIND and the penMIND estimator with the Lepskiĭ principle, which is introduced in Section 2.5. The homogeneous H^1-norm (i.e. $k = 1$) is chosen as the regularization term for both estimators. Concerning the implementation of the Lepskiĭ principle, we choose $q = 2$, $\alpha_0 = 0.01$, $\kappa = 1.1$, and $C_0 = 0.4$, see (2.33). Such choice of C_0 gives desirable performance uniformly over various signals and a range of sample sizes, in the current setting, i.e. $k = d = 1$. We compute penMIND using an ADMM algorithm, see Section 2.6 for details. Similar to the asymptotic analysis, MIND and penMIND with the Lepskiĭ principle works nearly the same for signals of different nature in practice. For example, we show in Figure 2.4 the results of MIND and penMIND with the Lepskiĭ principle on the estimation of the HeaviSine signal from the same data as the previous simulation study. There is almost no visual difference, and the relative difference between them with respect to L^2-norm is less than 0.18%. We note, however,

that the competent performance of penMIND with the Lepskiĭ principle relies heavily on the empirical choice of $C_0 = 0.4$, which is way smaller than the theoretical value, and needs to be adjusted for different values of k and d. We are not aware of any better strategy for the tuning of C_0 than the exhaustive search. Therefore, we recommend MIND for practical purpose, and only present the results for MIND in the coming experiments.

Robustness and stability in significance level

We first examine the robustness of SS, Nem, 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 2.5 we find that the MIND estimator is rather robust to the choice of

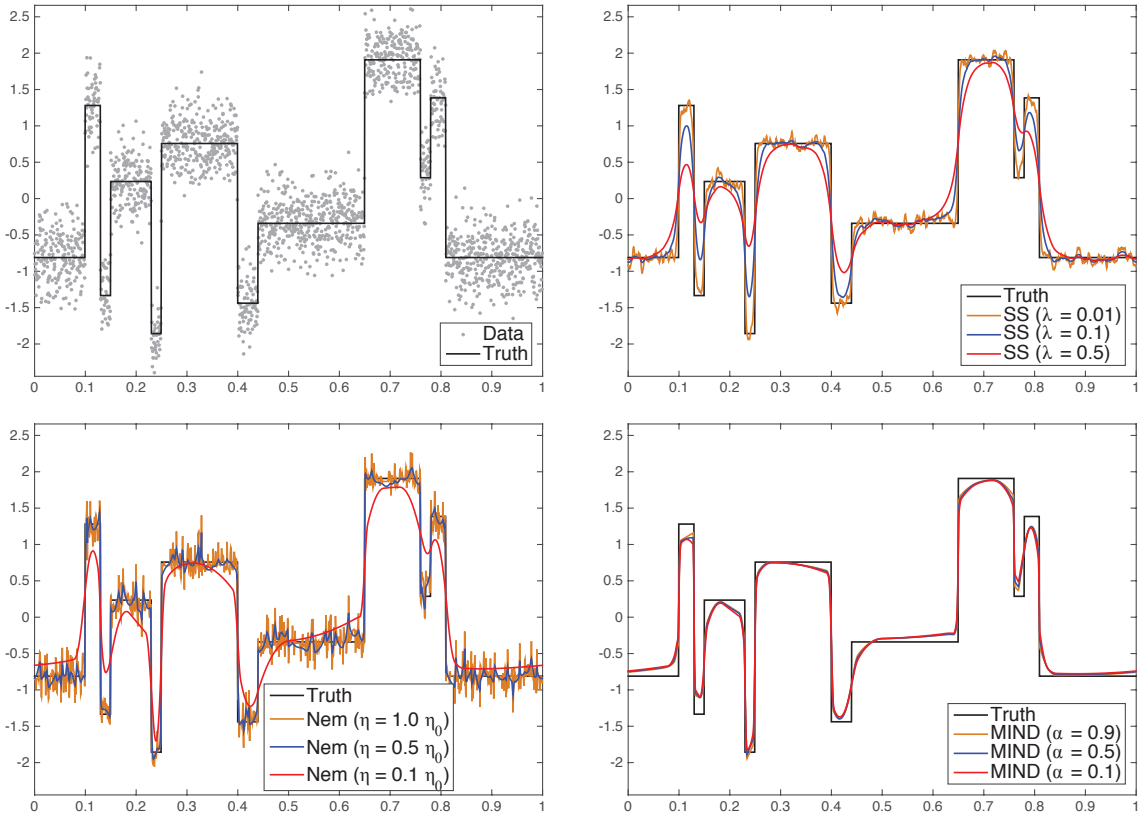

Figure 2.5.: Impact of parameter choices: various λ for SS $\hat{f}_\lambda$ in (2.43), various η for Nem $\hat{f}_\eta$ in (2.44), and various α for the MIND $\hat{f}_{\gamma_n}$ in (2.4) with $\gamma_n = \gamma_n(\alpha)$ in (2.41). (number of samples $n = 2^{11}$, noise level $\sigma = 0.3\|f\|_{L^2}$, and $\eta_0 = \|D^k f\|_{L^2}$).

significance level α, while SS and Nem 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 SS and Nem.

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 2.6, 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 2.6). Recall that the significance level α can be interpreted as an error control in the sense of (2.42). Thus, the additional power by an increased significance level comes at the expense of a lower confidence about the inference. It is natural to cast the choice of large α into the asymptotics that $\alpha_n \not\to 0$ as $n \to \infty$, which leads to inconsistency of MIND $\hat{f}_{\gamma_n}$ with $\gamma_n = \gamma_n(\alpha_n)$. Consequently, this makes impossible to derive convergence rates for such choice of α_n, but we conjecture that in such situation it might be possible to control instead the false discovery rate (see Li et al., 2014, for an answer in multiple change-point segmentation).

We next study the influence of the noise distribution on the choice of threshold $\gamma_n(\alpha)$, and in turn on the behavior of MIND. We still assume the noise level $\sigma := \mathrm{sd}(\xi_n)$ is known, but the common distribution of ξ_n is only known to have sub-Gaussian tails as in (2.3). This situation is often encountered in real applications, since compared to the exact distribution, the noise level is way easier to be estimated. Noting that the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ attains the worst bound in (2.3), one should simulate $\gamma_n(\alpha)$ by assuming that the noise distribution is Gaussian in order to guarantee the validity of the theoretical analysis, see also Section 2.7.1. To investigate how much we lose for such a conservative choice of $\gamma_n(\alpha)$, we consider the example that the data are collected from the Doppler signal with uniformly distributed noise, see Figure 2.7. The upper panel shows the result of MIND with $\gamma_n(\alpha)$ (precisely, $\gamma_n(0.1) = 4.02$) that is estimated from Gaussian distribution, as well as the noisy data and the true signal. As a comparison, we also illustrate in the lower panel the behavior of MIND when the threshold $\gamma_n(\alpha)$ (precisely, $\gamma_n(0.1) = 3.35$) is simulated from the exact distribution of ξ_n, namely the uniform distribution. These two choices of thresholds lead to in general comparable recovered signals, but with slight differences. In particular, one can see from the magnified region that MIND with $\gamma_n(\alpha)$ given by the exact noise distribution detects one additional peak from the underlying truth, which is marked by a shaded blue bar.

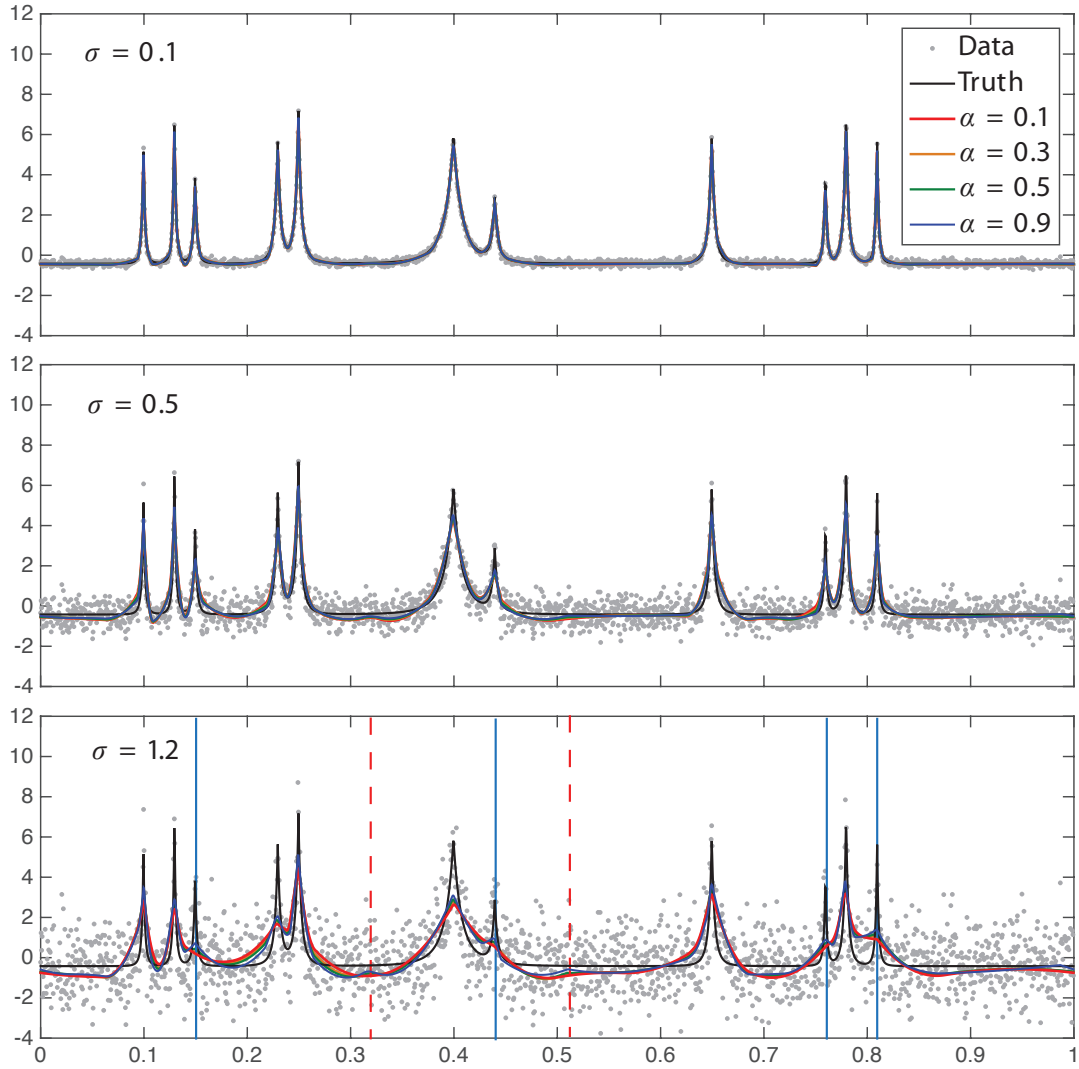

Figure 2.6.: Stability of MIND in significance level α and noise level σ. The reconstructions by MIND $\hat{f}_{\gamma_n}$ with $\gamma_n = \gamma_n(\alpha)$ for a range of α's are shown, together with the true signal and noisy data, in the cases of different noise levels (number of samples $n = 2^{11}$).

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).

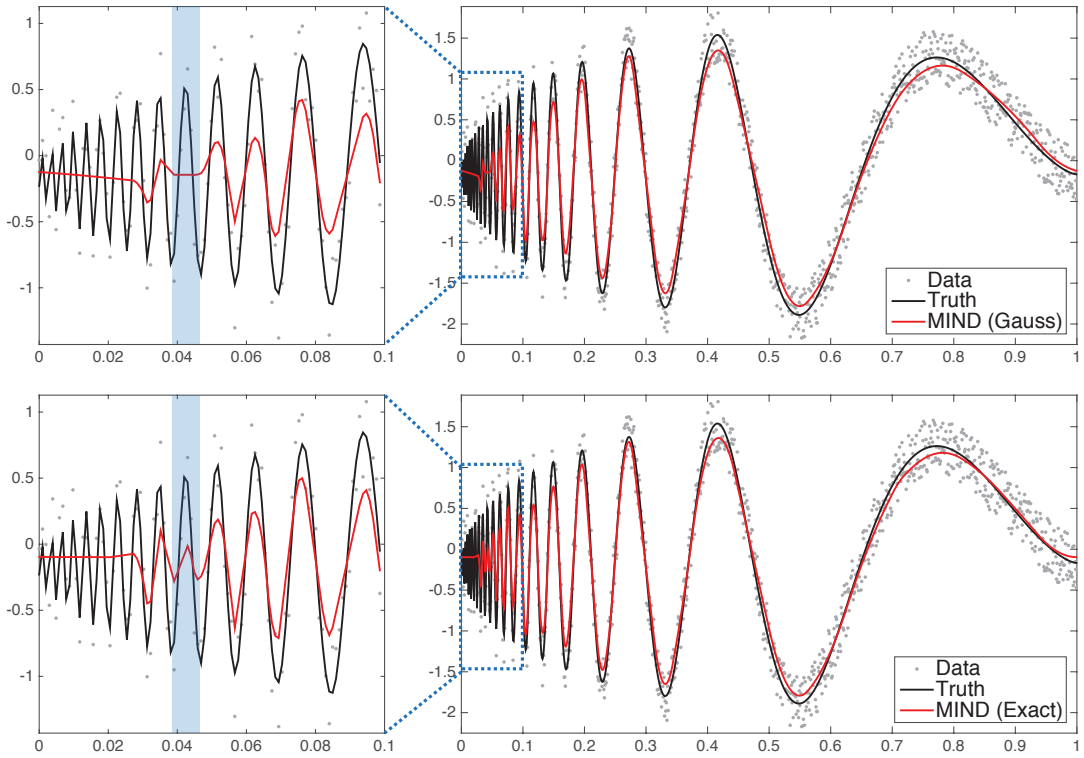

Figure 2.7.: Influence of noise distribution on MIND for the Doppler signal (number of samples $n = 2^{10}$, noise level $\sigma = 0.2\|f\|_{L^2}$).

Estimation of derivatives

Besides the estimation of the function itself, MIND also serves as an estimator for derivatives. We further evaluate the performance of MIND in this setting. For that purpose, we consider

$$f(x) := \operatorname{sgn}(x - 0.5)\sin^4(2\pi x) \in H_0^{4.5-\varepsilon}(\mathbb{T}) \qquad \text{for any } \varepsilon > 0$$

as the test signal, see (Griebel and Hamaekers, 2014). The smoothness order of the regularization term is set as $k = 3$. From the theoretical result, it follows that MIND is nearly minimax optimal for the estimation of the test signal above, and of its derivatives up to second order, see Section 2.4. The corresponding empirical performance is given in Figure 2.8, where the derivatives of the truth are calculated analytically, while those of MIND are estimated by finite differences. As one can see, MIND performs fairly well in the recovery of both the true signal and derivatives. In addition, we point out that the performance of MIND gets worse for the estimation of higher order derivative for a fixed sample size. This can be seen from the fact that MIND detects all major features

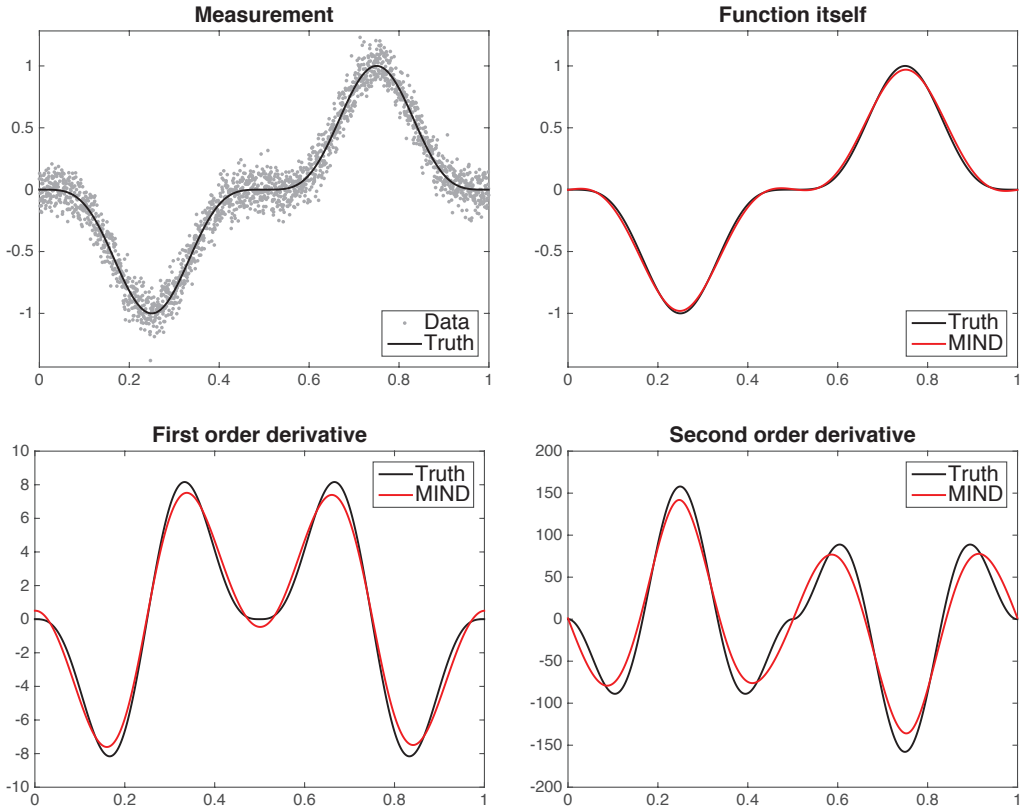

Figure 2.8.: Estimation of derivatives by MIND (number of samples $n = 2^{11}$, noise level $\sigma = 0.1\|f\|_{L^2}$).

of the signal and the first derivative, but misses the inflection point at 0.5 for the second derivative.

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 2.9) is chosen as the test signal, and the noise is independent Gaussian distributed with standard deviation $\sigma =$ The significance level for MIND is set to $\alpha = 0.1$. Figure 2.9 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 2.4.3. As already mentioned, one should, however, notice that

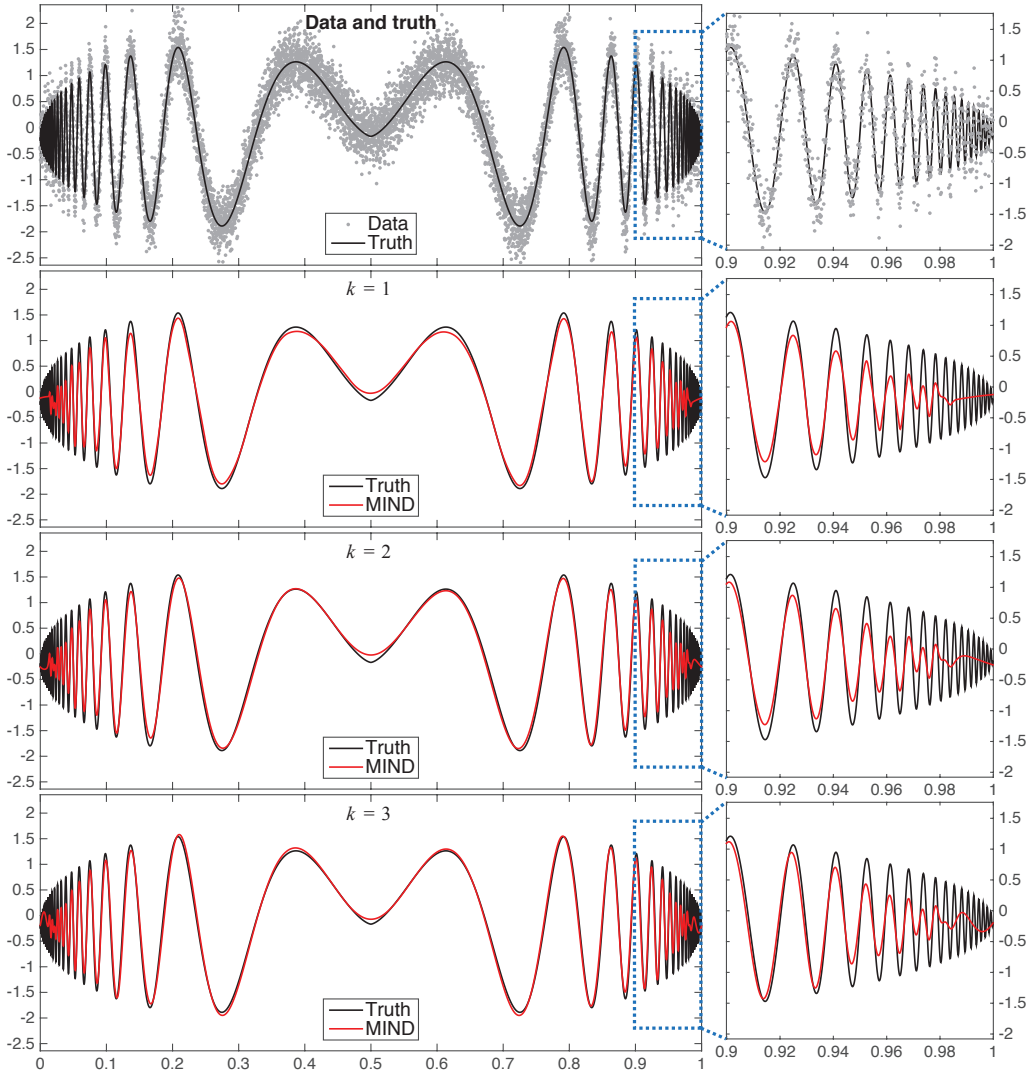

Figure 2.9.: Various choices of smoothness order k in the regularization term for MIND in (2.4) on symmetric 'Doppler' signal (number of samples $n = 2^{13}$, noise level $\sigma = 0.3\|f\|_{L^2}$).

the optimization problem becomes numerically more ill-conditioned as k increases.

3. Statistical Inverse Problems

In this chapter we consider the application of the MIND estimator to linear statistical inverse problems. We are particularly interested in the situations where the forward operator is β-smoothing over Hilbert scales $H_0^s(\mathbb{T}^d)$ with $s \geq 0$. This includes nonparametric regression, deconvolution, and reconstruction from Radon transform measurements as special cases. In such a setting, we study the asymptotic property of MIND as an extension of the analytical methodology introduced in the previous chapter. A first and crucial result is an interpolation inequality between Sobolev norms and the multiresolution norm applied to the range of the forward operator. This, together with a natural generalization of the multiscale distance functions, forms essentially the basis of all the coming results. The material is organized similarly as the previous chapter: we first give a general analysis based on abstract smoothness assumptions, then focus on examples in one dimension, and finally present some simulation results.

3.1. MIND as regularization methods

We study the inverse problem of solving an operator equation

$$Tf = y$$

with noisy right hand side y, where

$$T\colon L^2([0,1]^d) \to L^2([0,1]^d)$$

is a *bounded linear* operator. More precisely, we sample the data y on the regular grid Γ_n on $[0,1]^d$ (cf. (2.2)) and assume that we are given noisy data

$$y_n(x) = Tf(x) + \xi_n(x) \qquad \text{for } x \in \Gamma_n, \tag{3.1}$$

where the vector $(\xi_n(x))_{x\in\Gamma_n}$ is the realization of an i.i.d. centered sub-Gaussian noise process with scale parameter σ as in (2.3). For technical simplicity, we assume that the solution f can be periodically extended to $\mathbb{R}^d$ and has mean zero, and the noise level σ is known beforehand; that is, Assumption 1 holds.

In the following we study operators T that are defined on the Hilbert scale $H_0^s(\mathbb{T}^d)$, that is, for each $s \geq 0$ we can write T as an operator

$$T\colon H_0^s(\mathbb{T}^d) \to H_0^s(\mathbb{T}^d).$$

Moreover, we assume that the forward operator T satisfies the following property.

Definition 3.1.1. A linear operator T is called *β-smoothing* for some $\beta \geq 0$ if there are constants $C_1, C_2 > 0$, depending only on s and d, such that

$$C_1 \|f\|_{H_0^s} \leq \|Tf\|_{H_0^{s+\beta}} \leq C_2 \|f\|_{H_0^s}, \tag{3.2}$$

for every $f \in H_0^s(\mathbb{T}^d)$ and every $s \geq 0$.

Remark 3.1.2. Note that if $\beta > 0$, then every β-smoothing operator $T\colon H_0^s(\mathbb{T}^d) \to H_0^s(\mathbb{T}^d)$ is compact, due to the compact embedding theorem (Adams and Fournier, 2003, Theorem 6.1) of Sobolev spaces, and thus does not have a bounded inversion. For the clarity and simplicity of exposition, we restrict ourselves to Hilbert scales $H_0^s(\mathbb{T}^d)$, $s \geq 0$, and will not pursue the most general case. The aim is mainly to give an illustration on how to extend the analysis framework established in Chapter 2 to problems other than nonparametric regression.

Some examples of model (3.1) with β-smoothing operators T are collected below.

Example 4. (a) The nonparametric regression corresponds to a special case of (3.1) with $T = \mathrm{I}$, the identity operator. Obiviously, the identity operator I satisfies the estimate (3.2) with $\beta = 0$, i.e. is 0-smoothing.

(b) For deconvolution problem, the operator T is defined as $Tf = \rho * f$ with convolution kernel ρ. Given a fixed $\beta \geq 0$, if the Fourier coefficients $u_\lambda := \langle \rho, e^{-2\pi i \langle \lambda, \cdot \rangle} \rangle_{L^2}$ of the kernel ρ satisfy the growth condition

$$u_\lambda \sim |\lambda|^{-\beta} \qquad \text{for all } \lambda \in \mathbb{Z}^d \setminus \{0\},$$

then it can be easily seen that the convolution operator T is β-smoothing.

(c) The computerized tomography problem also lies in the model (3.1) with T being the Radon transform. Some rescaling is necessary to ensure that the Radon transform is an operator from $L^2(\mathbb{T}^d)$ to $L^2(\mathbb{T}^d)$. In this case, the d-dimensional Radon transform is actually $(d-1)/2$-smoothing, see e.g. (Natterer, 2001, Chapter II, Theorem 5.1).

To guarantee that the point evaluation in model (3.1) is well defined, we further assume that the true solution f of this inverse problem is such that $Tf \in H_0^s(\mathbb{T}^d)$ is continuous. Because of the Hilbert scale assumption, this will be the case if $s > d/2$. Actually, for β-smoothing operators, the condition $f \in H_0^s(\mathbb{T}^d)$ with $s > d/2 + \beta$ would be sufficient for the continuity of Tf. In such a setting, we can write the evaluation of Tf on the grid Γ_n as

a composition of T with the sampling operator S_n. Doing so, we can rewrite the problem we want to solve as the inverse problem of solving the equation

$$T_n f := (S_n \circ T) f = y_n$$

with noisy right hand side $y_n \in \mathbb{R}^{\Gamma_n}$ as given in (3.1). In order to solve the inverse problem, we propose the MIND estimator as a regularization method. To be precise, we consider the estimator $\hat{f}_{\gamma_n}$ given by

$$\hat{f}_{\gamma_n} = \operatorname*{arg\,min}_{f \in H_0^k(\mathbb{T}^d)} \frac{1}{2} \|D^k f\|_{L^2}^2 \qquad \text{subject to } \|T_n f - y_n\|_{\mathcal{B}} \leq \gamma_n, \tag{3.3}$$

which is a reformulation of (1.9) with abbreviation $T_n = S_n \circ T$. Here, the threshold γ_n is chosen in a universal way as

$$\gamma_n = C(\log n)^r \quad \text{for some } r \geq \frac{1}{2} \text{ and } C > \begin{cases} 0 & \text{if } r > \frac{1}{2} \\ \sigma\sqrt{6 + \frac{2k+2\beta}{d}} & \text{if } r = \frac{1}{2} \end{cases}, \tag{3.4}$$

and the system $\mathcal{B}$ of cubes satisfies Assumption 2, that is, the corresponding multiresolution norm being indeed a norm. Provided that

$$\operatorname{Ran}(T_n) = \mathbb{R}^{\Gamma_n} \qquad \text{for every } n \in \mathbb{N},$$

which we always assume, it follows directly that the multiscale constraint in (3.3) is non-empty for all $\gamma_n \geq 0$ and $y_n \in \mathbb{R}^{\Gamma_n}$. From the strict convexity and coercivity of $\|D^k \cdot\|_{L^2}$ on $H_0^k(\mathbb{T}^d)$, we further see that MIND $\hat{f}_{\gamma_n}$ as a solution to (3.3) always *exists* and is *unique*.

3.2. Convergence analysis

3.2.1. An interpolation inequality

We start by an interpolation inequality between H_0^s-norms ($s \geq 0$), and the multiresolution norm (cf. Section 2.2).

Lemma 3.2.1. *Assume that $0 \leq r < \lfloor s \rfloor$, $d < 2\lfloor s \rfloor$, $r, s \in \mathbb{R}$, $d \in \mathbb{N}$, and that $\mathcal{B}$ is a normal system of cubes. Then there exist constants C and n_0, depending only on s, d, and $\mathcal{B}$, such that for every $f \in H^s([0,1]^d)$ and for $n \geq n_0$,*

$$\|f\|_{H_0^r} \leq C \max\Big\{ n^{-\frac{s-r}{2s+d}} \|S_n f\|_{\mathcal{B}}^{\frac{2s-2r}{2s+d}} \|f\|_{H_0^s}^{\frac{2r+d}{2s+d}},\, n^{-1/2} \|S_n f\|_{\mathcal{B}},\, n^{-\frac{2\lfloor s \rfloor (s-r)}{d(2\lfloor s \rfloor + d)}} \|f\|_{H_0^s} \Big\}.$$

Proof. See Appendix B.1. □

The above inequality is an extension of Theorem 2.2 to homogeneous Sobolev norms with non-integer orders of smoothness, in the case of $p = q = 2$. Recall that $s > d/2$ is sufficient for the continuity of functions in $H^s([0,1]^d)$, which in turn guarantees the well-definedness of S_n. Thus the technical requirement $d < 2\lfloor s \rfloor$ is slightly stronger than necessary. It is imposed because the approach of the proof relies on a certain approximation by polynomials. We conjecture that the lemma holds also for $d < 2s$ if we instead use some well behaved functions for the approximation part in the proof.

Proposition 3.2.2. *Assume that the operator T is β-smoothing for some $\beta \geq 0$, that $k + \lfloor \beta \rfloor > d/2$, $k, d \in \mathbb{N}$, and that the system $\mathcal{B}$ is normal. Then there are constants $C > 0$ and $n_0 \in \mathbb{N}$ only depending on k, d, $\mathcal{B}$, and T, such that for every $f \in H_0^k([0,1]^d)$*

$$\|f\|_{L^2} \leq C \max\Big\{ \frac{\|T_n f\|_{\mathcal{B}}^{2\vartheta}}{n^{\vartheta}} \|f\|_{H_0^k}^{1-2\vartheta},\ \frac{\|T_n f\|_{\mathcal{B}}}{n^{1/2}},\ \frac{\|f\|_{H_0^k}}{n^{\vartheta'}} \Big\},$$

where

$$\vartheta := \frac{k}{2k + 2\beta + d}, \tag{3.5a}$$

$$\text{and} \quad \vartheta' := \frac{2k(k + \lfloor \beta \rfloor)}{d(2k + 2\lfloor \beta \rfloor + d)}. \tag{3.5b}$$

Proof. It follows by applying Lemma 3.2.1 to Tf with $r = \beta$ and $s = k + \beta$, and the estimate (3.2). □

It always holds that $\vartheta < \vartheta'$, since $k + \lfloor \beta \rfloor > d/2$. Note that it is actually possible to consider the H_0^θ-norm ($0 \leq \theta < k + \lfloor \beta \rfloor - \beta$) on the left hand side of the inequality from Proposition 3.2.2, if we choose $r = \beta + \theta$ in the proof. This will in turn lead to estimates for the H_0^θ-loss of MIND, and the analysis of MIND for the estimation of derivatives. However, for the sake of simplicity, we only consider the L^2-loss and the estimation of function itself here and later, which can be easily extended to the more general case.

3.2.2. Approximate source conditions

Similar as in Chapter 2, our analysis of convergence is based on approximate source conditions and the multiscale distance function corresponding to the operator T at the true solution f. In order to define the latter, we first recall that $T_n : H_0^k(\mathbb{T}^d) \to \mathbb{R}^{\Gamma_n}$ is bounded linear, and thus has a well-defined adjoint $T_n^* : \mathbb{R}^{\Gamma_n} \to H_0^k(\mathbb{T}^d)$ given by

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

Moreover, we recall the dual $\|\cdot\|_{\mathcal{B}^*}$ of the multiresolution norm from Section 2.3.1.

Definition 3.2.3. We define the *multiscale distance function* $d_n(t;T)$ with respect to f and a linear operator T by

$$d_n(t;T) := \min_{\|\omega\|_{\mathcal{B}^*}\leq t} \|D^k T_n^*\omega - D^k f\|_{L^2} = \min_{\|\omega\|_{\mathcal{B}^*}\leq t} \|T_n^*\omega - f\|_{H_0^k} \qquad \text{for } t \geq 0.$$

Remark 3.2.4. The definition of the multiscale distance function immediately implies that it is a decreasing function in t for each fixed n and T. Moreover

$$d_n(0;T) = \|f\|_{H_0^k},$$

which provides an upper bound of the multiscale distance function provided that the truth f is contained in $H_0^k(\mathbb{T}^d)$. Note that Definition 2.3.1 is a particular case of Definition 3.2.3, namely, $d_n(t) = d_n(t;\mathrm{I})$.

By the introduction of multiscale distance functions, we are able to derive an upper bound on the L^2-loss of MIND in case that the truth f is admissible to the multiscale constraint in (3.3).

Lemma 3.2.5. *Assume that the operator T is β-smoothing for some $\beta \geq 0$, that $k, d \in \mathbb{N}$, $k + \lfloor\beta\rfloor > d/2$, that the system $\mathcal{B}$ is normal, and that the inequality*

$$\|\xi_n\|_{\mathcal{B}} = \|T_n f - y_n\|_{\mathcal{B}} \leq \gamma_n$$

holds. Let $\hat{f}_{\gamma_n}$ be the MIND estimator in (3.3)*, and*

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

Then there are constants $C > 0$ and $n_0 \in \mathbb{N}$ only depending on k, d, $\mathcal{B}$, and T, such that

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

where ϑ and ϑ' are given in (3.5)*.*

Proof. The proof follows similarly as that of Theorem 2.3.3, while one should replace S_n by T_n, and the inequality in Theorem 2.2.7 by the one in Proposition 3.2.2. □

Based on such an estimate, we can obtain convergence rates for MIND under certain smoothness assumption, that is, *approximate source conditions* (cf. (2.17) and the discussion thereinafter).

Theorem 3.2.6. *Assume that the operator T is β-smoothing for some $\beta \geq 0$, that $k, d \in \mathbb{N}$, $k + \lfloor\beta\rfloor > d/2$, and that the system $\mathcal{B}$ is normal. Denote by $\hat{f}_{\gamma_n}$ the MIND estimator in* (3.3) *with γ_n given in* (3.4)*, and assume that*

$$\min_{t\geq 0}\big(d_n(t;T) + (\log n)^{r/2} t^{1/2}\big) = \mathcal{O}\left(n^{-\mu}\right) \qquad \text{as } n \to \infty \tag{3.6}$$

for some $0 \leq \mu < 1/2$. *Then it holds almost surely that*

$$\|\hat{f}_{\gamma_n} - f\|_{L^2} = \mathcal{O}\big(n^{-\mu(1-2\vartheta)-\vartheta}(\log n)^{2r\vartheta}\big) \qquad \text{as } n \to \infty, \tag{3.7}$$

with ϑ *in* (3.5a). *If the operator* T *in addition satisfies*

$$\mathcal{C}_0^\infty(\mathbb{T}^d) \subset \mathrm{Ran}(T), \tag{3.8}$$

where $\mathcal{C}_0^\infty(\mathbb{T}^d)$ *consists of all functions in* $\mathcal{C}^\infty(\mathbb{T}^d)$ *with mean zero, then the assertion* (3.7) *also holds in expectation.*

Proof. See Appendix B.2. □

Remark 3.2.7. Note that the additional requirement (3.8) for the convergence rate in expectation is clearly satisfied for all convolution operators that are β-smoothing. If such requirement is violated, we can still obtain an asymptotic estimate for the L^2-risk of MIND by means of a projection trick. More precisely, we consider the projection of MIND $\hat{f}_{\gamma_n}$ onto an L^2-ball of radius κ_n, that is,

$$\hat{f}_{\gamma_n,\kappa_n} := \mathrm{proj}\big(\hat{f}_{\gamma_n}, B_2(\kappa_n)\big) = \hat{f}_{\gamma_n} \min\Big\{\kappa_n/\|\hat{f}_{\gamma_n}\|_{L^2}, 1\Big\}$$

with $B_2(\kappa_n) := \big\{f \colon \|f\|_{L^2} \leq \kappa_n\big\}$. Let us assume that $\kappa_n \sim \log n$, and that n is sufficiently large such that the truth $f \in B(\kappa_n)$. Then it is easy to see that

$$\|\hat{f}_{\gamma_n,\kappa_n} - f\|_{L^2} \leq \|\hat{f}_{\gamma_n} - f\|_{L^2} \quad \text{ and } \|\hat{f}_{\gamma_n,\kappa_n} - f\|_{L^2} \leq 2\kappa_n. \tag{3.9}$$

It further implies that

$$\begin{aligned}\mathbb{E}\Big[\|\hat{f}_{\gamma_n,\kappa_n} - f\|_{L^2}\Big] =& \mathbb{E}\Big[\|\hat{f}_{\gamma_n,\kappa_n} - f\|_{L^2}; \|\xi_n\|_{\mathcal{B}} \leq \gamma_n\Big] + \mathbb{E}\Big[\|\hat{f}_{\gamma_n,\kappa_n} - f\|_{L^2}; \|\xi_n\|_{\mathcal{B}} > \gamma_n\Big] \\ \leq& \mathbb{E}\Big[\|\hat{f}_{\gamma_n} - f\|_{L^2}; \|\xi_n\|_{\mathcal{B}} \leq \gamma_n\Big] + 2\kappa_n \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} > \gamma_n\right\}.\end{aligned}$$

Combining it with Lemma 3.2.5 and Proposition 2.2.4, we can obtain by simple calculations exactly the same estimate for the L^2-risk of $\hat{f}_{\gamma_n,\kappa_n}$ as in (3.7). Furthermore, from (3.9) it follows that such an estimate of the L^2-loss of $\hat{f}_{\gamma_n,\kappa_n}$ also holds almost surely as $n \to \infty$.

Example 5. As an application of Theorem 3.2.6, we consider a simple example that the truth satisfies

$$f \in H_0^k(\mathbb{T}^d).$$

This is the "proper smoothing" case, where the smoothness order k in the regularization term of MIND matches perfectly to that of the underlying truth. From Remark 3.2.4, we obtain an estimate of the approximate source condition (3.6) with $\mu = 0$, i.e.

$$\min_{t \geq 0}\big(d_n(t;T) + (\log n)^{r/2} t^{1/2}\big) \leq d_n(0;T) \leq \|f\|_{H_0^k}.$$

It follows directly from the theorem that

$$\|\hat{f}_{\gamma_n} - f\|_{L^2} = \mathcal{O}\big(n^{-\vartheta}(\log n)^{2r\vartheta}\big) \qquad \text{with } \vartheta \text{ in (3.5a)}$$

holds almost surely and in expectation, provided that the operator T is β-smoothing and satisfies (3.8). In the following, we show that the above estimate is actually minimax optimal over ellipsoids in $H_0^k(\mathbb{T}^d)$ up to a log-factor. Note first that

$$\|\hat{f} - f\|_{L^2} \sim \|T\hat{f} - Tf\|_{H_0^\beta} \qquad \text{for any estimator } \hat{f},$$
$$\text{and} \quad \mathcal{C}_0^\infty(\mathbb{T}^d) \subset T\big(H_0^k(\mathbb{T}^d)\big) := \big\{Tf \colon f \in H_0^k(\mathbb{T}^d)\big\} \subset H_0^{k+\beta}(\mathbb{T}^d).$$

Following the proof of (Nemirovski, 1985, Theorem 1) closely, one can derive that the minimax rate for nonparametric regression with respect to H_0^β-loss over ellipsoids in $H_0^{k+\beta}(\mathbb{T}^d)$ is at least of order $n^{-k/(2k+2\beta+d)} = n^{-\vartheta}$, and will further find that this lower bound still holds for ellipsoids in $T\big(H_0^k(\mathbb{T}^d)\big)$ because it contains $\mathcal{C}_0^\infty(\mathbb{T}^d)$ as a subset. Therefore, the rate for estimating f from the measurement in (2.2) is no faster than $n^{-\vartheta}$, which leads to the almost minimax optimality of MIND in this setting.

3.3. Convergence rates in one dimension

In this section, on the basis of Proposition 2.4.1, we are able to translate the approximate source conditions (3.6) into classical Hölder-type source conditions that relate to the Sobolev smoothness, in the setting of recovering functions defined on one-dimensional domain $\mathbb{T}$. As a consequence, we derive explicit convergence rates for MIND, and discuss the minimax optimality of such rates, which leads naturally to an adaptation phenomenon.

3.3.1. Hölder-type source conditions

Theorem 3.3.1. *Let $d = 1$, $k \in \mathbb{N}$, $\beta \in \mathbb{N}_0$, and $k \geq \beta \geq 0$. Assume moreover that the operator $T \colon L^2(\mathbb{T}) \to L^2(\mathbb{T})$ and its adjoint $T^* \colon L^2(\mathbb{T}) \to L^2(\mathbb{T})$ are β-smoothing, that $\mathcal{B}$ is regular, and that*

$$f = T^*h \qquad \text{for some } h \in H_0^s(\mathbb{T}) \qquad \text{with } k - \beta + 1 \leq s \leq 2k. \tag{3.10}$$

Then the MIND estimator $\hat{f}$ in (3.3), with the threshold γ_n given by (3.4), satisfies almost surely and in expectation that

$$\|\hat{f}_{\gamma_n} - f\|_{L^2} = \mathcal{O}\Big(n^{-\frac{s+\beta}{2s+4\beta+1}}(\log n)^{\frac{2r(s+\beta)}{2s+4\beta+1}}\Big) \qquad \text{as } n \to \infty.$$

Proof. See Appendix B.3. □

Remark 3.3.2. It is worth pointing out that from the proof of Theorem 3.3.1 one can see that the constant hidden in the $\mathcal{O}$-notation takes the form

$$C\|h\|_{H_0^s}^{1-2\vartheta} \qquad \text{with } \vartheta \text{ given in (3.5a)},$$

where $C > 0$ depends only on k, s, and $\mathcal{B}$. The smoothness assumption (3.10) is often referred to as *Hölder-type source conditions*, which are the typical smoothness conditions for linear inverse problems. Under the setting of the theorem, one can show that the condition (3.10) is equivalent to

$$f \in T^*\big(H_0^s(\mathbb{T})\big) = (T^*T)^{1/2}\big(H_0^s(\mathbb{T})\big) = H_0^{s+\beta}(\mathbb{T}) \qquad \text{with } k - \beta + 1 \le s \le 2k.$$

Note that every β-smoothing convolution operator satisfies the requirements on T in the theorem, since its adjoint is again β-smoothing, see Example 4 (b). For such operators, the condition (3.10) can be further rewritten as

$$f \in T^*\big(H_0^s(\mathbb{T})\big) = (T^*T)^{\mu}\big(H_0^{s+(1-2\mu)\beta}(\mathbb{T})\big) \qquad \text{for every } \mu \ge 0.$$

Moreover, if $\mu \in \mathbb{N}_0$, the above relation holds for every operator T such that its adjoint and itself are β-smoothing.

3.3.2. Adaptation property

By $\mathcal{E}(n; \mathcal{F})$ we denote the minimax L^2-risk of the statistical inverse problem (3.1) over a class $\mathcal{F}$ of functions for a fixed sample size n (cf. Section 2.4.3). We consider in particular the *source sets* with respect to the operator T

$$H_L^s(T^*) := \Big\{ f \colon f = T^*h,\ \|h\|_{H_0^s} \le L \Big\} \qquad \text{for } s \ge 0 \text{ and } L > 0$$

as the function classes $\mathcal{F}$. Following the discussion in Example 5, one can show that the minimax L^2-risk $\mathcal{E}\big(n; H_L^s(T^*)\big)$ is of the same order in n as the minimax H^β-risk of nonparametric regression over $H_L^s(TT^*)$, provided that the operator T is β-smoothing. If the adjoint T^* is also β-smoothing, then we have for some constants L', L'',

$$S_{L'}^{s+2\beta,2} \subset H_L^s(TT^*) \subset S_{L''}^{s+2\beta,2} \qquad \text{with } S_L^{s,2} \text{ given by (2.24)}.$$

By (2.26) it further implies that for n sufficiently large

$$\mathcal{E}\big(n; H_L^s(T^*)\big) \ge C\Big(\frac{\sigma^2}{n}\Big)^{\frac{s+\beta}{2s+4\beta+1}} L^{\frac{2\beta+1}{2s+4\beta+1}},$$

where constant $C > 0$ depends only on s, β and T.

Comparing the lower bound of $\mathcal{E}\big(n; H_L^s(T^*)\big)$ with the upper bounds in Theorem 3.3.1 and Example 5, we see that the convergence rates of MIND in (3.3) are minimax optimal up to a log-factor over

$$H_L^s(T^*) \qquad \text{for every } s \in \{k-\beta\} \cup [k-\beta+1, 2k].$$

It is worth noting that our convergence rates do not rely on a precise knowledge of the smoothness class of the function f, because the threshold γ_n in (3.4) is independent of the truth f. This suggests that in practice the smoothing order k of the regularization term need only be a rough guess of the actual smoothness s of f. Furthermore, if the true smoothness s is available, one should choose $k = s/2$ rather than $k = s + \beta$ for the sake of numerical stability.

Put differently, the convergence result above says that MIND automatically adapts to the smoothness of the truth f over a range of source sets $H_L^s(T^*)$, including both the "proper smoothing" case, $s = k - \beta$ (cf. Example 5), and the "under-smoothing" case, $s \in [k-\beta+1, 2k]$, where the truth f is smoother than that is required by the regularization term. We refer to such property as *partial adaptation*. As a by-product, in the particular case of $T = \mathrm{I}$, it implies that MIND attains the minimax optimal rates up to a log-factor with respect to L^2-risk over Sobolev ellipsoids

$$H_L^s(\mathrm{I}^*) = S_L^{s,2} \qquad \text{with } s \in \{k\} \cup [k+1, 2k]$$

for nonparametric regression. This reproduces part of our results in the previous chapter, see Section 2.4 for a more comprehensive analysis in this special setting. Moreover, although we were not able to derive (nearly) optimal rates for the "over-smoothing" case, we believe that similar result as in Section 2.4.2 possibly holds if Sobolev spaces $W_0^{s,\infty}(\mathbb{T})$ are considered instead of $H_0^s(\mathbb{T})$.

Finally, we point out that one can also study the penalized version of MIND in (1.12), as a generalization of the penMIND estimator in (2.27) to statistical inverse problems (3.1), and derive similar convergence results as those developed in this and the previous sections, in an analogous way as in Section 2.5.

3.4. Numerical results

We now investigate the practical performance, i.e. the finite sample behavior, of the MIND estimator on some simulated examples. From Section 2.6.1, we see that the MIND estimator defined by (3.3), being the solution of a non-smooth convex optimization problem, can be efficiently solved by an ADMM algorithm. As a practical adjustment to enhance the performance of MIND for finite sample sizes, we choose the threshold γ_n as the α-quantile $\gamma_n(\alpha)$ of the multiresolution norm of random error by (2.41), rather than the one

in (3.4), see Section 2.7.1 for explanation. In the coming simulations, the significance level α for $\gamma_n(\alpha)$ is set to 0.1, and the quantile $\gamma_n(\alpha)$ is estimated by 10^5 independent random draws of Monte-Carlo simulations. The random error is assumed to be i.i.d. Gaussian distributed with zero mean and a known variance. Moreover, for the definition of the multiresolution norm, we always select the 2-partition system (see Definition 2.2.3), which is indeed the sparsest system that satisfies the requirement of all the asymptotic analysis. Implementation is provided in the MATLAB package "MOP" (see Section 2.6.2).

3.4.1. Deconvolution in one dimension

First of all, we evaluate the performance of MIND for a one-dimensional deconvolution problem if the convolution kernel is 1-smoothing. The spatially variable function Bumps (Donoho and Johnstone, 1994) is chosen as the test signal. For comparison, we consider a Sobolev regularization method (SOB) defined by

$$\min_f \|T_n f - y_n\|_2 + \lambda \|D^k f\|_{L^2}^2 \tag{3.11}$$

which can be regarded as a generalization of the smoothing spline estimator (2.43) to statistical inverse problems. Moreover, by DAN we denote a variant of the Dantzig selector in (1.6), which is given by

$$\min_f \|D^k f\|_{L^2} \qquad \text{subject to } \|T_n^*(T_n f - y_n)\|_{L^\infty} \leq \gamma. \tag{3.12}$$

In order to study the impact of different data fidelity terms, we pick the same regularization term, H_0^k-norm, for the three estimators, and choose in particular $k = 1$ for the smoothness order. The balancing parameter λ for SOB is manually tuned to give the best performance, and the threshold γ for DAN is chosen similarly as MIND, that is, the 0.1-quantile of $\|T_n^* \xi_n\|_{L^\infty}$. The simulation results are collected in Figure 3.1. It shows that MIND is able to recover all the bumps of various scales and locations, while presents slight distortions form the truth. This reflects our theoretical finding that MIND attains nearly optimal rates for $H_0^s(\mathbb{T})$ with $s = 1$, or $2 \leq s \leq 3$ in this particular setting. By a sharp contrast, SOB and DAN only detect the isolated bumps, and fail to discern bumps that are closely located.

For a robustness study, we repeat the previous experiment but with a Gaussian convolution kernel. This lies outside of our asymptotic analysis, since such a convolution operator is not β-smoothing for any $\beta \geq 0$. From Figure 3.2, one can see that MIND detects most of the bumps, and still outperforms both SOB and DAN. Compared to SOB, DAN recovers more bumps, but meanwhile introduces many artificial bumps in the nearly constant part of the true signal. The SOB performs similarly as the previous experiment, finding only bumps of large scales.

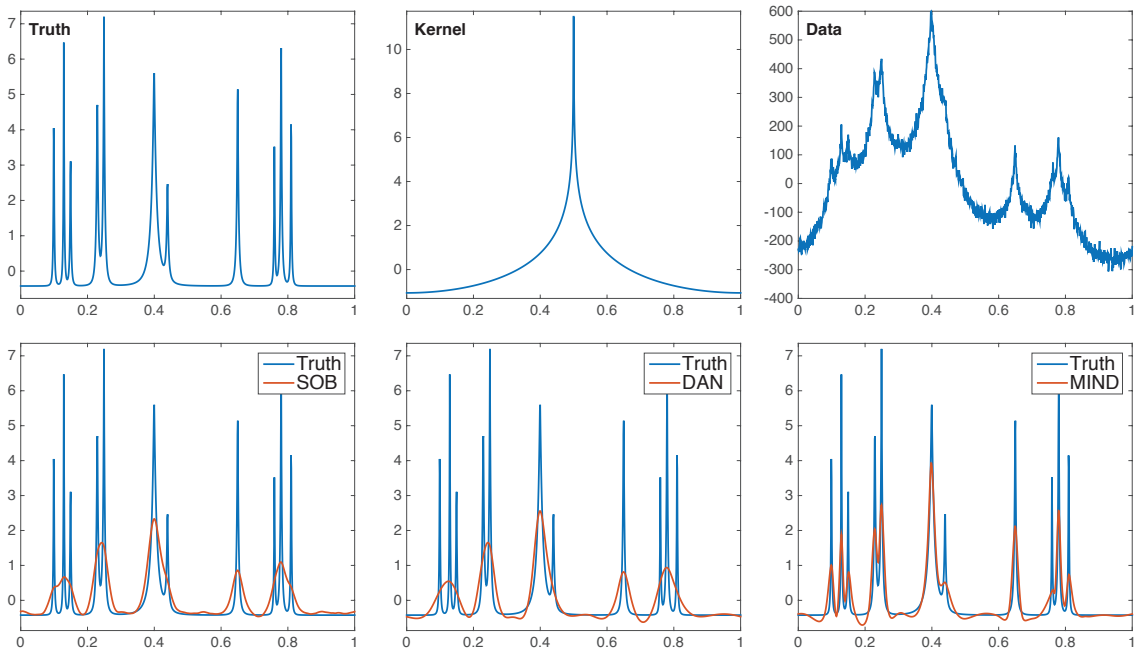

Figure 3.1.: Deconvolution of Bumps signal with 1-smoothing kernel by various methods: SOB in (3.11), DAN in (3.12), and the proposed MIND in (3.3) (number of samples $n = 2^{10}$, noise level $\sigma = 0.08\|Tf\|_{L^2}$).

3.4.2. Imaging in two dimension

We next continue to study the multiresolution norm as the data fidelity term in imaging applications. In particular, let us consider the recovery of the Shepp-Logan phantom (Shepp and Logan, 1974) from noisy measurements of the Radon transform, i.e. the model (3.1) with T chosen as the Radon transform. This is a standard example for computerized tomography (see Natterer, 2001, for an overview).

Recall that the total variation (TV) semi-norm of functions $f\colon \mathbb{T}^d \to \mathbb{R}$ is defined as

$$\|f\|_{\mathrm{TV}} := \sup_{g\in\mathcal{V}} \int_{\mathbb{T}^d} (-f \operatorname{div} g)dx,$$

where the set of test functions

$$\mathcal{V} := \left\{ g\colon \mathbb{T}^d \to \mathbb{R}^d \,;\, g \text{ is differentiable, and } |g(x)| \le 1 \text{ for all } x \in \mathbb{T}^d \right\}.$$

In particular, for differentiable functions f, it reduces to the L^1-norm of the gradient of f, i.e. $\|f\|_{\mathrm{TV}} \equiv \|Df\|_{L^1}$, see e.g. (Ambrosio et al., 2000) for further details. It is well known

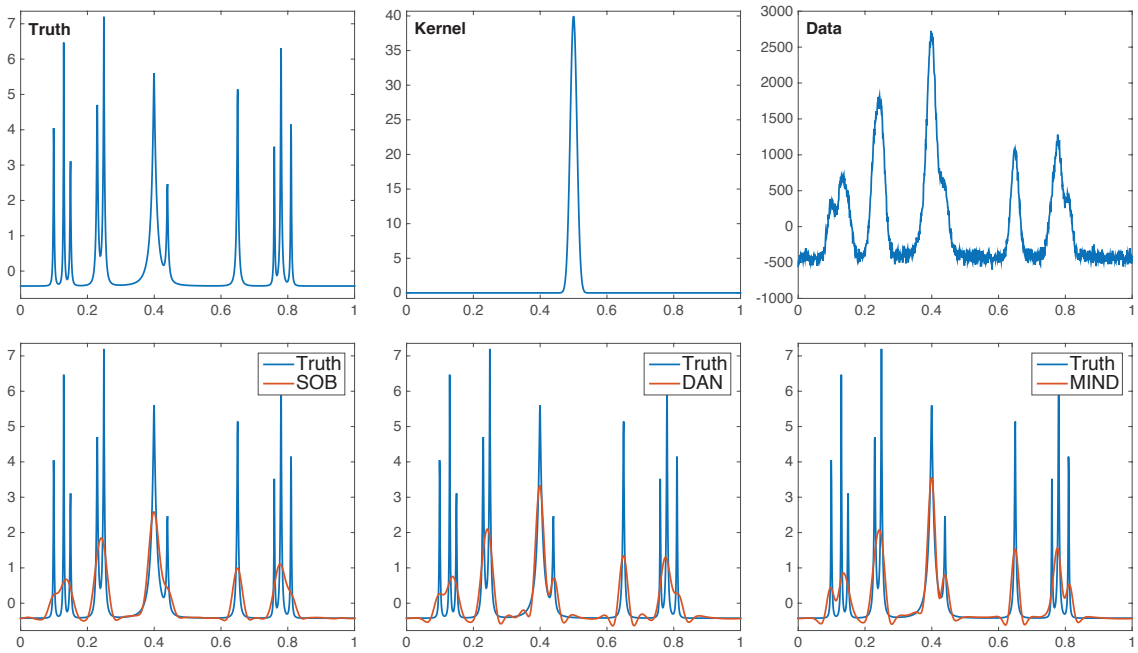

Figure 3.2.: Deconvolution of Bumps signal with Gauss kernel $\rho(x) \sim \exp(-\frac{x^2}{2\sigma^2})$ with $\sigma = 0.01$ by various methods: SOB in (3.11), DAN in (3.12), and the proposed MIND in (3.3) (number of samples $n = 2^{10}$, noise level $\sigma = 0.08\|Tf\|_{L^2}$).

that the TV semi-norm favors carton-like images (including Shepp-Logan phantom as an example) if applied as the regularization term, because it is effective at preserving sharp edges whilst smoothing away noise in flat regions, mainly due to the sparsity enhancing nature of L^1-norm. Thus, as an extension of MIND, we introduce the *MIND-TV* estimator defined by

$$\min_f \|f\|_{\mathrm{TV}} \qquad \text{subject to } \|T_n f - y_n\|_{\mathcal{B}} \leq \gamma. \tag{3.13}$$

Similar to MIND, the threshold parameter γ here is selected as $\gamma_n(\alpha)$ by (2.41) with $\alpha = 0.1$. Meanwhile, we consider the famous *filtered back projection* (FBP) reconstruction method, and the classical *TV-regularization* (TVreg) method given by

$$\min_f \|T_n f - y_n\|_2 + \lambda \|f\|_{\mathrm{TV}}. \tag{3.14}$$

The Ram-Lak filter is used for FBP, see (Natterer, 2001, Section V.1) for details. The penalization parameter λ of TVreg is manually tuned for the best visual quality of the reconstruction.

A comparison result among the three methods is given in Figure 3.3. It shows that TVreg and MIND-TV are comparable, and perform significantly better than FBP, concerning both

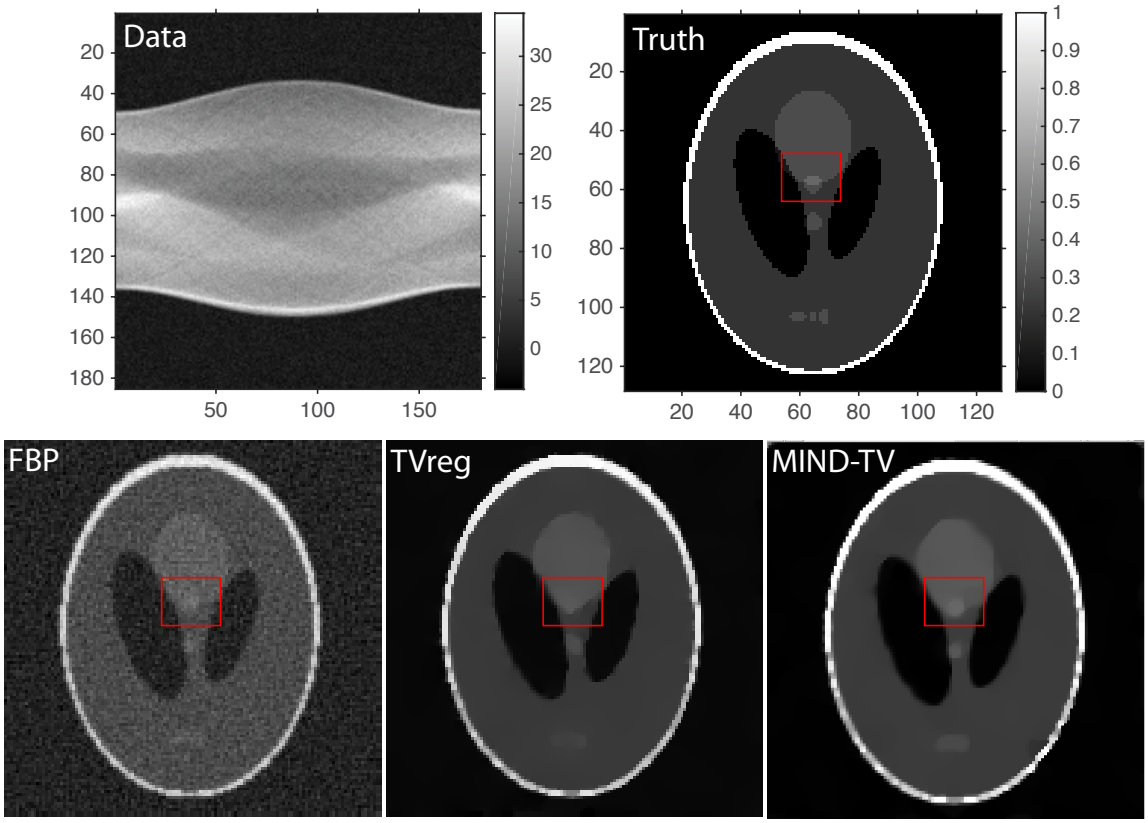

Figure 3.3.: Reconstruction for Radon data by FBP, TVreg in (3.14) and MIND-TV (3.13), with noise level $\sigma = 1$.

the removal of noise and the recovery of features. Furthermore, the MIND-TV detects one more feature of the truth, which is marked by a red rectangle, compared to TVreg. This is a consequence of the favorable multiscale nature of the multiresolution norm, which is not shared by the global method TVreg. To get an impression about the performance of Algorithm 1, we further illustrate the details of each iteration for the computation of MIND-TV in terms of objective values and gaps of the multiscale constraint in Figure 3.4.

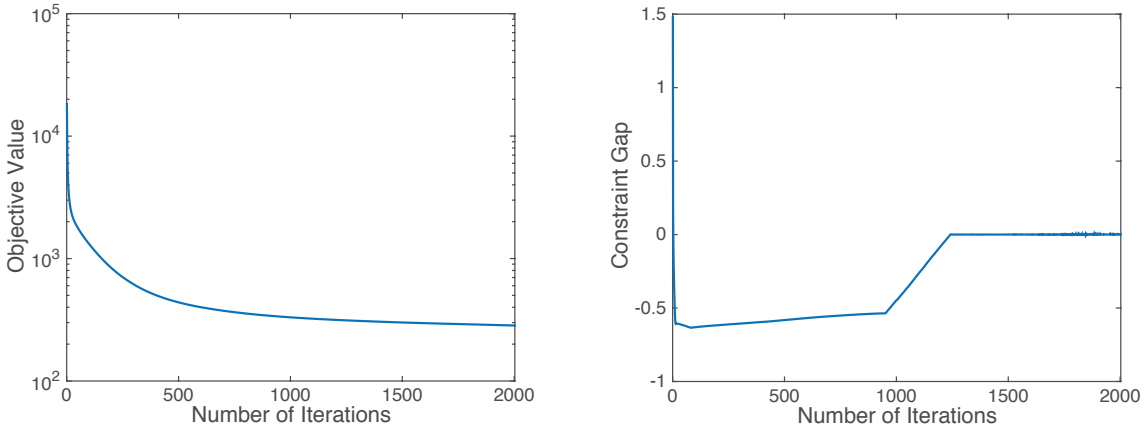

Figure 3.4.: Convergence curves of ADMM iterations for the computation of MIND-TV from the data in Figure 3.3. The objective value $\|f_l\|_{\mathrm{TV}}$ of the l-th iteration, as well as the corresponding gap of the multiscale constraint $\|T_n f_l - y_n\|_{\mathcal{B}} - \gamma$, is plotted versus the iteration l, cf. Algorithm 1.

4. Discussion and Outlook

In the settings of nonparametric regression and statistical inverse problems, we have introduced a constrained variational estimator, MIND, which minimizes the L^2-norm of the k-th order derivatives (i.e. the homogenous H^k-norm), 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 threshold γ_n depending on the sample size n. The idea behind this approach is that the multiresolution norm effectively allows to differentiate between smooth functions and noise, as the multiresolution norm of a continuous function 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 $r > 1/2$ 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 work was to underpin the already known empirically good performance of MIND in several special cases by some theoretical evidence. For a general dimension d, from interpolation inequalities for the multiresolution norm and Sobolev norms, we derive asymptotic convergence rates of MIND for nonparametric regression and statistical inverse problems, provided that $f \in H_0^k(\mathbb{T}^d)$ and one applies regularization with the homogeneous H^k-norm. Moreover, these rates turn out to be minimax optimal up to a logarithmic factor. In order to derive convergence rates for different smoothness classes or source sets, 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. With this approach we have obtained conditions that guarantee certain convergence rates. 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 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. For nonparametric regression of functions and derivatives, this mainly 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. In the case of statistical inverse problems, we have derived by a similar argument optimal convergence rates for functions f in source sets $T^*\big(H_0^s(\mathbb{T})\big) \equiv H_0^{s+\beta}(\mathbb{T})$ for every $k - \beta + 1 \leq s \leq 2k$, provided that the forward operator T and its adjoint T^* are both β-smoothing. However, these results are only concerned with functions f that are of higher regularity than assumed a-priori. By contrast, for nonparametric regression (i.e. $T = \mathrm{I}$ the identity operator) 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 $s \in [1, k] \cup [k+1, 2k]$ for nonparametric regression, and $s \in \{k\} \cup [k+1, 2k+\beta]$ for statistical inverse problems), making it truly an adaptive method. Note further that all the theoretical results also hold for the practical adjustment $\gamma_n := \gamma_n(\alpha)$ the α-quantile of $\|\xi_n\|_{\mathcal{B}}$, provided that the significance level α goes to zero sufficiently fast. This allows for the statistical inference that the truth is no smoother than the estimator with probability at least $1 - \alpha$, namely the smoothness guarantee. 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.

In addition, we stress that the argumentation developed here can apply to the analysis of a penalized version of MIND, the penMIND estimator, as well. As an illustration, we have presented convergence rates of penMIND for nonparametric regression. Moreover, with the Lepskiĭ principle for the choice of balancing parameter, the penMIND estimator performs adaptively over the same range of smoothness classes as MIND, while its finite sample behavior is still improvable due to a certain constant that is not tight. If we replace such a theoretical constant by a well-tuned one, the practical performance of penMIND will be significantly enhanced and even comparable to that of MIND. For comparison, we recall that no adaptation result is available for the Nemirovski's estimator, which can be viewed as a smoothness-constrained version of MIND. Thus, it is rather intricate to calibrate this estimator in practice. Empirically, the Nemirovski's estimator shows an inadequate performance regarding the removal of noise, even with the oracle choice of parameter $\eta := \|D^k f\|_{L^p}$ of the truth f. An overview of these three related estimators is given in Table 4.1.

Table 4.1.: A summary of the three multiscale variational statistical estimations.

	MIND	penMIND	Nemirovski's
Choice of parameters	Quantiles of $\|\xi_n\|_{\mathcal{B}}$	Lepskiĭ principle	Oracle
Statistical property	Partial adaptation & Smoothness guarantee	Partial adaptation	No adaptation
Practical performance	Satisfactory	(Less) satisfactory	Unsatisfactory

There are several questions still open concerning MIND. First of all, almost all concrete results concerning convergence rates in this work were derived for a one-dimensional setting. In higher dimensions we only have the (somehow generic) result mentioned in Examples 3 and 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. Moreover, in the literature, there is few results on the size of approximation coefficients, which are necessary for our analysis (cf. the discussion below Example 3). 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. As an alternative, one could consider functions that satisfy zero boundary conditions. Throughout the analysis, one only needs to adjust the approximation B-splines at boundaries. Since B-splines have compact supports, the impact of such modification, we guess, will vanish asymptotically. Furthermore, there are still some regularity classes, for which we do not know whether our proposed method provides optimal convergence rates. 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. In contrast, we believe that the upper limit of smoothness order for the adaptation range is sharp, i.e. the full adaptation may not be possible for MIND.

Finally, we collect some stimulating directions for extending both our methodology and the theoretical analysis, which we plan to explore in our future research.

(a) Recall from one dimensional nonparametric regression that we do obtain convergence

rates for functions $f \in W_0^{s,p}(\mathbb{T})$ with $p < 2$, but 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, which can be naturally generalized to TV semi-norms if $k = 1$. In addition, the combination of the TV regularization and the multiscale constraint is shown to be quite promising in various imaging and image processing applications, see Section 3.4.2, and (Frick et al., 2012, 2013) for example. Note that the point evaluation becomes problematic for bounded variation functions when $d \geq 2$. Thus, as a first step, we should reinterpret the measurements as local averages rather than point-wise values, that is,

$$y_n(x) = \int_{x+[-\frac{1}{2m},\frac{1}{2m})} (Tf)(z)dz + \xi_n(x) \qquad \text{for } x \in \Gamma_n,$$

where $m := n^{1/d}$. The definitions of the multiresolution norm and its dual need to be adjusted in a similar way as well.

(b) For nonparametric regression and statistical inverse problems we always assume that the random error is distributed according to a sub-Gaussian law. In many applications (see e.g. Frick et al., 2013; Aspelmeier et al., 2015), this assumption is only an approximation, and sometimes unsatisfactory. It is therefore interesting to extend our argumentation to measurement models with other noise distributions. For instance, one could consider the standard *exponential family* distributions, which have densities $\exp\big(\langle\theta,\cdot\rangle - \psi(\theta)\big)$ for some parameter $\theta \in \mathbb{R}^d$. Motivated by the relation between the multiresolution norm and multiple testing (see Section 1.2), we suggest to redefine the multiscale constraint on residuals by

$$\mathcal{L}_{\mathcal{B}}(Tf, y_n) := \sup_{B\in\mathcal{B}} \sqrt{2(\#B \cap \Gamma_n) J\big(\bar{y}_B, (\overline{Tf})_B\big)} \leq \gamma_n,$$

where $J(x,\theta) := \psi^*(x) + \psi(\theta) - \langle x,\theta\rangle$ with $\psi^*(x) := \sup_\theta \langle x,\theta\rangle - \psi(\theta)$ the Legendre-Fenchel conjugate of ψ, and $(\cdot)_B$ denotes the average over $B\cap\Gamma_n$ (cf. Frick et al., 2014). Note that this reduces to $\|Tf - y_n\|_{\mathcal{B}}$ when the noise distribution is Gaussian. A first and key step is to establish an interpolation-type inequality between the multiscale functional $\mathcal{L}_{\mathcal{B}}$, the loss functional, and the regularization functional, which is probably more involved since $\mathcal{L}_{\mathcal{B}}$ is in general not necessarily a norm.

(c) The statistical optimality we are concerned throughout lies in the classical minimaxity paradigm with respect to the L^q-loss for $1 \leq q \leq \infty$. In the development over nearly half a century, there have already been a rich and diverse collection of optimal statistical approaches for nonparametric regression and statistical inverse problems (cf. Section 1.1.1). However, the optimality theory does not fully reflect the empirically different performances of various approaches for a given sample size. We

doubt that this might be a consequence of the choice of loss functionals, which have a global nature due to the integration, and for instance cannot differentiate random deviations from deterministic ones, as we see in Example 1. A perhaps reasonable choice would be something like the multiresolution norm that takes a range of scales into account. As an alternative remedy, instead of a fixed signal, we suggest to consider a sequence of signals that depends on the sample size n and tends to be more challenging to recover as n increases. In this way, we believe that the real difficulty for finite sample sizes will be better reflected in the asymptotic analysis.

A. Proofs of Chapter 2

A.1. Nemirovski's interpolation inequality

In this section, we will prove the interpolation inequality (in Theorem 2.2.7) between multiresolution norms and Sobolev semi-norms. For brevity, by $\|f\|_{B,*}$ we denote $\|f\mathbf{1}_B\|_*$ for any set B, and norm $\|\cdot\|_*$. We first need some basic properties of d-dimensional polynomials, which are summarized in the following two lemmata. Since these results are known, we only present a proof when there is no proper reference.

Lemma A.1.1. *Let $B_{\theta,r} := \{x \in \mathbb{R}^d; \|x-\theta\|_\infty \le r\}$, and $\mathcal{P}_m := \{$polynomials of degree $\le m$ on $\mathbb{R}^d\}$. Then for every $p \in \mathcal{P}_m$, $\gamma \in (0,1)$, and every B_{θ_0,r_0}, there exists a sub-cube $B_{\theta_1,r_1} \subset B_{\theta_0,r_0}$ such that*

$$\min_{x\in B_{\theta_1,r_1}} |p(x)| \ge \gamma \max_{x\in B_{\theta_0,r_0}} |p(x)|,$$

and the ratio r_1/r_0 depends on γ, m and d only.

Proof. We w.l.o.g. assume $m \ge 1$. Let $x^* := \arg\max_{x\in B_{\theta_0,r_0}} |p(x)|$, $r_1 := r_0(1-\gamma)/(2dm^2)$, and θ_1 satisfy that $B_{\theta_1,r_1} \ni x^*$. Let also $x^{**} := \arg\min_{x\in B_{\theta_1,r_1}} |p(x)|$. Note that

$$\begin{aligned}
|p(x^*)| - |p(x^{**})| \le& |p(x^*) - p(x^{**})| = |\int_0^1 \frac{d}{dt} p\big(x^{**} + t(x^* - x^{**})\big)dt| \\
\le& \sum_{i=1}^d \int_0^1 |\partial_i p\big(x^{**} + t(x^* - x^{**})\big)||x^*_{(i)} - x^{**}_{(i)}|dt \\
\le& 2r_1 \sum_{i=1}^d \max_{x\in B_{\theta_1,r_1}} |\partial_i p(x)| \le 2r_1 \sum_{i=1}^d \max_{x\in B_{\theta_0,r_0}} |\partial_i p(x)| \\
\le& \frac{2r_1 dm^2}{r_0} |p(x^*)|,
\end{aligned}$$

where the last inequality is because of the Markov brothers' inequality (see e.g. DeVore and Lorentz, 1993, Chapter 4, Theorem 1.4). It follows that

$$|p(x^{**})| \ge \big(1 - \frac{2r_1 dm^2}{r_0}\big)|p(x^*)| = \gamma |p(x^*)|,$$

which is equivalent to the assertion. □

Lemma A.1.2 (Brenner and Scott (2008), Proposition 4.3.2). *Let $B \subset [0,1]^d$ be a cube,* $\mathrm{diam}(B)$ *its diameter, and $U_r \subset B$ an ℓ^2-ball with radius r. Then, for every $f \in W^{k,p}([0,1]^d)$ with $k > d/p$ or $k = d$ and $p = 1$, there is a polynomial f_{U_r} of degree $\leq k-1$, depending on f and U_r only, such that*

$$\|f - f_{U_r}\|_{B,L^\infty} \leq C\mathrm{diam}(B)^{k-d/p}\|D^k f\|_{B,L^p},$$

where the constant C depends on k, d, p and the ratio $\mathrm{diam}(B)/r$ *only.*
Remark A.1.3. The polynomial f_{U_r} is actually the *averaged Taylor polynomial* of order k (i.e. degree $k-1$) over $U_r := \{x \in \mathbb{R}^d; \|x - x_0\|_2 \leq r\}$, that is,

$$f_{U_r}(x) := \int_{U_r} T_y^k f(x)\phi(y)dy,$$

where the Taylor polynomial

$$T_y^k f(x) := \sum_{|\alpha|<k, \alpha \in \mathbb{N}_0^d} \frac{1}{\alpha!} D^\alpha f(y)(x-y)^\alpha,$$

and

$$\phi(x) := \begin{cases} C\exp\Big(-r^2/\big(r^2 - \|x - x_0\|_2^2\big)\Big) & \text{if } x \in U_r \\ 0 & \text{otherwise} \end{cases}$$

with C such that $\int_{\mathbb{R}^d} \phi(x)dx = 1$.

We also need an interpolation inequality between derivatives of different order.
Lemma A.1.4 (Gagliardo-Nirenberg interpolation inequality (Nirenberg, 1959)). *Let $1 \leq p, r \leq \infty$, $l \in \{0, \ldots, k-1\}$, $k \in \mathbb{N}$, and $q, \gamma \in \mathbb{R}$ such that*

$$\frac{d}{q} - l = \gamma\frac{d}{r} + (1-\gamma)(\frac{d}{p} - k) \quad \text{and } 0 \leq \gamma \leq \frac{k-l}{k}.$$

Then, for every $f \in W^{k,p}([0,1]^d)$, it holds that

$$\|D^l f\|_{L^q} \leq C_1\|f\|_{L^r}^\gamma \|D^k f\|_{L^p}^{1-\gamma} + C_2\|f\|_{L^s}, \tag{A.1}$$

where $s > 0$ is arbitrary (constants C_1, C_2 depend only on d, k, p, r) with the following exceptional cases:

(i) If $l = 0$, $k < \frac{d}{p}$ and $r = \infty$, we assume in addition that either f tends to zero at boudary or $f \in L^{\tilde{q}}(\mathbb{R}^d)$ for some finite $\tilde{q} > 0$;

(ii) If $1 < p < \infty$ and $k - l - \frac{d}{p}$ is a nonnegative integer, then the inequality (A.1) *does not hold for $\gamma = 0$.*

Proposition A.1.5. *Let $0 < \lambda < \infty$, $1 \le p, q, r \le \infty$, $l \in \{0, \dots, k-1\}$, and $k \in \mathbb{N}$. Let also $k > d/p$ or $k = d$ and $p = 1$. Then, for every $f \in W^{k,p}([0,1]^d)$, it holds that either*

$$\|D^k f\|_{L^p} \le \lambda \|f\|_{L^\infty} \quad \text{and then } \|D^l f\|_{L^q} \le C_1 \|f\|_{L^\infty},$$

or

$$\|D^l f\|_{L^q} \le C_2 \|f\|_{L^r}^{\gamma} \|D^k f\|_{L^p}^{1-\gamma},$$

where

$$\gamma \equiv \gamma(d,k,l,p,q,r) := \begin{cases} \frac{k-l}{k} & \text{if} \quad \frac{k}{q} \ge \frac{k-l}{r} + \frac{l}{p}, \\ \frac{k-l-d/p+d/q}{k-d/p+d/r} & \text{if} \quad \frac{k}{q} \le \frac{k-l}{r} + \frac{l}{p}; \end{cases}$$

and constants C_1, C_2 depend only on d, k, p, r and λ.

Proof. This is an application of Lemma A.1.4. Obviously, exception (i) does not happen. We first consider the case when $\frac{k}{q} \le \frac{k-l}{r} + \frac{l}{p}$. By the choice of $\gamma = \gamma(d,k,l,p,q,r)$, we have

$$\frac{d}{q} - l = \gamma \frac{d}{r} + (1-\gamma)(\frac{d}{p} - k) \quad \text{and } 0 < \gamma \le \frac{k-s}{k},$$

which implies exception (ii) does not happen either. It follows, in particular, that

$$\|D^l f\|_{L^q} \le C_1 \|f\|_{L^r}^{\gamma} \|D^k f\|_{L^p}^{1-\gamma} + C_2 \|f\|_{L^r}.$$

If $\|D^k f\|_{L^p} \le \lambda \|f\|_{L^\infty}$, then

$$\|D^l f\|_{L^q} \le C_1 \|f\|_{L^\infty}^{\gamma} (\lambda \|f\|_{L^\infty})^{1-\gamma} + C_2 \|f\|_{L^\infty} \le C_3 \|f\|_{L^\infty}.$$

If $\|D^k f\|_{L^p} > \lambda \|f\|_{L^\infty}$, then

$$\|D^l f\|_{L^q} \le C_1 \|f\|_{L^r}^{\gamma} \|D^k f\|_{L^p}^{1-\gamma} + C_2 \|f\|_{L^r}^{\gamma} (\lambda^{-1} \|D^k f\|_{L^p})^{1-\gamma} \le C_4 \|f\|_{L^r}^{\gamma} \|D^k f\|_{L^p}^{1-\gamma}.$$

The case when $\frac{k}{q} \ge \frac{k-l}{r} + \frac{l}{p}$ follows from the first case by noticing that in this case γ does not depend on q and that $\|D^l f\|_{L^q} \le \|D^l f\|_{L^{q'}}$ if $q \le q'$. □

Remark A.1.6. The facts of Lemma A.1.1 and Proposition A.1.5 also appear in (Nemirovski, 1985), while here we give elementary proofs based on results that are relatively more well-known.

We now are ready to present the proof of Theorem 2.2.7.

Proof (of Theorem 2.2.7). In this proof, by C (with subscripts) we denote positive constants depending on c, k, d and p only. Let us define

$$\tilde{B} := B \cap [0,1]^d \qquad \text{for each cube } B.$$

It follows from Lemma A.1.2 that, for every $f \in W^{k,p}([0,1]^d)$ and every cube B with its center in $[0,1]^d$, there exists a polynomial f_B of degree $(k-1)$, corresponding to a maximal ball $U_r \subset \tilde{B} \equiv B \cap [0,1]^d$ such that

$$\|f - f_B\|_{\tilde{B},L^\infty} \le C_1 \mathrm{diam}(\tilde{B})^{k-d/p} \|D^k f\|_{\tilde{B},L^p}.$$

Let us call a cube B *regular*, if

$$\|f\|_{\tilde{B},L^\infty} \geq 4C_1 \mathrm{diam}(\tilde{B})^{k-d/p} \|D^k f\|_{\tilde{B},L^p}.$$

It implies for a regular cube B,

$$\|f - f_B\|_{\tilde{B},L^\infty} \leq \frac{1}{4}\|f\|_{\tilde{B},L^\infty}. \tag{A.2}$$

In the following we consider separately two cases.

Case I. The cube $[0,1]^d$ is not regular.

Let $U := \{x \in (0,1)^d; f(x) \neq 0\}$. It is easily seen that for every $x \in U$ there is a *maximal* regular cube B_x containing x as its center (i.e. the one with the largest radius). Since $[0,1]^d$ is not regular, every maximal regular cube B_x must satisfy

$$\|f\|_{\tilde{B}_x,L^\infty} = 4C_1 \mathrm{diam}(\tilde{B}_x)^{k-d/p} \|D^k f\|_{\tilde{B}_x,L^p}.$$

It follows that

$$\|f\|_{\tilde{B}_x,L^\infty} \leq C_2 |\tilde{B}|^{\frac{k}{d}-\frac{1}{p}} \|D^k f\|_{\tilde{B},L^p}.$$

By Besicovitch's covering theorem (Besicovitch, 1945, 1946, 1947), we can extract a countable sub-system $\mathcal{A}$ from $\{B_x; x \in U\}$, such that

$$U \subset \bigcup_{B\in\mathcal{A}} \tilde{B} \qquad \text{and } \#\{B \ni x; B \in \mathcal{A}\} \leq C_3 \quad \text{for every } x \in U.$$

Let us set

$$r \in [\frac{2k+d}{d}p, \infty), \quad A := \sup_{B\in\mathcal{A}} |\tilde{B}|^{1/2} \|f\|_{\tilde{B},L^\infty} \quad \text{and } \zeta := \frac{1/r + k/d - 1/p}{1/2 + k/d - 1/p}.$$

Then we have

$$\begin{aligned}
\|f\|_{L^r}^r &= \int_U |f(x)|^r dx \leq \sum_{B\in\mathcal{A}} \int_{\tilde{B}} |f(x)|^r dx \\
&\leq \sum_{B\in\mathcal{A}} |\tilde{B}| \|f\|_{\tilde{B},L^\infty}^r = \sum_{B\in\mathcal{A}} \Big(|\tilde{B}|^{1/2}\|f\|_{\tilde{B},L^\infty}\Big)^{r\zeta} \Big(|\tilde{B}|^{-(\frac{k}{d}-\frac{1}{p})}\|f\|_{\tilde{B},L^\infty}\Big)^{r-r\zeta} \\
&\leq C_2^{r-r\zeta} A^{r\zeta} \sum_{B\in\mathcal{A}} \|D^k f\|_{\tilde{B},L^p}^{r-r\zeta} \\
&\leq C_2^{r-r\zeta} A^{r\zeta} \Big(\sum_{B\in\mathcal{A}} \|D^k f\|_{\tilde{B},L^p}^p\Big)^{(r-r\zeta)/p} \qquad (\text{since } r - r\zeta \geq p) \\
&\leq C_2^{r-r\zeta} C_3^{(r-r\zeta)/p} A^{r\zeta} \|D^k f\|_{L^p}^{r-r\zeta}.
\end{aligned}$$

That is

$$\|f\|_{L^r} \le C_4^{1-\zeta} A^{\zeta} \|D^k f\|_{L^p}^{1-\zeta}.$$

By letting $r \to \infty$, we see that the above inequality is valid for $r \in [\frac{2k+d}{d}p, \infty]$.

By definition of A, there is a regular cube B such that

$$\|f\|_{\tilde{B},L^\infty} |\tilde{B}|^{1/2} \ge \frac{1}{2} A.$$

Since B is regular, we have from (A.2) that

$$\frac{3}{4}\|f\|_{\tilde{B},L^\infty} \le \|f_B\|_{\tilde{B},L^\infty} \le \frac{5}{4}\|f\|_{\tilde{B},L^\infty}.$$

Since f_B is a polynomial of degree $\le k-1$, by Lemma A.1.1 there is a cube $B_* \subset \tilde{B}$ such that

$$|B_*| \ge C_5 |\tilde{B}| \qquad \text{and } |f_B(x)| \ge \frac{1}{2}\|f\|_{\tilde{B},L^\infty} \text{ for every } x \in B_*. \tag{A.3}$$

It together with (A.2) implies that

$$|f(x)| \ge \frac{1}{4}\|f\|_{\tilde{B},L^\infty} \text{ if } x \in B_* \implies A \le 8 C_5^{-1/2} |B_*|^{1/2} \min_{x \in B_*} |f(x)|.$$

Thus, we have for $r \in [\frac{2k+d}{d}p, \infty]$,

$$\begin{aligned} \|f\|_{L^r} \le & C_4^{1-\zeta} \big(8 C_5^{-1/2}\big)^{\zeta} \Big(|B_*|^{1/2} \min_{x \in B_*} |f(x)| \Big)^{\zeta} \|D^k f\|_{L^p}^{1-\zeta} \\ \le & C_6 \Big(|B_*|^{1/2} \min_{x \in B_*} |f(x)| \Big)^{\zeta} \|D^k f\|_{L^p}^{1-\zeta}. \end{aligned} \tag{A.4}$$

If $|B_*| > cn^{-1}$, since $\mathcal{B}$ is a c-normal system, there is a cube $B_{**} \in \mathcal{B}$ such that $|B_{**}| \ge c^{-1}|B_*|$ and $B_{**} \subset B_*$, and we get

$$\|S_n f\|_{\mathcal{B}} \ge (\#B_{**} \cap \Gamma_n)^{1/2} \min_{x \in B_{**}} |f(x)| \ge C_7 n^{1/2} |B_{**}|^{1/2} \min_{x \in B_{**}} |f(x)|,$$

where Γ_n is the grid. Then, we have

$$\|f\|_{L^r} \le C_8 n^{-\zeta/2} \|S_n f\|_{\mathcal{B}}^{\zeta} \|D^k f\|_{L^p}^{1-\zeta}.$$

If $|B_*| \le cn^{-1}$, inequality (A.4) with $r = \infty$ yields

$$\begin{aligned} \|f\|_{L^\infty} & \le C_6 \Big(|B_*|^{1/2} \min_{x \in B_*} |f(x)| \Big)^{\zeta_*} \|D^k f\|_{L^p}^{1-\zeta_*} \\ & \le C_6 \Big(|B_*|^{1/2} \|f\|_{B_*,L^\infty} \Big)^{\zeta_*} \|D^k f\|_{L^p}^{1-\zeta_*}, \end{aligned}$$

where $\zeta_* = \frac{k/d-1/p}{1/2+k/d-1/p}$. It follows that

$$\|f\|_{L^\infty} \le C_6^{1/(1-\zeta_*)} |B_*|^{\frac{\zeta_*}{2(1-\zeta_*)}} \|D^k f\|_{L^p},$$

which together with (A.4) implies that

$$\begin{aligned} \|f\|_{L^r} &\le C_6^{1+\frac{\zeta}{1-\zeta_*}} |B_*|^{\frac{\zeta}{2(1-\zeta_*)}} \|D^k f\|_{L^p} \\ \Longrightarrow \|f\|_{L^r} &\le C_9 n^{-(\frac{k}{d}-\frac{1}{p}+\frac{1}{r})} \|D^k f\|_{L^p}. \end{aligned}$$

Therefore, we obtain that for $r \in [\frac{2k+d}{d}p, \infty]$

$$\|f\|_{L^r} \le C_{10} \max\Big\{ n^{-\vartheta_0} \|S_n f\|_{\mathcal{B}}^{2\vartheta_0} \|D^k f\|_{L^p}^{1-2\vartheta_0}, n^{-\vartheta_0'} \|D^k f\|_{L^p} \Big\}, \tag{A.5}$$

with $\vartheta_0 = \vartheta_0(k,d,p,r)$ given in (2.10), and $\vartheta_0' = \vartheta_0'(k,d,p,r)$ in (2.11).

Since $[0,1]^d$ is not regular, it holds that

$$\|f\|_{L^\infty} < 4C_1 d^{\frac{k}{2}-\frac{d}{2p}} \|D^k f\|_{L^p}.$$

Assume $q \ge \frac{2k+d}{2l+d}p$ and choose $r = \frac{2k+d}{d}p$. Then it follows from (A.5) and Proposition A.1.5 with $\lambda^{-1} = 4C_1 d^{\frac{k}{2}-\frac{d}{2p}}$ that

$$\|D^l f\|_{L^q} \le C_{11} \max\Big\{ n^{-\vartheta_l} \|S_n f\|_{\mathcal{B}}^{2\vartheta_l} \|D^k f\|_{L^p}^{1-2\vartheta_l}, n^{-\vartheta_l'} \|D^k f\|_{L^p} \Big\}, \tag{A.6}$$

with $\vartheta_l = \vartheta_l(k,d,p,q)$ and $\vartheta_l' = \vartheta_l'(k,d,p,q)$.

Since the values $\vartheta_l(k,d,p,q)$, $\vartheta_l'(k,d,p,q)$ for $q < \frac{2k+d}{2l+d}p$ are the same for $q = \frac{2k+d}{2l+d}p$, inequality (A.6) is indeed valid for all $q \in [1,\infty]$.

Case II. The cube $[0,1]^d$ is regular.

Note that (A.2) with $B = [0,1]^d$ implies $\|f - f_B\|_{L^\infty} \le \frac{1}{4}\|f\|_{L^\infty}$. Same as the argument for (A.3), there is a cube B_o such that

$$|B_o| \ge C_{12} \qquad \text{and } |f(x)| \ge \frac{1}{4}\|f\|_{L^\infty} \quad \text{ for } x \in B_o.$$

Since $\mathcal{B}$ is a normal system with c, there is $B_{oo} \in \mathcal{B}$ such that $B_{oo} \subset B_o$ and $|B_{oo}| \ge c^{-1}|B_o|$. Let $n_0 := \lfloor c/C_{12} \rfloor + 1$. If $n \ge n_0 (\Longrightarrow \#B_{oo} \cap \Gamma_n \ge 1)$, we get

$$\|S_n f\|_{\mathcal{B}} \ge C_{13}^{-1} n^{1/2} \|f\|_{L^\infty}.$$

The regularity of $[0,1]^d$ implies that

$$\|f\|_{L^\infty} \ge 4C_1 d^{\frac{k}{2}-\frac{d}{2p}} \|D^k f\|_{L^p}.$$

By Proposition A.1.5 with $\lambda^{-1} = 4C_1 d^{\frac{k}{2}-\frac{d}{2p}}$, we have

$$\|D^l f\|_{L^q} \le C_{14}\|f\|_{L^\infty} \le C_{14}C_{13}n^{-1/2}\|S_n f\|_{\mathcal{B}}. \tag{A.7}$$

Combining (A.6) and (A.7), we complete the proof. □

Remark A.1.7. The proof above follows more or less the idea from (Nemirovski, 1985), but with sharpened tools: one is to use averaged Taylor polynomials (Lemma A.1.2) instead of Taylor polynomials; the other is to select Besicovitch cover rather than Vitali cover.

A.2. General convergence analysis

This section gives proofs of convergence results under approximate source conditions.

A.2.1. Good noise case

The convergence rate is derived provided that the noise ξ_n is good, namely, $\|\xi_n\|_{\mathcal{B}} \le \gamma_n$.

Proof (of Theorem 2.3.3). The assumption $\|\xi_n\|_{\mathcal{B}} = \|S_n f - y_n\|_{\mathcal{B}} \le \gamma_n$ implies that f is admissible for the minimization problem (2.4), which in turn implies that

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

As a consequence, we obtain the estimate

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

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

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

Since

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

we obtain the inequality

$$\begin{aligned}\|D^k\hat{f}_{\gamma_n} - D^k f\|_{L^2} &\le d_n(t) + \sqrt{d_n(t)^2 + 2t\|S_n\hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}}} \\ &\le 2d_n(t) + (2t)^{1/2}\|S_n\hat{f}_{\gamma_n} - S_n f\|_{\mathcal{B}}^{1/2} \\ &\le 2d_n(t) + 2(\gamma_n t)^{1/2}\end{aligned} \tag{A.8}$$

for every $t \ge 0$. We now recall the interpolation inequality in Theorem 2.2.7 with $p = 2$

$$\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^q} \le C\max\Bigg\{\frac{\|S_n(\hat{f}_{\gamma_n} - f)\|_{\mathcal{B}}^{2\vartheta_l}}{n^{\vartheta_l}}\|D^k(\hat{f}_{\gamma_n} - f)\|_{L^2}^{1-2\vartheta_l}, \\ \frac{\|S_n(\hat{f}_{\gamma_n} - f)\|_{\mathcal{B}}}{n^{1/2}}, \frac{\|D^k(\hat{f}_{\gamma_n} - f)\|_{L^2}}{n^{\vartheta_l'}}\Bigg\} \tag{A.9}$$

for n sufficiently large, with some $C > 0$.

If the maximum in (A.9) is attained at the first term, the estimate (A.8) implies that

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

On the other hand, if the maximum in (A.9) is attained at the second term, we have

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

Finally, if the maximum in (A.9) is attained at the third term, we have

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

The assertion of Theorem 2.3.3 follows from combining the above three cases. □

A.2.2. Estimate of L^q-risk

The crucial part of the proof below is to show that the loss of MIND asymptotically vanishes fast enough when the noise is not good, i.e. $\|\xi\|_{\mathcal{B}} > \gamma_n$.
Proof (of Theorem 2.3.5). Denote in the following

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

Using Theorem 2.3.3, we see that we can estimate, for n sufficiently large,

$$\mathbb{E}\left[\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^q}\right] \leq \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \leq \gamma_n\right\} C n^{-\mu(1-2\vartheta_l)-\vartheta_l} (\log n)^{2r\vartheta_l} + \int_{\gamma_n}^{\infty} \sup\left\{\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^q}; \|\xi_n\|_{\mathcal{B}} = t\right\} dp(t). \quad \text{(A.10)}$$

In the following, we will show that the second term on the right hand side of (A.10) 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

$$\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^q} \leq \|D^l \hat{f}_{\gamma_n}\|_{L^q} + \|D^l f\|_{L^q} \leq C\|D^k \hat{f}_{\gamma_n}\|_{L^2} + \|D^l f\|_{L^\infty} \quad \text{(A.11)}$$

for some constant C depending only on 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

$$h_{n,x}(z) = h(n^{1/d}(z - x)) \qquad \text{for } x - 1/2 \leq z \leq x + 1/2.$$

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

$$\begin{aligned}
\|D^k g\|_{L^2} &= \sum_{x \in \Gamma_n} |f(x) + \xi_n(x)| \|D^k h_{n,x}\|_{L^2} = \sum_{x \in \Gamma_n} |f(x) + \xi_n(x)| n^{\frac{2k-d}{2d}} \|D^k h\|_{L^2} \\
&= n^{\frac{2k-d}{2d}} \|S_n f + \xi_n\|_1 \|D^k h\|_{L^2} \leq n^{\frac{2k+d}{2d}} \left(\|f\|_{L^\infty} + \|\xi_n\|_\infty\right) \|D^k h\|_{L^2} \\
&\leq C n^{\frac{2k+d}{2d}} \left(\|D^l f\|_{L^\infty} + \|\xi_n\|_\infty\right) \|D^k h\|_{L^2} \qquad \text{[by the same argument as in (A.11)]}.
\end{aligned}$$

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

$$\sup\left\{\|D^l\hat{f}_{\gamma_n} - D^l f\|_{L^q}; \|\xi_n\|_{\mathcal{B}} = t\right\} \leq Cn^{\frac{2k+d}{2d}}(\|D^l f\|_{L^\infty} + t).$$

As a consequence, we can estimate the last term in (A.10) by

$$\begin{aligned}\int_{\gamma_n}^{\infty} \sup\left\{\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^\infty}; \|\xi_n\|_{\mathcal{B}} = t\right\} dp(t) &\leq \int_{\gamma_n}^{\infty} Cn^{\frac{2k+d}{2d}}(\|D^l f\|_{L^\infty} + t)\, dp(t) \\ &= Cn^{\frac{2k+d}{2d}}(\|D^l f\|_{L^\infty} + \gamma_n)(1 - p(\gamma_n)) - Cn^{\frac{2k+d}{2d}} \int_{\gamma_n}^{\infty} (p(t) - 1)\, dt.\end{aligned}$$

From Proposition 2.2.4 we obtain that

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

for sufficiently large n. Thus we see that

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

for sufficiently large n. Now the choice of γ_n in (2.5) implies that

$$n^{\frac{2k+5d}{2d}} \gamma_n e^{-\frac{\gamma_n^2}{2\sigma^2}} = \mathcal{O}(n^{-\frac{1}{2}-\varepsilon})$$

as $n \to \infty$ for some $\varepsilon > 0$. This shows that the second term in (A.10) tends to zero faster as $n \to \infty$, which concludes the proof of Theorem 2.3.5. □

A.2.3. Removal of zero mean requirement

Proof (of Proposition 2.3.6). (i) It is clear that for every $f \in H^k(\mathbb{T}^d)$, function $f^0 := f - \int_{\mathbb{T}^d} f(z)dz \in H_0^k(\mathbb{T}^d)$, and $D^k f^0 = D^k f$. Let $m := n^{1/d}$, and note that for each $x \in \Gamma_n$

$$\begin{aligned}\left|\int_{x+[-\frac{1}{2m},\frac{1}{2m})^d} f(z)dz - \frac{1}{n}f(x)\right| &= \left|\int_{x+[-\frac{1}{2m},\frac{1}{2m})^d} \int_0^1 \langle (Df)(x + t(z - x)), z - x\rangle dt dz\right| \\ &\leq \int_0^1 \int_{x+[-\frac{1}{2m},\frac{1}{2m})^d} |\langle (Df)((1-t)x + tz)), z - x\rangle| dz dt\end{aligned}$$

$$\leq \frac{\sqrt{d}}{2m}\frac{1}{n}\|f\|_{\mathcal{C}^1}.$$

It follows that

$$\begin{aligned}\Big|\int_{\mathbb{T}^d} f(z)dz - \frac{1}{n}\sum_{x\in\Gamma_n} f(x)\Big| &\leq \sum_{x\in\Gamma_n}\Big|\int_{x+[-\frac{1}{2m},\frac{1}{2m})^d} f(z)dz - \frac{1}{n}f(x)\Big| \\ &\leq \frac{\sqrt{d}}{2m}\sum_{x\in\Gamma_n}\frac{1}{n}\|f\|_{\mathcal{C}^1} = \frac{\sqrt{d}}{2m}\|f\|_{\mathcal{C}^1}.\end{aligned}$$

Then

$$\begin{aligned}\|S_n f^0 - (y_n - \bar{y}_n)\|_{\mathcal{B}} \leq& \|S_n f - y_n\|_{\mathcal{B}} + \|\int_{\mathbb{T}^d} f(z)dz - \frac{1}{n}\sum_{x\in\Gamma_n} f(x)\|_{\mathcal{B}} + \|\frac{1}{n}\sum_{x\in\Gamma_n} f(x) - \bar{y}_n\|_{\mathcal{B}} \\ =& \|\xi_n\|_{\mathcal{B}} + \sqrt{n}\Big|\int_{\mathbb{T}^d} f(z)dz - \frac{1}{n}\sum_{x\in\Gamma_n} f(x)\Big| + \sqrt{n}|\bar{\xi}_n| \\ \leq& 2\|\xi_n\|_{\mathcal{B}} + \mathcal{O}(n^{1/2-1/d}) = \mathcal{O}(\gamma_n). \qquad \text{(A.12)}\end{aligned}$$

We now apply Theorem 2.3.3 to $(\hat{f}^0_{\gamma_n} - f^0)$ and obtain that

$$\|\hat{f}^0_{\gamma_n} - f^0\|_{L^q} = \mathcal{O}\Big(\max\Big\{\frac{\gamma_n^{2\vartheta_l} c_n^{1-2\vartheta_l}}{n^{\vartheta_l}}, \frac{\gamma_n}{n^{1/2}}, \frac{c_n}{n^{\vartheta_l'}}\Big\}\Big)$$

It further implies

$$\begin{aligned}\|\hat{f}_{\gamma_n} - f\|_{L^q} \leq& \|\hat{f}^0_{\gamma_n} - f^0\|_{L^q} + \|\int_{\mathbb{T}^d} f(z)dz - \frac{1}{n}\sum_{x\in\Gamma_n} f(x)\|_{L^q} + \|\frac{1}{n}\sum_{x\in\Gamma_n} f(x) - \bar{y}_n\|_{L^q} \\ =& \|\hat{f}^0_{\gamma_n} - f^0\|_{L^q} + |\int_{\mathbb{T}^d} f(z)dz - \frac{1}{n}\sum_{x\in\Gamma_n} f(x)| + |\bar{\xi}_n| \\ \leq& \|\hat{f}^0_{\gamma_n} - f^0\|_{L^q} + \mathcal{O}(n^{-1/d}) + n^{-1/2}\|\xi_n\|_{\mathcal{B}} \qquad \text{(A.13)} \\ =& \mathcal{O}\Big(\max\Big\{\frac{\gamma_n^{2\vartheta_l} c_n^{1-2\vartheta_l}}{n^{\vartheta_l}}, \frac{\gamma_n}{n^{1/2}}, \frac{c_n}{n^{\vartheta_l'}}\Big\}\Big).\end{aligned}$$

(ii) Based on the estimates (A.12) and (A.13), the assertion follows from Corollary 2.3.4 and Theorem 2.3.5 in a similar way as (i). □

A.3. Results in one dimension

We now give the proofs of Proposition 2.4.1 and Theorem 2.4.4. 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.

A.3.1. Approximation properties of splines

The following result is a generalization of a known result for splines on $\mathbb{R}$ (Scherer and Shadrin, 1999) to periodic splines.

Proposition A.3.1 (Condition number of B-splines). *Assume that*

$$\{Q_i^m(x), i = 0, \dots, n-1\}$$

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

$$\|(c)_{i=0}^{n-1}\|_p \le m2^m n^{1/p} \|\sum_{i=0}^{n-1} c_i Q_i^m\|_{L^p} \qquad \text{for } 1 \le p \le \infty. \tag{A.14}$$

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

$$\tilde{c}_i := c_{i \bmod n} \qquad \text{for } i = -m+1, \dots, ln-1.$$

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 that

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

or

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

By letting $l \to \infty$, we obtain (A.14) for $1 \le p < \infty$.

The case $p = \infty$ follows by taking $p \to \infty$. □

Proposition A.3.2 (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*

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

Proof. By the fact that $(L^1(\mathbb{T}), L^\infty(\mathbb{T}))_{1-1/p,\,p} = L^p(\mathbb{T})$ for $1 < p < \infty$ (cf. Freitag, 1978) and Proposition 2.1.2, 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 A.3.3. 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 A.3.4 (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 $u \in L_0^1(\mathbb{T})$*

$$\|u - Au\|_{W_0^{r,q}} \le C_1 \frac{\|u\|_{B_{p,0}^{s,p'}}}{n^{s-r-(1/p-1/q)_+}} \qquad \text{with } 1 \le s \le m,\ 0 \le r \le \lfloor s-1 \rfloor,$$
$$\|Au\|_{W_0^{r,q}} \le C_2 \frac{\|u\|_{B_{p,0}^{s,p'}}}{n^{s-r-(1/p-1/q)_+}} \qquad \text{with } 1 \le s \le \lceil s \rceil \le r \le m-1,$$

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

Remark A.3.5. 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 Proposition 2.1.2.

Proposition A.3.6 (Finite differences and $W^{1,p}(\mathbb{T})$). *Let $h > 0$ and $1 \le p \le \infty$. Then*

$$\|D_{h,+} f\|_{L^p} = \|D_{h,-} f\|_{L^p} \le h \|Df\|_{L^p} \qquad \text{for } f \in W^{1,p}(\mathbb{T}). \tag{A.15}$$

Proof. The case of $p = \infty$ is obviously true. Now consider $1 \le p < \infty$. Since $\|D_{h,+} f\|_{L^p} = \|D_{h,-} f\|_{L^p}$, it is sufficient to prove (A.15) 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

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

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

A.3.2. Regular systems of intervals

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 (cf. Definition 2.2.3). These piecewise constant vectors will appear as spline coefficients for certain approximation splines needed for the proof of Proposition 2.4.1.

Lemma A.3.7. *Assume that $m \in \mathbb{N}$, $m \geq 2$, and that $n \in \mathbb{N}$ is written as*

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

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

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

From $d_{r+1} \leq m-1$ and $m - 1 + 2(\sqrt{m}+1) = (\sqrt{m}+1)^2$, it follows that the claim holds for $r+1$, which concludes the proof. □

Lemma A.3.8. *Assume that the family $\mathcal{B}$ is m-regular for some fixed $m \geq 2$. 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*

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

Proof. Let $r = \lceil \log_m n \rceil$. Let $\ell_- \in \mathbb{N}$ be maximal such that $\ell_- m^{-r} \leq i_0/n$, 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

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

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

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

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 $i \in \Gamma_n$ and therefore

$$\|c\|_{\mathcal{B}^*} \le \sum_{B \in \mathcal{B}} |c_B| \sqrt{\#\Gamma_n \cap B} \le \sum_{B \in \hat{\mathcal{B}}} \sqrt{\#\Gamma_n \cap B}.$$

Now note that

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

Therefore Lemma A.3.7 and the definition of $\hat{\mathcal{B}}$ imply that

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

□

Remark A.3.9. Note that the estimate in the previous lemma can be improved to $\|c\|_{\mathcal{B}^*} \leq (\sqrt{m}+1)\sqrt{2p}$ 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 $\|c\|_{\mathcal{B}^*} \leq \sqrt{p}$ in case the family $\mathcal{B}$ contains all intervals.

A.3.3. Estimate of multiscale distance functions

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. As illustrated in Figure A.1, 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 A.3.10). 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 2.4.1.

Lemma A.3.10. *Let $1 \leq q \leq \infty$, $k \in \mathbb{N}$, $\beta \in \mathbb{N}_0$, and $k \geq \beta \geq 0$. Let also $\Gamma \subset \mathbb{T}$ be a finite set such that*

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

Denote

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

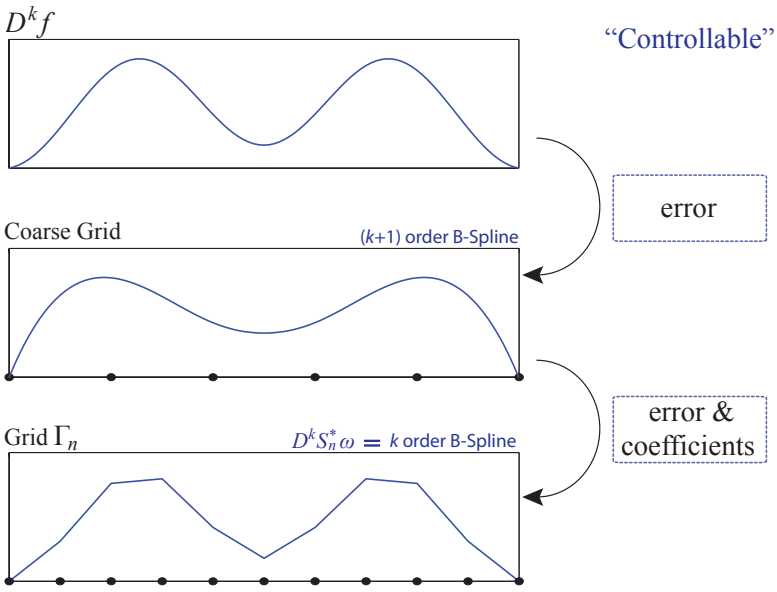

Figure A.1.: Idea for the proof of Proposition 2.4.1.

and assume that

$$g \in \mathcal{S}_{k+\beta+1}(\Gamma;\mathbb{T}) \qquad \textit{with} \int_{\mathbb{T}} g\,dx = 0,$$

and that $\mathcal{B}$ is m-regular. Then there exists $c \in \mathbb{R}^{\Gamma_n}$ such that

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

where constants C_1, $C_2 > 0$ only depend on k, q, and m.

Proof. Let h be the best approximation of g in $\operatorname{span}\{\psi_{i,n}^{k+\beta} : i = 0,\dots,n-1\}$ in the L^2 sense, see (2.22). Then we can write

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

for some coefficients $\tilde{c}_i \in \mathbb{R}$. Because the functions $\psi_{i,n}^{k+\beta}$ 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

$$h = \sum_{i=0}^{n-1} \tilde{c}_i \psi_{i,n}^{k+\beta} = \sum_{i=0}^{n-1} \tilde{c}_i Q_i^{k+\beta}. \tag{A.16}$$

Now note that the fact $\int_{\mathbb{T}} g\,dx = 0$ implies that h is at the same time the best approximation of g in $\mathcal{S}_{k+\beta}(\Gamma_n;\mathbb{T})$. Thus (A.16) shows that, actually, the coefficients $\tilde{c}_i$ are the coefficients of the $(k+\beta)$-th order spline that approximates g best in the L^2-sense. Thus it follows from Proposition A.3.4 that

$$\|h-g\|_{L^2} \le C_1 \frac{\|g^{(k+\beta)}\|_{L^2}}{n^{k+\beta}}.$$

Now let

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

which implies that

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

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

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

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

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

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

for all r. From this, we obtain that

$$\langle h, D_{\frac{1}{n},+}^r Q_j^{k+\beta}\rangle_{L^2} = \langle g, D_{\frac{1}{n},+}^r Q_j^{k+\beta}\rangle_{L^2}$$

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

$$\langle D_{\frac{1}{n},-}^r h, Q_j^{k+\beta}\rangle_{L^2} = \langle D_{\frac{1}{n},-}^r g, Q_j^{k+\beta}\rangle_{L^2} \tag{A.17}$$

for all j and all r. Next we note that

$$D_{\frac{1}{n},-}^{k+\beta+1} h = \sum_{i=0}^{n-1} (D_-^{k+\beta+1}\tilde{c})_i Q_i^{k+\beta} = (-1)^k n^{1-k-\beta} \sum_{i=0}^{n-1} (D_- c)_i Q_i^{k+\beta}. \tag{A.18}$$

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

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

As a consequence, we obtain from (A.17) with $r = k+\beta+1$ and (A.18) that

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

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

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

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

Now denote by $I \subset \Gamma_n$ the set of all points i/n for which $i/n \notin \Gamma + (-(k+\beta)/n, (k+\beta+1)/n)$. 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 A.3.8 implies that

$$\|c\|_{\mathcal{B}^*} \leq \sum_{j=1}^{\#\Gamma} C|\omega_j| \sqrt{\#I_j \cap \Gamma_n} + \sum_{i/n \notin I} |c_i| \tag{A.19}$$

for some constant $C > 0$ only depending on m. Now define

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

Then the right hand side term in (A.19) can also be written as a sum over all products $t_i|c_i|$, $i = 0, \dots, n-1$. Therefore

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

Applying Hölder's inequality gives

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

$$\leq \|c\|_q \Big(2(k+\beta)\#\Gamma + C^{q_*} \sum_{j=1}^{\#\Gamma} (\#I_j \cap \Gamma_n)^{1-q_*/2} \Big)^{1/q_*}$$

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

$$\begin{aligned} \|c\|_{\mathcal{B}^*} &\leq \|c\|_q \Big(2(k+\beta)\#\Gamma + C^{q_*} \#\Gamma (n\tau_{\max})^{(1-q_*/2)_+} \Big)^{1/q_*} \\ &\leq C\|c\|_q (\#\Gamma)^{1/q_*} (n\tau_{\max})^{(1/q_*-1/2)_+} = C\|c\|_q (\#\Gamma)^{1-1/q} (n\tau_{\max})^{(1/2-1/q)_+}. \quad \text{(A.20)} \end{aligned}$$

Now note that (A.17) implies that $D^{k+\beta}_{1/n,-}h$ is the best approximating spline in the L^2-sense of $D^{k+\beta}_{1/n,-}g$. Thus the definition of c and Propositions A.3.1, A.3.2 and A.3.6 imply that

$$\begin{aligned} \|c\|_q = n^{k+\beta-1} \|D^{k+\beta}_- \tilde{c}\|_q &\leq C n^{k+\beta-1+1/q} \|D^{k+\beta}_{1/n,-} h\|_{L^q} \\ &\leq C n^{k+\beta-1+1/q} \|D^{k+\beta}_{1/n,-} g\|_{L^q} \leq C n^{-1+1/q} \|g^{(k+\beta)}\|_{L^q}. \end{aligned}$$

Together with (A.20) this shows that

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

for some constant $C > 0$. □

Proof (of Proposition 2.4.1). Assume first that $p \geq 2$. Proposition A.3.4 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

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

Next we obtain from Lemma A.3.10 the existence of a vector $c \in \mathbb{R}^{\Gamma_n}$ such that

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

provided that λ is sufficiently large (here we use that, in the notation of the lemma, $\beta = 0$, $\#\Gamma = \lambda$ and $\tau_{\max} = 1/\lambda$). Combining these estimates, it follows that, for

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

we have

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

Choosing

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

we obtain that

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

as $n \to \infty$.

Now let $p \le 2$. Again applying Proposition A.3.4 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

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

and we obtain, for λ sufficiently large, from Lemma A.3.10 the existence of $c \in \mathbb{R}^{\Gamma_n}$ with

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

This shows that, for

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

we have

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

Choosing

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

we obtain that

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

which proves the assertion.

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

A.3.4. Over-smoothing

We need a simple upper bound on the multiresolution norm, which has a similar flavor as Proposition 2.2.6.

Lemma A.3.11. *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

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

□

Proof (of Theorem 2.4.4). By C we denote a generic constant whose value may be different from place to place. 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 A.3.4, and $g_\lambda(t) := G'_\lambda(t)$. It follows that $g_\lambda \in H_0^k(\mathbb{T})$, and that

$$\|D^l(f - g_\lambda)\|_{L^q} = \|D^{l+1}(F - G_\lambda)\|_{L^q} \leq C \frac{\|F\|_{W^{s+1,\infty}}}{\lambda^{s-l}}$$

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

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

The second relation implies that

$$\tilde{d}_n(t) := \min_{\|w\|_{\mathcal{B}^*}\leq t}\|D^k S_n^* w - D^k g_\lambda\|_{L^2} \leq \tilde{d}_n(0) = \|D^k g_\lambda\|_{L^2}$$
$$\leq C\left(\frac{n}{\log n}\right)^{\frac{k-s}{2s+1}}(\log n)^{(k-s)\tilde{\epsilon}}\|f\|_{W^{s,\infty}} \quad \text{(A.21)}$$

for every $t \geq 0$. By Proposition A.3.4 and Lemma A.3.11, we have

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

for sufficiently large n. Consequently

$$\|S_n g_\lambda - y_n\|_{\mathcal{B}} \leq \|S_n g_\lambda - S_n f\|_{\mathcal{B}} + \|S_n f - y_n\|_{\mathcal{B}}$$
$$\leq C(\log n)^{\frac{1}{2}-(s+\frac{1}{2})\tilde{\epsilon}}\|f\|_{W^{s,\infty}} + \|\xi_n\|_{\mathcal{B}}. \quad \text{(A.22)}$$

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 (A.22), implies that $\|S_n g_\lambda - y_n\|_{\mathcal{B}} \leq \gamma_n$ for large enough n. In such case we can apply Theorem 2.3.3, but with f replaced by its approximation g_λ, and obtain the estimate

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

with $\vartheta = (k-l)/(2k+1)$ and $\vartheta' = 2k(k-l)/(2k+1)$. By (A.21), this further implies that, for sufficiently large n, the estimate

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

with $\epsilon = \frac{(2l+1)(k-s)\tilde{\epsilon}+(2r-1)(k-l)}{2k+1} > \frac{(2r-1)(k-l)}{2k+1}$. Note that $\lim_{n\to\infty} \mathbb{P}\{\|\xi_n\|_{\mathcal{B}} > \tilde{\gamma}_n\} = 0$ by Proposition 2.2.4. Thus we obtain that

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

almost surely as $n \to \infty$.

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

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

As in the proof of Theorem 2.3.5, we see that

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

which tends to zero faster than $(\log n)^{(s-l)/(2s+1)+\epsilon} n^{-(s-l)/(2s+1)}$ 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. □

A.4. Results for penMIND

In what follows we collect all the missing proofs of the results about the penMIND estimator, including Theorems 2.5.1, 2.5.3, Corollary 2.5.5, and Proposition 2.5.8.
Proof (of Theorem 2.5.1). The fact that $\hat{f}_\alpha$ solves (2.27) implies that

$$\|S_n\hat{f}_\alpha - y_n\|_{\mathcal{B}} + \frac{\alpha}{2}\|D^k\hat{f}_\alpha\|_{L^2}^2 \leq \|S_n f - y_n\|_{\mathcal{B}} + \frac{\alpha}{2}\|D^k f\|_{L^2}^2.$$

Consequently,

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

From this we further obtain the estimate

$$\begin{aligned}
\|D^k\hat{f}_\alpha - D^k f\|_{L^2} &\leq d_n\Big(\frac{1}{2\alpha}\Big) + \sqrt{d_n\Big(\frac{1}{2\alpha}\Big)^2 + \frac{4}{\alpha}\|\xi_n\|_{\mathcal{B}} - \frac{1}{\alpha}\|S_n\hat{f}_\alpha - S_n f\|_{\mathcal{B}}} \\
&\leq d_n\Big(\frac{1}{2\alpha}\Big) + \sqrt{2d_n\Big(\frac{1}{2\alpha}\Big)^2 + \frac{8}{\alpha}\|\xi_n\|_{\mathcal{B}}} - \frac{1}{\sqrt{\alpha}}\|S_n\hat{f}_\alpha - S_n f\|_{\mathcal{B}}^{1/2} \qquad \text{(A.23)} \\
&\leq (1+\sqrt{2})d_n\Big(\frac{1}{2\alpha}\Big) + \frac{2\sqrt{2}}{\sqrt{\alpha}}\|\xi_n\|_{\mathcal{B}}^{1/2} - \frac{1}{\sqrt{\alpha}}\|S_n\hat{f}_\alpha - S_n f\|_{\mathcal{B}}^{1/2}.
\end{aligned}$$

We now recall the interpolation inequality from Theorem 2.2.7

$$\|D^l w\|_{L^q} \leq C_1 \max\Big\{\frac{\|S_n w\|_{\mathcal{B}}^{2\vartheta_l}}{n^{\vartheta_l}}\|D^k w\|_{L^2}^{1-2\vartheta_l};\ \frac{\|S_n w\|_{\mathcal{B}}}{n^{1/2}};\ \frac{\|D^k w\|_{L^2}}{n^{\vartheta_l'}}\Big\}, \quad \text{for } n \geq n_0. \qquad \text{(A.24)}$$

We apply this inequality to $w = \hat{f}_\alpha - f$, and treat each term in the maximum separately. For the first term in the r.h.s. of (A.24), it follows from Young's inequality and (A.23) that

$$\frac{\|S_n\hat{f}_\alpha - S_n f\|_{\mathcal{B}}^{2\vartheta_l}}{n^{\vartheta_l}}\|D^k\hat{f}_\alpha - D^k f\|_{L^2}^{1-2\vartheta_l}$$

$$
\begin{aligned}
&= \frac{\alpha^{2\vartheta_l}}{n^{\vartheta_l}} \left(\Big(\frac{1}{\sqrt{\alpha}} \|S_n \hat{f}_\alpha - S_n f\|_{\mathcal{B}}^{1/2} \Big)^{\frac{4\vartheta_l}{1+2\vartheta_l}} \|D^k \hat{f}_\alpha - D^k f\|_{L^2}^{\frac{1-2\vartheta_l}{1+2\vartheta_l}} \right)^{1+2\vartheta_l} \\
&\leq \frac{\alpha^{2\vartheta_l}}{n^{\vartheta_l}} \left(\frac{4\vartheta_l}{1+2\vartheta_l} \frac{1}{\sqrt{\alpha}} \|S_n \hat{f}_\alpha - S_n f\|_{\mathcal{B}}^{1/2} + \frac{1-2\vartheta_l}{1+2\vartheta_l} \|D^k \hat{f}_\alpha - D^k f\|_{L^2} \right)^{1+2\vartheta_l} \\
&\leq \frac{\alpha^{2\vartheta_l}}{n^{\vartheta_l}} \left((1+\sqrt{2}) d_n\Big(\frac{1}{2\alpha}\Big) + \frac{2\sqrt{2}}{\sqrt{\alpha}} \|\xi_n\|_{\mathcal{B}}^{1/2} \right)^{1+2\vartheta_l} \\
&\leq C_2 \left(\frac{\alpha^{2\vartheta_l} d_n(\frac{1}{2\alpha})^{1+2\vartheta_l}}{n^{\vartheta_l}} + \frac{\|\xi_n\|_{\mathcal{B}}^{1/2+\vartheta_l}}{\alpha^{1/2-\vartheta_l} n^{\vartheta_l}} \right).
\end{aligned}
$$

For the second term in the r.h.s. of (A.24), we have by (A.23) that

$$
\begin{aligned}
\frac{\|S_n \hat{f}_\alpha - S_n f\|_{\mathcal{B}}}{n^{1/2}} &\leq n^{-1/2} \left((1+\sqrt{2}) \sqrt{\alpha} d_n\Big(\frac{1}{2\alpha}\Big) + 2\sqrt{2} \|\xi_n\|_{\mathcal{B}}^{1/2} \right)^2 \\
&\leq C_3 \left(\frac{\alpha d_n(\frac{1}{2\alpha})^2}{\sqrt{n}} + \frac{\|\xi_n\|_{\mathcal{B}}}{\sqrt{n}} \right).
\end{aligned}
$$

And for the last term in the r.h.s. of (A.24), we have again by (A.23) that

$$
\begin{aligned}
\frac{\|D^k \hat{f}_\alpha - D^k f\|_{L^2}}{n^{\vartheta_l'}} &\leq n^{-\vartheta_l'} \left((1+\sqrt{2}) d_n\Big(\frac{1}{2\alpha}\Big) + \frac{2\sqrt{2}}{\sqrt{\alpha}} \|\xi_n\|_{\mathcal{B}}^{1/2} \right) \\
&\leq C_4 \left(\frac{d_n(\frac{1}{2\alpha})}{n^{\vartheta_l'}} + \frac{(\|\xi_n\|_{\mathcal{B}})^{1/2}}{n^{\vartheta_l'} \sqrt{\alpha}} \right).
\end{aligned}
$$

Combining the three estimates above proves the assertion. □

Proof (of Theorem 2.5.3). We always assume that n is large enough, i.e. $n \geq \max\{n_0, n_1\}$. Note that for $i = 1, 2, 3$,

$$
\alpha \sup_{f \in \mathcal{C}_n} d_n(1/2\alpha)^2 \leq \sigma \sqrt{\log n} \iff \sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha) \leq \Psi_{i,n}(\alpha). \tag{A.25}
$$

Let us introduce the following notation

$$
\begin{aligned}
\alpha_{\mathrm{or}} &:= \arg\min_{\alpha} \max_{i \in \{1,2,3\}} \sup_{f \in \mathcal{C}_n} \big\{ \Phi_{i,n}(\alpha) + \Psi_{i,n}(\alpha) \big\}, \\
\alpha_* &:= \max\Big\{ \alpha \in \mathcal{A}_\kappa \colon \sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_\alpha - f)\|_{L^q} \leq 2C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha) \Big\}, \\
\text{and } \alpha_{**} &:= \max\big\{ \alpha \in \mathcal{A}_\kappa \colon \alpha \sup_{f \in \mathcal{C}_n} d_n(1/2\alpha)^2 \leq \sigma \sqrt{\log n} \big\}.
\end{aligned}
$$

If $\alpha_{\mathrm{or}} \leq \alpha_{**}$, then for $i = 1, 2, 3$,

$$
\sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha_{\mathrm{or}}) + \Psi_{i,n}(\alpha_{\mathrm{or}}) \geq \Psi_{i,n}(\alpha_{**}).
$$

If $\alpha_{**} < \alpha_{\mathrm{or}} < \alpha_{**}\kappa$, then for $i = 1, 2, 3$,

$$\sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha_{\mathrm{or}}) + \Psi_{i,n}(\alpha_{\mathrm{or}}) \geq \Psi_{i,n}(\alpha_{**}\kappa) \geq \kappa^{-1/2} \Psi_{i,n}(\alpha_{**}).$$

If $\alpha_{**}\kappa \leq \alpha_{\mathrm{or}}$, then for $i = 1, 2, 3$,

$$\sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha_{\mathrm{or}}) + \Psi_{i,n}(\alpha_{\mathrm{or}}) \geq \sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha_{**}\kappa) \geq \Psi_{i,n}(\alpha_{**}\kappa) \geq \kappa^{-1/2} \Psi_{i,n}(\alpha_{**}).$$

Combining all three cases, we obtain

$$\max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha_{**}) \leq \kappa^{1/2} \max_{i \in \{1,2,3\}} \sup_{f \in \mathcal{C}_n} \big\{ \Phi_{i,n}(\alpha_{\mathrm{or}}) + \Psi_{i,n}(\alpha_{\mathrm{or}}) \big\}. \tag{A.26}$$

From the definition of α_{**} and (A.25), it follows that

$$\sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha_{**}) \leq \Psi_{i,n}(\alpha_{**}) \implies \sup_{f \in \mathcal{C}_n} \Phi_{i,n}(\alpha_{**}) + \Psi_{i,n}(\alpha_{**}) \leq 2\Psi_{i,n}(\alpha_{**}).$$

Then we obtain by Theorem 2.5.1 that

$$\begin{aligned} \sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_{\alpha_{**}} - f)\|_{L^q} &\leq \sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_{\alpha_{**}} - f)\|_{L^q} \\ &\leq C_0 \sup_{f \in \mathcal{C}_n} \max_{i \in \{1,2,3\}} \big\{ \Phi_{i,n}(\alpha_{**}) + \Psi_{i,n}(\alpha_{**}) \big\} \leq 2C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha_{**}). \end{aligned}$$

It follows that $\alpha_{**} \leq \alpha_*$. This together with the definition of α_* implies that for for every $\tilde{\alpha} \leq \alpha \leq \alpha_{**}$,

$$\begin{aligned} \|D^l(\hat{f}_\alpha - \hat{f}_{\tilde{\alpha}})\|_{L^q} &\leq \sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_\alpha - f)\|_{L^q} + \sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_{\tilde{\alpha}} - f)\|_{L^q} \\ &\leq 2C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha) + 2C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\tilde{\alpha}) \\ &\leq 4C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\tilde{\alpha}). \end{aligned} \tag{A.27}$$

Note also that the condition (2.34) implies $\alpha_{**} < n$, so we obtain $\alpha_{**} \leq \alpha_L$.

Therefore,

$$\begin{aligned} &\sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_{\alpha_L} - f)\|_{L^q} \\ &\leq \|D^l(\hat{f}_{\alpha_L} - \hat{f}_{\alpha_{**}})\|_{L^q} + \sup_{f \in \mathcal{C}_n} \|D^l(\hat{f}_{\alpha_{**}} - f)\|_{L^q} \\ &\leq 4C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha_{**}) + 2C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha_{**}) \qquad \text{[by (A.27) and } \alpha_{**} \leq \alpha_*\text{]} \\ &\leq 6C_0 \max_{i \in \{1,2,3\}} \Psi_{i,n}(\alpha_{**}) \\ &\leq 6\kappa^{1/2} C_0 \max_{i \in \{1,2,3\}} \sup_{f \in \mathcal{C}_n} \big\{ \Psi_{i,n}(\alpha_{\mathrm{or}}) + \Psi_{i,n}(\alpha_{\mathrm{or}}) \big\} \qquad \text{[by (A.26)].} \end{aligned}$$

□

Proof (of Corollary 2.5.5). For the choice of θ in (2.29), we have by Proposition 2.2.4 that

$$\mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} > \theta\sigma\sqrt{\log n}\right\} \to 0 \qquad \text{as } n \to \infty,$$

This together with Theorem 2.5.3 implies that the assertion holds almost surely.

We next prove that the assertion holds in expectation. Let $p(t) := \mathbb{P}\{\|\xi_n\|_{\mathcal{B}} \le t\}$. It follows from Theorem 2.5.3 that

$$\begin{aligned}\sup_{f\in\mathcal{C}_n} \mathbb{E}\left[\|D^l(\hat{f}_\alpha - f)\|_{L^q}\right] \\ \le 6\sqrt{\kappa}C_0\mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \le \theta\sigma\sqrt{\log n}\right\} \min_{\alpha} \max_{i\in\{1,2,3\}} \sup_{f\in\mathcal{C}_n}\left\{\Phi_{i,n}(\alpha) + \Psi_{i,n}(\alpha)\right\} \\ + \sup_{f\in\mathcal{C}_n} \int_{\theta\sigma\sqrt{\log n}}^{\infty} \sup\left\{\|D^l(\hat{f}_\alpha - f)\|_{L^q}; \|\xi_n\|_{\mathcal{B}} = t\right\} dp(t). \quad \text{(A.28)}\end{aligned}$$

As in the proof of Theorem 2.3.5, we have for the second term in (A.28) that

$$\begin{aligned}\sup_{f\in\mathcal{C}_n} \int_{\theta\sigma\sqrt{\log n}} \sup\left\{\|D^l(\hat{f}_\alpha - f)\|_{L^q}; \|\xi_n\|_{\mathcal{B}} = t\right\} dp(t) \\ \le Cn^{\frac{2k+5d}{2d}}\Big(\sup_{f\in\mathcal{C}_n}\|f\|_{L^\infty} + \theta\sigma\sqrt{\log n}\Big)e^{-\frac{\theta^2\log n}{2}}.\end{aligned}$$

By the choice of θ in (2.29), the upper bound above tends to zero faster than $1/\sqrt{n}$ as $n \to \infty$. Note, on the other hand, that the first term in (A.28) goes to zero no faster than $\sqrt{\log n/n}$. This concludes the proof. □

Proof (of Proposition 2.5.8). By C we denote a generic constant whose value may change from place to place. Let us set

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

As in the proof of Theorem 2.4.4, there is $g_\lambda \in \mathcal{S}_{k+1}(\Gamma_\lambda; \mathbb{T}) \cap H_0^k(\mathbb{T})$ such that

$$\|D^l(f - g_\lambda)\|_{L^q} \le C\left(\frac{\log n}{n}\right)^{\frac{s-l}{2s+1}} \|f\|_{W^{s,\infty}}, \quad \text{(A.29a)}$$

$$\|D^k g_\lambda\|_{L^2} \le C\left(\frac{n}{\log n}\right)^{\frac{k-s}{2s+1}} \|f\|_{W^{s,\infty}}, \quad \text{(A.29b)}$$

$$\text{and} \quad \|S_n f - S_n g_\lambda\|_{\mathcal{B}} \le C\sqrt{\log n}\|f\|_{W^{s,\infty}}. \quad \text{(A.29c)}$$

We apply Theorem 2.5.1, but with f replaced by its approximation g_λ, and obtain that

$$\|D^l(\hat{f}_\alpha - g_\lambda)\|_{L^q} \le C \max\Bigg\{ \frac{\alpha^{2\vartheta}\tilde{d}_n(\frac{1}{2\alpha})^{1+2\vartheta}}{n^\vartheta} + \frac{\|\tilde{\xi}_n\|_{\mathcal{B}}^{1/2+\vartheta}}{\alpha^{1/2-\vartheta}n^\vartheta}; \\ \frac{\alpha\tilde{d}_n(\frac{1}{2\alpha})^2}{\sqrt{n}} + \frac{\|\tilde{\xi}_n\|_{\mathcal{B}}}{\sqrt{n}}; \frac{\tilde{d}_n(\frac{1}{2\alpha})}{n^{\vartheta'}} + \frac{(\|\tilde{\xi}_n\|_{\mathcal{B}})^{1/2}}{n^{\vartheta'}\sqrt{\alpha}} \Bigg\},$$

with $\vartheta = (k-l)/(2k+1)$ and $\vartheta' = 2k(k-l)/(2k+1)$. Here for every $t \ge 0$

$$\begin{aligned}\tilde{d}_n(t) :=& \min_{\|w\|_{\mathcal{B}^*}\le t} \|D^k(S_n^* w - g_\lambda)\|_{L^2} \\ \le& \tilde{d}_n(0) = \|D^k g_\lambda\|_{L^2} \le C\left(\frac{n}{\log n}\right)^{\frac{k-s}{2s+1}} \|f\|_{W^{s,\infty}} \qquad \text{[by (A.29b)]},\end{aligned}$$

and

$$\begin{aligned}\|\tilde{\xi}_n\|_{\mathcal{B}} :=& \|y_n - S_n g_\lambda\|_{\mathcal{B}} \\ \le& \|y_n - S_n f\|_{\mathcal{B}} + \|S_n f - S_n g_\lambda\|_{\mathcal{B}} \le \|\xi_n\|_{\mathcal{B}} + C\sqrt{\log n}\|f\|_{W^{s,\infty}} \qquad \text{[by (A.29c)]}.\end{aligned}$$

If $\|\xi_n\|_{\mathcal{B}} \le \theta\sigma\sqrt{\log n}$ with some $\theta > \sqrt{6+2k}$ and

$$\alpha \sim n^{-\frac{2(k-s)}{2s+1}} (\log n)^{\frac{2(k-s)}{2s+1}+\frac{1}{2}}$$

then it implies that, for sufficiently large n,

$$\begin{aligned}&\|D^l(\hat{f}_\alpha - g_\lambda)\|_{L^q} \\ \le& C \max\left\{ \frac{(\log n)^{\frac{s-l}{2s+1}}}{n^{\frac{s-l}{2s+1}}} \|f\|_{W^{s,\infty}}^{\frac{2l+1}{2k+1}}, \left(\frac{\log n}{n}\right)^{1/2}, \frac{(\log n)^{-\frac{k-s}{2s+1}}}{n^{\frac{s-l}{2s+1}}}\|f\|_{W^{s,\infty}} \right\} \\ \le& C\Big(\frac{\log n}{n}\Big)^{\frac{s-l}{2s+1}} \|f\|_{W^{s,\infty}}^{\frac{2l+1}{2k+1}}.\end{aligned}$$

Based on it and (A.29a), we obtain

$$\begin{aligned}\|D^l(\hat{f}_\alpha - f)\|_{L^q} &\le \|D^l(\hat{f}_\alpha - g_\lambda)\|_{L^q} + \|D^l(g_\lambda - f)\|_{L^q} \\ &\le C(\log n)^{\frac{s-l}{2s+1}} n^{-\frac{s-l}{2s+1}} \|f\|_{W^{s,\infty}}^{\frac{2l+1}{2k+1}} + C(\log n)^{\frac{s-l}{2s+1}} n^{-\frac{s-l}{2s+1}} \|f\|_{W^{s,\infty}} \\ &\le C(\log n)^{\frac{s-l}{2s+1}} n^{-\frac{s-l}{2s+1}} \max\{1, \|f\|_{W^{s,\infty}}\}.\end{aligned}$$

Note that $\lim_{n\to\infty} \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} > \theta\sqrt{\log n}\right\} = 0$ by Proposition 2.2.4, so the above error bound holds almost surely as $n \to \infty$.

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

$$\mathbb{E}\left[\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^q}\right] \le \mathbb{P}\left\{\|\xi_n\|_{\mathcal{B}} \le \theta\sigma\sqrt{\log n}\right\} C(\log n/n)^{\frac{s-l}{2s+1}} \max\{1, \|f\|_{W^{s,\infty}}\}$$

$$
\begin{aligned}
&\quad + \int_{\theta\sigma\sqrt{\log n}}^{\infty} \sup\left\{ \|D^l(\hat{f}_{\gamma_n} - f)\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t \right\} dp(t) \\
&\leq C(\log n/n)^{\frac{s}{2s+1}} \max\{1, \|f\|_{W^{s,\infty}}\} \\
&\quad + \int_{\theta\sigma\sqrt{\log n}}^{\infty} \sup\left\{ \|D^l(\hat{f}_{\gamma_n} - f)\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t \right\} dp(t).
\end{aligned}
$$

As in the proof of Theorem 2.3.5, we see that

$$
\int_{\theta\sigma\sqrt{\log n}}^{\infty} \sup\left\{ \|D^l(\hat{f}_{\gamma_n} - f)\|_{L^\infty} : \|\xi_n\|_{\mathcal{B}} = t \right\} dp(t) \leq C n^{\frac{2k+5d}{2d}} \sqrt{\log n} \exp\left(-\frac{\theta^2 \log n}{2} \right),
$$

which tends to zero faster than $(\log n/n)^{(s-l)/(2s+1)}$ as $n \to \infty$. Thus, we have

$$
\mathbb{E}\left[\|D^l(\hat{f}_{\gamma_n} - f)\|_{L^q} \right] = \mathcal{O}\left(\left(\log n/n\right)^{\frac{s-l}{2s+1}} \right).
$$

It is clear that the above argument also holds for $f \in B^{s,p'}_{\infty,0}(\mathbb{T})$ with $1 \leq p' \leq \infty$. □

B. Proofs of Chapter 3

B.1. Interpolation inequality

Proof (of Lemma 3.2.1). The proof essentially follows that of Theorem 2.2.7, but we replace the usage of the Gagliardo-Nirenberg interpolation inequality (see Lemma A.1.4) by that of the interpolation inequality between homogeneous Sobolev norms

$$\|f\|_{H_0^{\alpha(1-\theta)+\beta\theta}} \le \|f\|_{H_0^\alpha}^{1-\theta} \|f\|_{H_0^\beta}^{\theta} \qquad \text{for } \alpha, \beta \in \mathbb{R} \text{ and } \theta \in (0,1),$$

which follows from the Hölder's inequality, and the definition of such norms in (2.7).

The detail goes as follows. Let C be a generic constant, which depends at most on s, d, and $\mathcal{B}$. Let also $k := \lfloor s \rfloor$. Consider first the case that $[0,1]^d$ is not regular. It follows from

$$\|f\|_{H_0^k} \le \|f\|_{L^2}^{1-\frac{k}{s}} \|f\|_{H_0^s}^{\frac{k}{s}}$$

and (A.5) that

$$\|f\|_{L^2} \le C \max\Big\{ n^{-\frac{s}{2s+d}} \|S_n f\|_{\mathcal{B}}^{\frac{2s}{2s+d}} \|f\|_{H_0^s}^{\frac{d}{2s+d}},\ n^{-\frac{2k(s-r)}{d(2k+d)}} \|f\|_{H_0^s} \Big\}.$$

This together with

$$\|f\|_{H_0^r} \le \|f\|_{L^2}^{1-\frac{r}{s}} \|f\|_{H_0^s}^{\frac{r}{s}}$$

further implies that

$$\|f\|_{H_0^r} \le C \max\Big\{ n^{-\frac{s-r}{2s+d}} \|S_n f\|_{\mathcal{B}}^{\frac{2(s-r)}{2s+d}} \|f\|_{H_0^s}^{1-\frac{2(s-r)}{2s+d}},\ n^{-\frac{2k(s-r)}{d(2k+d)}} \|f\|_{H_0^s} \Big\}.$$

Next we consider the other case that $[0,1]^d$ is regular. As the argument for (A.7), we obtain that

$$\|f\|_{H_0^r} \le \|f\|_{L^2}^{1-\frac{r}{k}} \|f\|_{H_0^k}^{\frac{r}{k}} \le C\|f\|_{L^\infty} \le C n^{-\frac{1}{2}} \|S_n f\|_{\mathcal{B}}.$$

Combining the above two cases, we conclude the proof. □

B.2. General analysis

Proof (of Theorem 3.2.6). The almost sure convergence result comes from Lemma 3.2.5 and the fact from Proposition 2.2.4 that

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

for the choice of γ_n in (3.4).

We now consider the convergence rate in expectation. Let $h \in \mathcal{C}_0^\infty(\mathbb{T}^d)$ satisfy that $h(0) = 1$ and $\operatorname{supp} h \subset [-1/2, 1/2]^d$. For $n \in \mathbb{N}$, $x \in \Gamma_n$, we further introduce the functions $h_{n,x} : \mathbb{T}^d \to \mathbb{R}$ by

$$h_{n,x}(z) = h(n^{1/d}(z - x)) \qquad \text{for } x - 1/2 \le z \le x + 1/2.$$

Note that $\mathcal{C}_0^\infty(\mathbb{T}^d) \subset \operatorname{Ran}(T)$. Thus, for fixed $n \in \mathbb{N}$ and $x \in \Gamma_n$, there exists a function g such that

$$Tg = \sum_{x \in \Gamma_n} y_n(x) h_{n,x} = \sum_{x \in \Gamma_n} \big((Tf)(x) + \xi_n(x)\big) h_{n,x}.$$

Since $T_n g = y_n$, we have by construction of MIND that

$$\|\hat{f}_{\gamma_n}\|_{H_0^k} \le \|g\|_{H_0^k}. \tag{B.1}$$

Moreover, from the estimate (3.2) and the fact that the functions $h_{n,x}$, $x \in \Gamma_n$ have pairwise disjoint supports, we obtain that

$$\begin{aligned}
\|g\|_{H_0^k} \le& C\|Tg\|_{H_0^{k+\beta}} = C \sum_{x \in \Gamma_n} |(Tf)(x) + \xi_n(x)| \|h_{n,x}\|_{H_0^{k+\beta}} \\
\le& C n^{\frac{2k+2\beta-d}{2d}} \|T_n f + \xi_n\|_1 \|h\|_{H_0^{k+\beta}} \\
\le& C n^{\frac{2k+2\beta+d}{2d}} \big(\|Tf\|_{L^\infty} + \|\xi_n\|_\infty\big) \|h\|_{H_0^{k+\beta}} \\
\le& C n^{\frac{2k+2\beta+d}{2d}} \big(\|Tf\|_{H_0^{1+\beta}} + \|\xi_n\|_\infty\big) \|h\|_{H_0^{k+\beta}} \\
\le& C n^{\frac{2k+2\beta+d}{2d}} \big(\|f\|_{H_0^1} + \|\xi_n\|_{\mathcal{B}}\big) \|h\|_{H_0^{k+\beta}}.
\end{aligned} \tag{B.2}$$

The last second inequality above is due to the Sobolev embedding theorem (Adams and Fournier, 2003, Theorem 4.12) and the Poincaré inequality (Ziemer, 1989, Theorem 4.4.2). By the same argument, we can also derive that

$$\|\hat{f}_{\gamma_n} - f\|_{L^2} \le \|\hat{f}_{\gamma_n}\|_{L^2} + \|f\|_{L^2} \le C \left(\|\hat{f}_{\gamma_n}\|_{H_0^k} + \|f\|_{H_0^1}\right).$$

Together with (B.1) and (B.2), it further implies

$$\|\hat{f}_{\gamma_n} - f\|_{L^2} \le C n^{\frac{2k+2\beta+d}{2d}} \big(\|f\|_{H_0^1} + \|\xi_n\|_{\mathcal{B}}\big) \|h\|_{H_0^{k+\beta}} + C\|f\|_{H_0^1}$$

$$\leq C n^{\frac{2k+2\beta+d}{2d}} \left(\|f\|_{H_0^1} + \|\xi_n\|_{\mathcal{B}} \right) \|h\|_{H_0^{k+\beta}}. \tag{B.3}$$

By Lemma 3.2.5, we can estimate the L^2-risk of MIND, for n large enough,

$$\mathbb{E}\left[\|\hat{f}_{\gamma_n} - f\|_{L^2} \right] \leq \mathbb{P}\left\{ \|\xi_n\|_{\mathcal{B}} \leq \gamma_n \right\} C n^{-\mu(1-2\vartheta)-\vartheta} (\log n)^{2r\vartheta} + \int_{\gamma_n}^{\infty} \sup \left\{ \|\hat{f}_{\gamma_n} - f\|_{L^2}; \|\xi_n\| = t \right\} dp(t), \tag{B.4}$$

with $p(t) := \mathbb{P}\left\{ \|\xi_n\|_{\mathcal{B}} \leq t \right\}$. Using (B.3), we derive an upper bound for the second term in the right hand side of (B.4)

$$\begin{aligned} \int_{\gamma_n}^{\infty} \sup \left\{ \|\hat{f}_{\gamma_n} - f\|_{L^2}; \|\xi_n\| = t \right\} dp(t) &\leq C n^{\frac{2k+2\beta+d}{2d}} \int_{\gamma_n}^{\infty} \left(\|f\|_{H_0^1} + t \right) \|h\|_{H_0^{k+\beta}} dp(t) \\ &\leq C n^{\frac{2k+2\beta+5d}{2d}} \gamma_n e^{-\frac{\gamma_n^2}{2\sigma^2}} \qquad \text{[by Proposition 2.2.4].} \end{aligned}$$

The choice of γ_n in (3.4) implies such an estimate is of order $n^{-1/2-\varepsilon}$ for some $\epsilon > 0$, which goes to 0 than the first term in the right hand side of (B.4) as $n \to \infty$. It follows immediately that for n sufficiently large

$$\mathbb{E}\left[\|\hat{f}_{\gamma_n} - f\|_{L^2} \right] \leq C n^{-\mu(1-2\vartheta)-\vartheta} (\log n)^{2r\vartheta}. \qquad \square$$

B.3. Concrete rates for $d = 1$

Proof (of Theorem 3.3.1). Applying Proposition A.3.4 with $u = h^{(k-\beta)}$, $q = 2$ and $m = k + \beta + 1$ yields for every $\lambda \in \mathbb{N}$ the existence of a spline $g \in \mathcal{S}_{k+\beta+1}(\Gamma_\lambda; \mathbb{T})$ such that

$$\begin{aligned} \|g - h^{(k-\beta)}\| &\leq C \frac{\|h\|_{H_0^s}}{\lambda^{s-k+\beta}}, \\ \|g^{(k+\beta)}\| &\leq C \frac{\|h\|_{H_0^s}}{\lambda^{s-2k}}. \end{aligned}$$

From Lemma A.3.10 with $q = 2$, $\#\Gamma = \lambda$ and $\tau_{\max} = 1/\lambda$, it follows that there exists a vector $c \in \mathbb{R}^{\Gamma_n}$ such that

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

if n is sufficiently large. Combining these estimates, we have, for

$$t \geq C\lambda^{1/2-s+2k} n^{-1/2} \|h\|_{H_0^s} \geq C\lambda^{1/2} n^{-1/2} \|g^{(k+\beta)}\|_{L^2} \geq \|c\|_{\mathcal{B}^*}$$

it holds that

$$\begin{aligned} d_n(t;T) \leq & \|T^* S_n^* c - T^* h\|_{H_0^k} \leq \|S_n^* c - h\|_{H_0^{k-\beta}} \qquad \text{[since } T^* \text{ is } \beta\text{-smoothing]} \\ \leq & \|(S_n^* c)^{(k-\beta)} - g\|_{L^2} + \|g - h^{(k-\beta)}\|_{L^2} \\ \leq & C n^{-k-\beta} \|g^{(k+\beta)}\|_{L^2} + C\lambda^{-s+k-\beta} \|h\|_{H_0^s} \\ \leq & C\lambda^{-s+k-\beta} \big(1 + (\lambda/n)^{k+\beta}\big) \|h\|_{H_0^s}. \end{aligned}$$

Choosing

$$\lambda \sim n^{1/(2s+4\beta+1)} (\log n)^{-2r/(2s+4\beta+1)},$$

we obtain that

$$\min_{t \geq 0} \big(d_n(t;T) + (\log n)^{r/2} t^{1/2}\big) = \mathcal{O}\big(n^{-\mu} (\log n)^{2r\mu}\big) \qquad \text{with} \quad \mu = \frac{s-k+\beta}{2s+4\beta+1} \tag{B.5}$$

as $n \to \infty$.

Note that the fact that T and T^* are β-smoothing implies that $\mathcal{C}_0^\infty(\mathbb{T}) \subset \mathrm{Ran}(T)$. Thus, the assertion follows by (B.5) and Theorem 3.2.6. □

List of Symbols

$\#S$	The number of elements in set S
$\lvert S\rvert$	The Lebesgue measure of set S
Γ_n	The regular grid on $[0,1]^d$, page 11
$\mathbb{N}_0$	The set of non-negative integers, i.e. $\mathbb{N} \cup \{0\}$
$\lVert\cdot\rVert_{\mathcal{B}}$	The multiresolution norm w.r.t. the system $\mathcal{B}$ of cubes, page 16
$\lVert\cdot\rVert_{\mathrm{TV}}$	The total variation (TV) semi-norm, page 67
$\lVert\cdot\rVert_{H_0^k}$, $\lVert\cdot\rVert_{W_0^{k,p}}$	The homogeneous Sobolev norm, page 15
$\lVert\cdot\rVert_{L^p}$	The L^p-norm w.r.t. the Lebesgure measure
$\lVert\cdot\rVert_p$	The ℓ^p-norm w.r.t. the counting measure
$\lVert\cdot\rVert_{W^{s,p}}$, $\lVert\cdot\rVert_{B_p^{s,p'}}$	The Besov/Sobolev norms, page 14
$\mathrm{Ran}(T)$	The range of operator T
$\mathbb{T}^d$	The d-dimensional torus $\mathbb{R}^d/\mathbb{Z}^d$
$D^l f$	The partial weak derivatives $\left(D^\alpha f\right)_{\lvert\alpha\rvert=l,\, \alpha\in\mathbb{N}_0^d}$ of order $l \in \mathbb{N}_0$ for function f, page 13
$d_n(t)$	The multiscale distance function for nonparametric regression, page 22
$d_n(t;T)$	The multiscale distance function for statistical inverse problems, page 61
$H_0^s(\mathbb{T}^d)$	The Sobolev space $W_0^{s,2}(\mathbb{T}^d)$, page 15
S_n	The point evaluation (or the sampling operator) on the regular grid Γ_n, page 12
T_n	The operator $S_n \circ T$, page 59
$W^{s,p}(\mathbb{T}^d)$, $B_p^{s,p'}(\mathbb{T}^d)$	The Sobolev/Besov spaces, page 14
$W_0^{s,p}(\mathbb{T}^d)$, $B_{p,0}^{s,p'}(\mathbb{T}^d)$	The subspace of Sobolev/Besov spaces consisting of functions with zero mean, page 15

Bibliography

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.

Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, 96(455):939–967. With discussion and a rejoinder by the authors.

Aspelmeier, T., Egner, A., and Munk, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.*, 2:163–202.

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

Bernholt, T., Eisenbrand, F., and Hofmeister, T. (2007). A geometric framework for solving subsequence problems in computational biology efficiently. In *Computational geometry (SCG'07)*, pages 310–318. ACM, New York.

Bernholt, T., Eisenbrand, F., and Hofmeister, T. (2009). Constrained Minkowski sums: a geometric framework for solving interval problems in computational biology efficiently. *Discrete Comput. Geom.*, 42(1):22–36.

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.

Besicovitch, A. S. (1947). Corrigenda to the paper "A general form of the covering principle and relative differentiation of additive functions. II.". *Proc. Cambridge Philos. Soc.*, 43:590.

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

Birgin, E. G. and Raydan, M. (2005). Robust stopping criteria for Dykstra's algorithm. *SIAM J. Sci. Comput.*, 26(4):1405–1414 (electronic).

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.

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A. (2006). *Numerical optimization*. Universitext. Springer-Verlag, Berlin, second edition. Theoretical and practical aspects.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122.

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 (Iowa City, Iowa, 1985)*, volume 37 of *Lecture Notes in Statist.*, pages 28–47. Springer, Berlin.

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.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Butucea, C., Matias, C., and Pouet, C. (2009). Adaptive goodness-of-fit testing from indirect observations. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(2):352–372.

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 Process.*, 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. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19.

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.

Chan, T. F. and Shen, J. (2005). *Image processing and analysis*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Variational, PDE, wavelet, and stochastic methods.

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.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of numerical integration*. Computer Science and Applied Mathematics. Academic Press, Inc., Orlando, FL, second edition.

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.

Deutsch, F. and Hundal, H. (1994). The rate of convergence of Dykstra's cyclic projections algorithm: the polyhedral case. *Numer. Funct. Anal. Optim.*, 15(5-6):537–565.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

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.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.*, 78(384):837–842.

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.

Ekeland, I. and Témam, R. (1999). *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition. Translated from the French.

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.

Evans, S. N. and Stark, P. B. (2002). Inverse problems as statistics. *Inverse Problems*, 18(4):R55–R97.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

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.

Fortin, M. and Glowinski, R. (1983). *Augmented Lagrangian methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam. Applications to the numerical solution of boundary value problems, Translated from the French by B. Hunt and D. C. Spicer.

Freitag, D. (1978). Real interpolation of weighted L_p-spaces. *Math. Nachr.*, 86:15–18.

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, Inc., Boston, MA.

Glaz, J., Pozdnyakov, V., and Wallenstein, S., editors (2009). *Scan statistics*. Statistics for Industry and Technology. Birkhäuser Boston, Inc., Boston, MA. Methods and applications.

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

Goldenshluger, A. and Pereverzev, S. V. (2000). Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory Related Fields*, 118(2):169–186.

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.

Grasmair, M., Li, H., and Munk, A. (2015). Variational multiscale nonparametric regression: smooth functions. *arXiv:1512.01068*.

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.

Griebel, M. and Hamaekers, J. (2014). Fast discrete fourier transform on generalized sparse grids. In Garcke, J. and Pflüger, D., editors, *Sparse Grids and Applications - Munich 2012*, volume 97 of *Lecture Notes in Computational Science and Engineering*, pages 75–107. Springer International Publishing.

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.

Haltmeier, M. and Munk, A. (2015). A variational view on statistical multiscale estimation. In preparation (via private communication).

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.

Holzmann, H., Bissantz, N., and Munk, A. (2007). Density testing in a contaminated sample. *J. Multivariate Anal.*, 98(1):57–75.

Ingster, Y. I., Sapatinas, T., and Suslina, I. A. (2012). Minimax signal detection in ill-posed inverse problems. *Ann. Statist.*, 40(3):1524–1549.

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.

Kabluchko, Z. and Munk, A. (2009). Shao's theorem on the maximum of standardized random walk increments for multidimensional arrays. *ESAIM Probab. Stat.*, 13:409–416.

Kaipio, J. and Somersalo, E. (2005). *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

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

Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

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

Kuipers, L. and Niederreiter, H. (1974). *Uniform distribution of sequences*. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney. Pure and Applied Mathematics.

Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods*, 26(6):1481–1496.

Laurent, B., Loubes, J.-M., and Marteau, C. (2011). Testing inverse problems: a direct or an indirect problem? *J. Statist. Plann. Inference*, 141(5):1849–1861.

Laurent, B., Loubes, J.-M., and Marteau, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electron. J. Stat.*, 6:91–122.

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.

Li, H., Sieling, H., and Munk, A. (2014). FDR-control in multiscale change-point segmentation. *axXiv:1412.5844*.

Lions, P.-L. and Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979.

Mair, B. A. and Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5):1424–1444.

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.

Marnitz, P. (2010). *Statistical Multiresolution Estimatiors in Linear Inverse Problems – Foundations and Algorithmic Aspects*. PhD thesis, Georg-August-Universität Göttingen.

Mathé, P. (2006). The Lepskiĭ principle revisited. *Inverse Problems*, 22(3):L11–L15.

Mathé, P. and Pereverzev, S. V. (2006). Regularization of some linear ill-posed problems with discretized random noisy data. *Math. Comp.*, 75(256):1913–1929 (electronic).

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.

Natterer, F. (2001). *The mathematics of computerized tomography*, volume 32 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1986 original.

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. and Nemirovskii, A. (1994). *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Nirenberg, L. (1959). On elliptic partial differential equations. *Ann. Scuola Norm. Sup. Pisa (3)*, 13:115–162.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527. With comments and a rejoinder by the author.

Peetre, J. (1963a). Nouvelles propriétés d'espaces d'interpolation. *C. R. Acad. Sci. Paris*, 256:1424–1426.

Peetre, J. (1963b). Sur le nombre de paramétres dans la définition de certains espaces d'interpolation. *Ricerche Mat.*, 12:248–261.

Pein, F., Sieling, H., and Munk, A. (2015). Heterogeneuous change point inference. *arXiv:1505.04898*.

Phillips, D. L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *J. Assoc. Comput. Mach.*, 9:84–97.

Plaskota, L. (1996). *Noisy information and computational complexity*. Cambridge University Press, Cambridge.

Potra, F. A. and Wright, S. J. (2000). Interior-point methods. *J. Comput. Appl. Math.*, 124(1-2):281–302. Numerical analysis 2000, Vol. IV, Optimization and nonlinear equations.

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 Mathematical Library. 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*.

Shepp, L. and Logan, B. (1974). The fourier reconstruction of a head section. *IEEE Trans. Nucl. Sci.*, 21(3):21–43.

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.

Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559.

Sudakov, V. and Khalfin, L. (1964). Statistical approach to ill-posed problems in mathematical physics. *Soviet Math. Doklady*, 157:1094–1096.

Tenorio, L. (2001). Statistical regularization of inverse problems. *SIAM Rev.*, 43(2):347–366 (electronic).

Teuber, T., Steidl, G., and Chan, R. H. (2013). Minimization and parameter estimation for seminorm regularization models with I-divergence constraints. *Inverse Problems*, 29(3):035007, 28.

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

Tikhonov, A. N. (1963a). Regularization of incorrectly posed problems. *Soviet Math. Doklady*, 4:1624–1627.

Tikhonov, A. N. (1963b). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady*, 4:1035–1038.

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.

Curriculum Vitae

Name	Housen Li
Address	Max Planck Institute for Biophysical Chemistry Am Fassberg 11 37077 Göttingen, Germany
Email:	housen.li@mpibpc.mpg.de

Personal Details

Gender	Male
Date of birth	December 1, 1985
Place of birth	Liaoning, China
Citizenship	Chinese

Education

Since 10/2011	Ph.D. student of mathematics at the University of Göttingen and the Max Planck Institute for Biophysical Chemistry, Germany Supervisor: Prof. Dr. Axel Munk, Prof. Dr. Markus Haltmeier
03/2011-09/2011	Ph.D. student of mathematics at the National University of Defense Technology, China Supervisor: Prof. Dr. Lizhi Cheng
09/2008-12/2010	Master student of mathematics at the National University of Defense Technology, China Supervisor: Prof. Dr. Lizhi Cheng
09/2004-06/2008	Student of mathematics at the National University of Defense Technology, China

09/2001-06/2004	Secondary school "Fushun No. 12 middle school" in Liaoning, China
09/1998-06/2001	Middle school "Fushun No. 41 middle school" in Liaoning, China
09/1992-06/1998	Primary school "Fushun Jiangjun No.3 school" in Liaoning, China

---

Research Experience

Since 10/2015	Member of the Research Training Group 2088 "Discovering structure in complex data: Statistics meets Optimization and Inverse Problems"
07/2015-09/2015	Member of the Felix Bernstein Institute for Mathematical Statistics in the Bioscience (FBMS)
09/2011-08/2015	Scholarship under the State Scholarship Fund by the China Scholarship Council (CSC)
12/2012-06/2015	Scientific assistant of the SFB 755 "Nanoscale Photonic Imaging"