

DFG-SNF Research Group FOR916

Statistical Regularization and Qualitative Constraints

Klaus Frick

Philipp Marnitz

Axel Munk

Shape Constrained Regularisation by Statistical Multiresolution for Inverse Problems: Asymptotic Analysis

Preprint FOR916 10-33

Updated Version (01-06-2011)

Preprint-Series of the Research Group FOR916

**SHAPE CONSTRAINED REGULARISATION BY STATISTICAL MULTIREOLUTION
FOR INVERSE PROBLEMS: ASYMPTOTIC ANALYSIS
(LONG VERSION INCLUDING SUPPLEMENTARY MATERIAL)**

KLAUS FRICK

*Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen*

PHILIPP MARNITZ

*Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen*

AXEL MUNK

*Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen*

*Max Planck Institute for Biophysical Chemistry
Am Faßberg 11, 37077 Göttingen*

ABSTRACT. This paper is concerned with a novel regularisation technique for solving linear ill-posed operator equations in Hilbert spaces from data that is corrupted by white noise. We combine convex penalty functionals with extreme-value statistics of projections of the residuals on a given set of subspaces in the image-space of the operator. We prove general consistency and convergence rate results in the framework of Bregman-divergences which allows for a vast range of penalty functionals.

Various examples that indicate the applicability of our approach will be discussed. Especially it will turn out that in the context of image processing the presented method constitutes a fully data-driven method for denoising that additionally exhibits locally adaptive behaviour.

E-mail addresses: frick@math.uni-goettingen.de, marnitz@math.uni-goettingen.de, munk@math.uni-goettingen.de.

Key words and phrases. Statistical Inverse Problems; Multiresolution; Extreme-Value Statistics; Shape Constrained Regularisation; Bregman-divergence.

Correspondence to frick@math.uni-goettingen.de .

1. INTRODUCTION

In this paper, we are concerned with the solution of the equation

$$(1) \quad Ku = g,$$

where $K : U \rightarrow V$ is a linear and bounded operator mapping between two Hilbert-spaces U and V . Equations of type (1) are called *well-posed* if for given $g \in V$ there exists a unique solution $u \in U$ that depends continuously on the right-hand side g . If one of these conditions is not satisfied, the problem is called *ill-posed*. In the case of ill-posedness, arbitrary small deviations in the right hand side g may lead to useless solutions u (if solutions exist). These deviations are commonly modelled as random: They are due to indispensable numerical errors as well as to the random nature of the measurement process itself. (*Statistical*) *regularisation methods* aim at computing stable approximations of true solutions u from a (statistically) perturbed signal g .

In this paper we assume that $u \in U$ is a solution of (1) and that we are given the observation

$$(2) \quad Y = Ku + \sigma\varepsilon.$$

Here, $\sigma > 0$ denotes the noise-level and $\varepsilon : V \rightarrow L^2(\Omega, \mathfrak{A}, \mathbb{P})$ a Gaussian white noise process, i.e. ε is linear and continuous and for all $v, w \in V$ one has

$$\varepsilon(v) \sim \mathcal{N}(0, \|v\|^2) \quad \text{and} \quad \mathbf{Cov}(\varepsilon(v), \varepsilon(w)) = \langle v, w \rangle.$$

Model (2) is denoted as a white noise model and it is very common in the theory of statistical inverse problems [see e.g. 53, 66, 23, 20, 7, 25]. In model (2) it is technically easier to analyze estimation methods for u compared to the sampling model

$$(3) \quad Y_i = (Ku)(x_i) + \sigma\varepsilon_i,$$

for i.i.d. errors ε_i and sampling points $x_i, i = 1, \dots, n$. From an asymptotic point of view both models are equivalent in Le Cam's sense. See [15, 70] for the regression case (i.e. when K equals the embedding operator) and [65] for a specific inverse problems setting. Here $\sigma = 1/\sqrt{n}$, and the asymptotics $n \rightarrow \infty$ in the sampling model (3) corresponds to $\sigma \rightarrow 0$ in the white noise model (2).

Moreover, we consider the Poisson analogue to (3), i.e. we observe $Y_i \sim \text{Pois}(Ku(x_i)), i = 1, \dots, n$. Then it is well known (see [44]) that this again is asymptotically equivalent to a white noise model with expectation $2\sqrt{Ku}$ (for one dimensional observations Y_i). This is also highlighted by the fact that a Poisson variable Y can be approximated well by a normal one provided the intensity is not to small, e.g. by Anscombe's transformation $\sqrt{Y + 3/8}$ or variants of it ([14]). Another option is to exploit the standardization $(Y - \lambda)/\sqrt{\lambda}$ as we will explain in Example 1.2.

Hence, model (2) can be regarded as reasonable approximation relevant for many areas of applications. The simplest case is classical nonparametric regression and its amplitude of applications. Here K is an embedding operator $K : U \hookrightarrow L^2$ modeling the smoothness of the function u (cf. [7]). More complicated instances of K are given by various optical systems, where K is given by a convolution kernel $k(x - y)$ (in engineering and physical terminology denoted as a point spread function). This includes the correction of the optical aberration of the Hubble Space Telescope ([46]) which has been one of the most spectacular success stories of statistical deconvolution in astronomy. In general the specific form of the PSF depends on the specification of the optical system, e.g. the lenses and position and shape of the mirrors. Other examples includes interferometry ([5]) or the estimation of the luminosity of the Milky Way based on satellite data ([8]), to mention a few. Methodologically related to this is the rapidly growing area of far field fluorescence microscopy (see Example 1.2) where the PSF typically is created by a diffraction limited excitation spot (generated by a laser beam) focusing a plane wave. Recently, there has been achieved a revolution in this area by breaking the classical Abbe

diffraction limit by rather sophisticated techniques, e.g. by stimulated emission depletion microscopy (see [47, 48] for a survey) which leads to a different PSF than for confocal microscopy which is well known to be close to the normal convolution (cf. Example 1.2).

More general, most analysis methods of modern biophysical chemistry rely heavily on estimation methods in models of type (2). This includes Nuclear Magnetic Resonance (NMR) spectroscopy, which is one of the most sensitive techniques for structure determination of molecules (see [21]) or mass spectrometry such as matrix assisted laser desorption ionization or electrospray ionisation ([74]). Recently, single molecule measurement techniques have been developed with great success, which all require data processing by proper deconvolution techniques. We mention patch clamp for ion channel recordings ([79]), fluorescence correlation spectroscopy ([55]) for the analysis of biomolecular motors or the use of optical tweezers for force investigations exerted by single molecules ([10]). Finally, we stress that beyond convolution general Fredholm operators K occur in various other applications, e.g. in seismic engineering (see e.g. [22]), in material sciences ([69]), Magnetic resonance imaging ([9]), signal processing, tomography and network analysis ([57, 77]).

According to the amplitude of applications, literature on statistical regularisation methods is vast and similar methods have often been invented independently in various communities. We only give a few, selective references: Penalised least-squares estimation (that includes Tikhonov-Phillips and maximum entropy regularisation) [6, 67, 78], wavelet based methods [31, 32, 51, 54, 52], estimation in Hilbert-scales [7, 43, 59, 61, 62, 63] and regularisation by projection [19, 20, 25, 61, 50] to name but a few.

In this work, we study a variational estimation scheme that defines estimators \hat{u} as solutions of

$$(4) \quad \inf_{u \in U} J(u) \quad \text{subject to} \quad T(\sigma^{-1}(Y - Ku)) \leq q(\alpha).$$

Here, the function T induces some notion of *distance* on the image space V that measures the deviation of the data Y and the estimated image Ku and J is some *regularisation functional* that is supposed to measure the regularity of candidate estimators $u \in U$. The parameter $q(\alpha)$ is chosen to be the $(1 - \alpha)$ -quantile of the statistic $T(\varepsilon)$ and governs the trade-off between data-fit and regularity. Hence the *admissible region*

$$(5) \quad \mathcal{A}(\alpha) = \{u \in U : T(\sigma^{-1}(Y - Ku)) \leq q(\alpha)\}$$

constitutes a $(1 - \alpha)$ -confidence region for a solution \hat{u} of (4). This gives the estimation procedure (4) a precise statistical interpretation: Since for each solution u^\dagger of (1) one has $u^\dagger \in \mathcal{A}(\alpha)$ with probability at least $1 - \alpha$ it follows from (4) that

$$\mathbb{P}(J(\hat{u}) \leq J(u^\dagger)) \geq 1 - \alpha.$$

Summarizing, regularisation methods of type (4) pick among all estimators \hat{u} for which the distance between $K\hat{u}$ and the data Y does not exceed the threshold value $q(\alpha)$ one with largest regularity. The probability that this particular estimator is more regular than any solution of (1) is bounded from below by $1 - \alpha$. This is in contrast to many other regularisation techniques where regularisation parameters merely govern the trade-off between fit-to-data and smoothness and do not allow such an interpretation.

Whereas most of the literature is concerned with the proper choice of the regularisation functional J , in this work we will discuss the issue of the data fidelity term T . We claim that from a statistical perspective the choice of T is of equal importance as the choice of J . Whereas J comprises prior smoothness or regularity assumptions on u , T measures the statistically significant deviations between data and these regularity assumptions.

In this paper we will introduce a particular family of distance functions T , the so called *multiresolution statistics* (*MR-statistics*) within the framework of statistical inverse problems. In their simplest form (see also Definition 3.1), MR-statistics coincide with extreme-value statistics of projections of the residuals $Y - Ku$ onto a set of linear sub-spaces $\{\lambda \phi_n : \lambda \in \mathbb{R}\}$ for given elements $\phi_n \in V$ (with $\|\phi_n\| = 1$ and $n \in \mathbb{N}$), that is we define $T = T_N$ where

$$T_N(v) = \sup_{1 \leq n \leq N} |\langle v, \phi_n \rangle|, \quad v \in V.$$

Under the hypothesis that u is the true solution of (1), we have that $T_N(\sigma^{-1}(Y - Ku))$ does not exceed the threshold $q(\alpha)$ with probability of at least $1 - \alpha$. If, however, u is wrongly specified the residual $Y - Ku$ contains a non-random signal and for some $1 \leq n_0 \leq N$

$$(6) \quad \mathbf{E}(\langle Y - Ku, \phi_{n_0} \rangle) \neq 0.$$

As an effect the statistic $T_N(\sigma^{-1}(Y - Ku))$ becomes relatively large and u happens to lie outside the admissible domain of the optimisation problem (4). When $T = T_N$ is an MR-statistic, we call a solution \hat{u} of (4) *statistical multiresolution estimator (SMRE)* (see also Definition 3.3).

The choice of the dictionary $\{\phi_1, \phi_2, \dots\}$ is subtle, since it should not miss any non-random information in the residual, if present. In principle, T_N would be most sensible against a large variety of signals u , if we employ a large number N such that the image space V is approximated sufficiently well by $\text{span}\{\phi_1, \dots, \phi_N\}$. This approach, however, turns (4) into an optimisation problem with a huge number of constraints which is hard to tackle numerically and is treated separately [see 41]. Besides these numerical difficulties, there is also a statistical limitation which will be a major issue to be discussed in this paper: If the entropy of the system $\{\phi_n\}_{1 \leq n \leq N}$ becomes too large (as $N \rightarrow \infty$), the asymptotic distribution of T_N will degenerate and would hence be useless for our purposes. In practical situations, a priori knowledge on the true solution of (1) can be used in order to design dictionaries whose entropy guarantees a non-degenerate limit of T_N and in addition allows to derive rates of convergence of the SMRE to the true signal. A similar comment applies to the choice of the regularisation functional J which models a priori information on the regularity of the true solution.

Example 1.1. In the context of signal and image processing U and V usually represent suitable spaces of functions on some domain $\Omega \subset \mathbb{R}^d$ (such as Sobolev spaces). A dictionary to be investigated in this paper in more detail is the class $\{\phi_1, \phi_2, \dots\}$ of normalised indicator functions w.r.t. a system of subsets $\{S_1, S_2, \dots\}$ of Ω , i.e.

$$\phi_n(x) = \frac{1}{\lambda_d(S_n)} \begin{cases} 1 & \text{if } x \in S_n \\ 0 & \text{else.} \end{cases}$$

Here, the sets S_n are supposed to encode a priori information on the shape of the unknown signal u and λ_d denotes the d -dimensional Lebesgue-measure on Ω . The MR-statistic T_N thus reads

$$T_N(v) = \sup_{1 \leq n \leq N} \frac{1}{\lambda_d(S_n)} \left| \int_{S_n} v(x) dx \right|, \quad v \in V.$$

Hence choosing $T = T_N$ in (4) can be considered as *shape constraint* and the resulting estimation method is capable of adapting the amount of regularization in a *locally adaptive* manner. This is in contrast to the widely used squared norm fidelity $T(v) = \|v\|^2$ that does not allow for adaptation to local structures. We will give a detailed account of this particular instance of MR-statistics in Section 4.2.

The regularisation scheme (4) with MR-statistic T_N was studied in [30] for the specific case of non-parametric regression in one space dimension and the total-variation semi-norm as regularisation

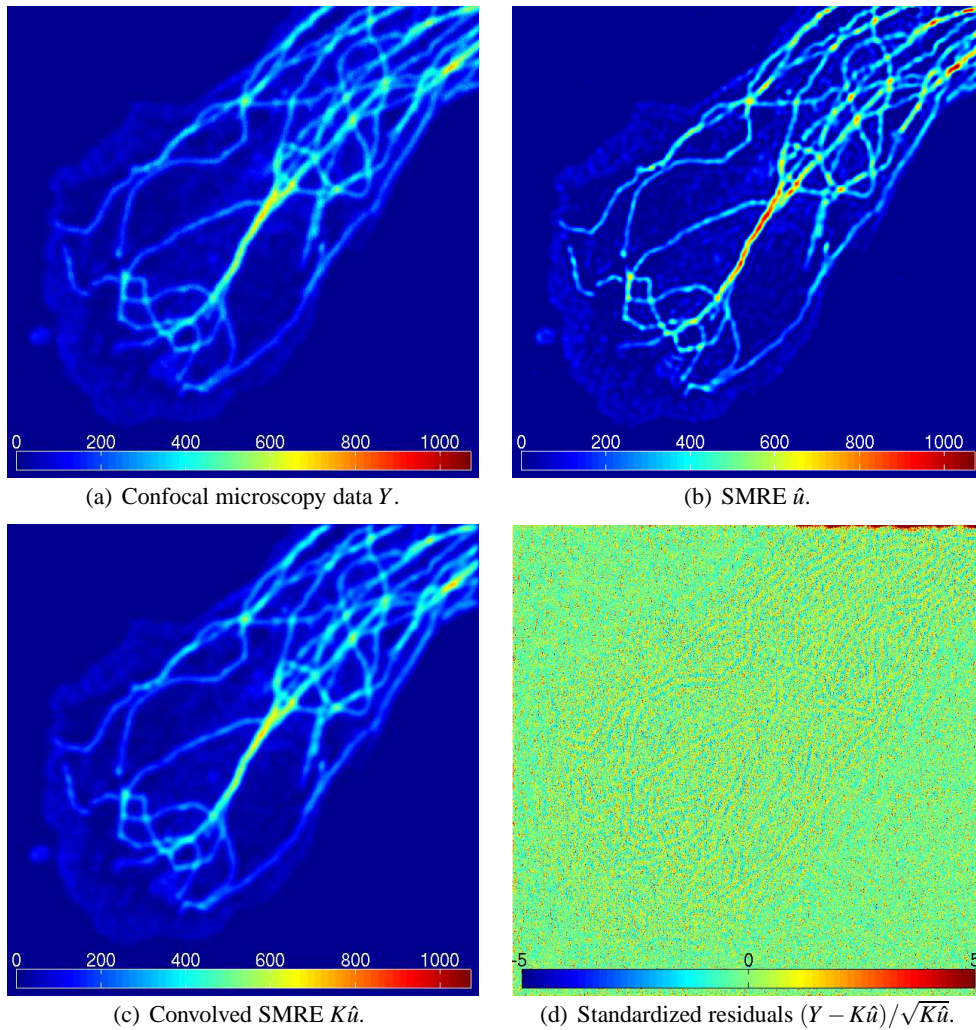


FIGURE 1. SMRE for image deblurring in confocal fluorescence microscopy (cf. Example 1.2)

functional J . In the following example we illustrate that the general formulation in (3) reveals the SMRE as a powerful regularisation method far beyond this situation: It can be extended to space dimensions larger than one as well as to inverse problems with general K as in (2) including deconvolution problems. We emphasise that the MR-constraint $T_N(\sigma^{-1}(Y - Ku)) \leq q(\alpha)$ has the appealing property that it adapts to local features of the signal (image) as it is required in many imaging problems.

Example 1.2. Figure 1(a) depicts a confocal microscopy recording of the β -tubulin distribution in a PtK2 cell taken from a kidney of *potorous tridactylus*¹. The image shows an area of $18 \times 18 \mu\text{m}^2$ at a resolution of 798×798 pixel. Before the recording, the protein β -tubulin is tagged with a fluorescent

¹By kind permission of the department of NanoBiophotonics at the Max Planck Institute for Biophysical Chemistry, Göttingen.

tracer. After point illumination by a laser beam the probe is scanned and the resulting photons are registered (for more details on confocal microscopy see [68]).

For each pixel (i, j) the observation Y_{ij} in Figure 1(a) can be modelled as a Poisson random variable with (unknown) intensity $(Ku)_{ij}$ where the operator K denotes the convolution with a point spread function (psf) of the optical system. It is custom to model the psf as a symmetric Gaussian kernel, in the present case with a full width at half maximum of 230nm. In the interior of the PtK2 cell the photon count-rates are sufficiently large (> 50) such that it is justified to assume that $(Y - Ku)_{ij}/\sqrt{(Ku)_{ij}} \sim \mathcal{N}(0, 1)$. In the outer region only background noise is detected that does not carry any information (photon count-rates up to 10 the most). Figure 1(b) shows the solution \hat{u} of

$$\inf_{u \in U} J(u) \quad \text{subject to} \quad T_N \left((Y - Ku)/\sqrt{Ku} \right) \leq q(\alpha)$$

with T_N being an MR-statistic w.r.t. the dictionary in Example 1.1 where the sets S_n are chosen to consist of all discrete squares of sidelength $1, \dots, 25$ pixel. Moreover, J is chosen to be the total variation semi-norm (cf. Section 4.3) and α is set to 0.1 such that a confidence level of 0.9 results.

As it becomes apparent from Figure 1(b), the reconstruction exhibits an appealing locally adaptive behaviour due to the shape constraint in (4): The gaps between the β -tubulin filaments, that actually make up the multiscale nature of the image, are reconstructed equally well independent of their size. Figure 1(c) depicts the convolved estimator $K\hat{u}$ and Figure 1(d) shows the standardized residual $(Y - K\hat{u})/\sqrt{K\hat{u}}$. Visually, the residual in the interior of the PtK2 cell resembles white noise as desired. This example indicates that SMRE is a promising approach for locally adaptive image reconstruction. Theoretical evidence for this will be given in Section 4.3 and particularly in Example 4.15. For more examples, details on implementation and algorithms we refer to [41].

In this paper we present very general consistency and convergence rates results for SMRE in the context of statistical inverse problems and discuss their impact on particular applications. To our best knowledge, results of this type have never been obtained before. It is necessary to assume additional regularity of the true solution of (1) in order to come up with convergence rates results. In the context of inverse problems, this is usually done by formulating so-called *source conditions*. These determine smoothness classes of solutions for (1) that guarantee risk bounds and fast convergence of the estimator to the true signal. In this work we study the standard source conditions used in the framework of Bregman-divergences that yield for each penalty functional J in (4) *one* specific smoothness class. As shown in Section 4 this can be considered as a generalization of the Sobolev-class of functions with exponent $1/2$. The formulation of conditions that give optimal convergence rates in a *scale* of smoothness classes for a general but *fixed* J to our knowledge is still open and will not be treated in this work.

This paper is organised as follows. After reviewing some basic definitions from convex analysis and the theory of inverse problems in Section 2 we develop a general scheme for estimation of solutions of (1) in Section 3. We use the regularisation scheme (4) where we employ MR-statistics T_N as distance measures T (Section 3.1). In Section 3.2 we then prove consistency and convergence rate results in terms of the Bregman-divergence w.r.t. the regularisation functional J . In Section 4 we study the performance of the so constructed estimators for typical examples, as the Gaussian sequence model (Section 4.1) and linear inverse regression problems (Section 4.2). In Section 4.3 we investigate the particular situation when the regularisation functional J is chosen to be the total-variation semi-norm, which has a particular appeal for imaging problems. Finally, some examples that illustrate the notions of source-condition and Bregman-divergence are given in Appendix A and the proofs of the main results as well as some auxiliary lemmata are collected in Appendix B.

2. BASIC DEFINITIONS

In this section we summarise some relevant definitions and assumptions needed throughout the paper. We start with stating minimal assumptions on the functional analytic setting

Assumption 2.1. (i) U and V denote separable Hilbert spaces. The norms on U and V are not further specified, and will be always denoted by $\|\cdot\|$, since the meaning is clear from the context. (ii) Let $J : U \rightarrow \overline{\mathbb{R}}$ be a convex functional from U into the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. The domain of J is defined by

$$D(J) = \{u \in U : J(u) \neq \infty\}.$$

J is called proper if $D(J) \neq \emptyset$ and $J(u) > -\infty$ for all $u \in U$. Throughout this paper J denotes a convex, proper and lower semi-continuous (l.s.c.) functional with dense domain $D(J)$.

(iii) $K : U \rightarrow V$ is a linear and bounded operator. By $\text{ran}(K) = K(U)$ we denote the range of K .

In the course of this paper we will frequently make use of tools from convex analysis. For a standard reference see [36].

- The sub-differential (or generalised derivative) $\partial J(u)$ of J at u is the set of all elements $p \in U$ satisfying

$$J(v) - J(u) - \langle p, v - u \rangle \geq 0 \quad \text{for all } v \in U.$$

The domain $D(\partial J)$ of the sub-gradient consists of all $u \in U$ for which $\partial J(u) \neq \emptyset$.

- We will prove consistency of estimators with respect to the Bregman-divergence. For $u \in D(J)$ the Bregman-divergence of J between u and v is defined by

$$D_J(v, u) = J(v) - J(u) - J'(v)(v - u)$$

where $J'(v)(v - u)$ denotes the directional derivative of J at v in direction $v - u$. The directional derivative is defined as

$$J'(v)(w) = \lim_{h \rightarrow 0^+} \frac{J(v + hw) - J(v)}{h}.$$

and is well defined for convex functions (possibly with values in $[-\infty, \infty]$).

- For $u \in D(\partial J)$ the Bregman-divergence of J between u and v w.r.t. $\xi \in \partial J(u)$ is defined as

$$D_J^\xi(v, u) = J(v) - J(u) - \langle \xi, v - u \rangle.$$

The following basic estimates hold

$$0 \leq D_J(v, u) \leq D_J^\xi(v, u), \quad \text{for all } \xi \in \partial J(u).$$

Remark 2.1. Clearly, the Bregman-divergence does not define a (quasi-)metric on U : It is non-negative but in general it is neither symmetric nor satisfies the triangle inequality. The big advantage, however, of formalising asymptotic results w.r.t. to the Bregman-divergence (such as consistency or convergence rates) for estimators defined by a variational scheme of type (4), is the fact, that the regularising properties of the used penalty functional J are incorporated automatically. If, for example, the functional J is slightly more than strictly convex, it was shown in [71] that convergence w.r.t. the Bregman-divergence already implies convergence in norm. If, however, J fails to be strictly convex (e.g. if it is of linear growth) it is in general hard to establish norm-convergence results but convergence results w.r.t. the Bregman-divergence, though weaker, may still be at hand. In Examples A.1-A.4 as well as in Section 4.3 we compute the Bregman-divergence for some particular choices of J .

The concept of Bregman-divergence in optimisation was introduced in [12] and has recently attracted much attention e.g. in the inverse problems community [see 17, 42, 27] or in statistical and machine learning [26, 56, 80].

Next, we introduce different classes of solutions for Equation (1) discussed in this paper.

Definition 2.2. (i) Let $u \in D(J)$ be a solution of (1). Then g is called *attainable*.

(ii) An element $u \in D(J)$ is called *J-minimising solution* of (1), if u solves (1) and

$$J(u) = \inf \{J(\tilde{u}) : K\tilde{u} = g\}.$$

(iii) Let $g \in V$ be attainable. An element $p \in V$ is called a *source element* if there exists a *J-minimising solution* u of (1) such that

$$(7) \quad K^*p \in \partial J(u).$$

Then, we say that u satisfies the *source condition* (7).

It is well-known in the theory of inverse problems with deterministic noise [see 37] that the source condition (7) is sufficient for establishing convergence rates for regularisation methods. It can be understood as a regularity condition for *J-minimising solutions* of Equation (1). Put differently, for each regularisation functional J and each operator K , the source condition (7) characterises *one particular* smoothness-class of solutions for (1) for which fast reconstruction is guaranteed. We clarify the notions *Bregman-divergence* and *source condition* by some examples in Appendix A.

Under fairly general conditions existence of *J* minimising solution can be guaranteed. We formalise these conditions in the following result, however, we omit the proof since it is standard in convex analysis [see 36, Chap. II Prop. 2.1].

Proposition 2.3. Let $g \in V$ be attainable and assume that for all $c \in \mathbb{R}$ the sets

$$(8) \quad \{u \in U : \|Ku\| + J(u) \leq c\}$$

are sequentially weakly compact. Then, there exist a *J-minimising solution* of (1).

3. A GENERAL SCHEME FOR ESTIMATION

In this section we construct a family of estimators \hat{u} for *J*-minimising solutions (cf. Definition 2.2) of Equation (1) from noisy data Y given by the white noise model (2). We define the estimators in a variational framework and prove consistency as well as convergence rates results in a rather general setting.

3.1. MR-Statistic and SMR-Estimation. We introduce a class of similarity measures in order to determine whether the residuals $Y - K\hat{u}$ for a given estimator $\hat{u} \in U$ resemble a white noise process or not. We will consider the extreme-value distribution of projections of the residuals onto a predefined collection of lines in V . To this end, assume that

$$\Phi = \{\phi_1, \phi_2, \dots\} \subset \overline{\text{ran}(K)} \setminus \{0\}$$

is a fixed dictionary such that $\|\phi_n\| \leq 1$ for all $n \in \mathbb{N}$. For the sake of simplicity, we will frequently make use of the abbreviation $\phi_n^* = \phi_n / \|\phi_n\|$.

Definition 3.1. Let $\{t_N : \mathbb{R}^+ \times (0, 1] \rightarrow \mathbb{R}\}_{N \in \mathbb{N}}$ be a sequence of functions that satisfy the following conditions

(i) For all $r \in (0, 1]$, the function $s \mapsto t_N(s, r)$ is convex, increasing and Lipschitz-continuous with Lipschitz-constants L_{Nr} such that $L_{Nr} \leq L < \infty$ for all $N \in \mathbb{N}$ and

$$(9) \quad 0 > \lambda_N(r) := \inf_{s \in \mathbb{R}^+} t_N(s, r) > -\infty.$$

(ii) There exist constants $c_1, c_2 > 0$ and $\sigma_0 \in (0, 1)$ such that for all $0 < \sigma < \sigma_0$

$$(10) \quad t_N(s, r) \geq c_1 s + c_2 t_N(\sigma s, r) \quad \text{for } (s, r) \in \mathbb{R}^+ \times (0, 1] \text{ and } N \in \mathbb{N}.$$

Then, for $N \in \mathbb{N}$, the mapping $T_N : V \rightarrow \mathbb{R}$ defined by

$$T_N(v) = \sup_{1 \leq n \leq N} t_N(|\langle v, \phi_n^* \rangle|, \|\phi_n\|)$$

is called a *multiresolution statistic (MR-statistic)*.

Remark 3.1. Let $\varepsilon : V \rightarrow L^2(\Omega, \mathfrak{A}, \mathbb{P})$ be a white noise process and consider the random variables

$$T_N(\varepsilon) = \sup_{1 \leq n \leq N} t_N(|\varepsilon(\phi_n^*)|, \|\phi_n\|).$$

Then, for a level $\alpha \in (0, 1)$ we denote the $(1 - \alpha)$ -quantile of $T_N(\varepsilon)$ by $q_N(\alpha)$, that is,

$$(11) \quad q_N(\alpha) := \inf \{q \in \mathbb{R} : \mathbb{P}(T_N(\varepsilon) \leq q) \geq 1 - \alpha\}$$

Definition 3.1 allows for a vast class of MR-statistics and the conditions in (i) and (ii) appear rather technical. The following example sheds some light on a special class of MR-statistics that later on will be studied in more detail. We note, however, that our general setting also applies to more involved statistics, as e.g. introduced in [34, 35].

Example 3.2. Assume that $\{f_N : (0, 1] \rightarrow \mathbb{R}\}_{N \in \mathbb{N}}$ is a sequence of positive functions and define

$$t_N(s, r) := s - f_N(r).$$

Then, the assumptions in Definition 3.1 are satisfied; to be more precise, we can set $L = 1$, $\lambda_N(r) = -f_N(r)$ and $c_1 = 1 - \sigma_0$ and $c_2 = 1$, where $\sigma_0 \in (0, 1)$ is arbitrary but fixed.

Our key paradigm is that an estimator \hat{u} for a solution of (1) fits the data Y sufficiently well, if the statistic $T_N(Y - K\hat{u})$ does not exceed the threshold $q_N(\alpha)$ ($\alpha \in (0, 1)$ and $N \in \mathbb{N}$ fixed). Among all those estimators we shall pick the *most parsimonious* by minimising the functional J .

Definition 3.3. Let $N \in \mathbb{N}$ and $\alpha \in (0, 1)$. Moreover, assume that T_N is an MR-statistic and that Y is given by (2). Then every element $\hat{u}_N(\alpha) \in U$ solving the convex optimisation problem

$$(12) \quad \inf_{u \in U} J(u) \quad \text{s.t.} \quad T_N(\sigma^{-1}(Y - Ku)) \leq q_N(\alpha)$$

is called a *statistical multiresolution estimator (SMRE)*.

An SMRE $\hat{u}_N(\alpha)$ depends on the regularisation parameters $N \in \mathbb{N}$ and $\alpha \in (0, 1)$ that determine the admissible region $\mathcal{A}_N(\alpha)$ for $T = T_N$ in (5).

In order to guarantee existence of a solution of the convex problem in Definition 3.3, that is existence of an SMRE, it is necessary to impose further standard assumptions:

Assumption 3.4. *There exists $N_0 \in \mathbb{N}$ such that for all $c \in \mathbb{R}$ the sets*

$$\Lambda(c) = \left\{ u \in U : \sup_{1 \leq n \leq N_0} |\langle Ku, \phi_n^* \rangle| + J(u) \leq c \right\}$$

are sequentially weakly compact.

Assumption 3.4 guarantees (weak) compactness of the level sets of the objective functional J restricted to the admissible region $\mathcal{A}_N(\alpha)$. We note, that if J is strongly coercive (e.g. when J is as in Example A.1 or A.4) then Assumption 3.4 is satisfied without any restrictions on the operator K . If J

lacks strong coercivity (as it is e.g. the case with the total-variation semi-norm studied in Section 4.3) additional properties of K are required in order to meet Assumption 3.4.

Application of standard arguments from convex optimisation yields

Proposition 3.5. *Assume that Assumption 3.4 holds and let $N \geq N_0$ and $\alpha \in (0, 1]$. Then, an SMRE $\hat{u}_N(\alpha)$ exists.*

Finally, we note that Assumption 3.4 already implies the requirements in Proposition 2.3 and consequently existence of J -minimising solutions.

3.2. Consistency and Convergence Rates. We investigate the asymptotic behaviour of $\hat{u}_N(\alpha)$ as the noise level σ in (2) tends to zero. According to the reasoning following Definition 3.3, the parameters $N \in \mathbb{N}$ and $\alpha \in (0, 1)$ can be interpreted as regularisation parameters and have to be chosen accordingly: The model parameter N has to be increased in order to guarantee a sufficiently accurate approximation of the image space V , whereas the test-level α tends to 0 such that the true solution (asymptotically) satisfies the constraints of (12) almost surely. We formulate consistency and convergence rate results by means of the Bregman-divergence of the SMRE $\hat{u}_N(\alpha)$ and a true solution u^\dagger in terms of almost sure convergence.

Throughout this section we shall assume that Assumptions 2.1 and 3.4 hold and that $\{\sigma_k\}_{k \in \mathbb{N}}$ is a sequence of positive noise-levels in (2) such that $\sigma_k \rightarrow 0^+$ as $k \rightarrow \infty$. Moreover, we assume that $\{\alpha_k\}_{k \in \mathbb{N}} \subset (0, 1)$ is a sequence of significance levels and that $N_k \geq N_0$ is such that

$$(13) \quad \sum_{k=1}^{\infty} \alpha_k < \infty \quad \text{and} \quad \lim_{k \rightarrow \infty} N_k = \infty.$$

Theorem 3.6. *Let u^\dagger be a J -minimising solution of (1) where $g \in \overline{\text{span}\Phi}$ and assume that*

$$\sup_{N \in \mathbb{N}} T_N(\varepsilon) < \infty$$

and

$$(14) \quad \zeta_k := \sigma_k \max \left(\inf_{1 \leq n \leq N_k} \lambda_{N_k}(\|\phi_n\|), \sqrt{-\log \alpha_k} \right) \rightarrow 0.$$

Then, for $\hat{u}_k := \hat{u}_{N_k}(\alpha_k)$ as in (12) one has

$$(15) \quad \sup_{k \in \mathbb{N}} \|\hat{u}_k\| < \infty, \quad J(\hat{u}_k) \rightarrow J(u^\dagger) \quad \text{and} \quad D_J(u^\dagger, \hat{u}_k) \rightarrow 0 \quad \text{a.s.}$$

as well as

$$(16) \quad \limsup_{k \rightarrow \infty} \sup_{1 \leq n \leq N_k} \frac{|\langle \phi_n^*, K\hat{u}_k - Ku^\dagger \rangle|}{\zeta_k} < \infty \quad \text{a.s.}$$

Theorem 3.6 states that if for a given vanishing sequence of noise levels σ_k , suitable (in the sense of (14)) sequences of regularisation parameters N_k and α_k can be constructed, then the sequences of corresponding SMRE converges to a true J -minimising solution u^\dagger w.r.t. the Bregman-divergence. We note that the assumption on the boundedness of MR-statistic $T_N(\varepsilon)$ is crucial and in general non-trivial to show.

It is well known that without further regularity restrictions on u^\dagger , the speed of convergence in (15) can be arbitrarily slow. *Source conditions* as in Definition 2.2 (iii) are known to constitute sufficient regularity conditions with quadratic fidelity T (cf. [59, 7, 58]). In our situation, where the fidelity controls the maximum over all residuals, we additionally have to assume that the source elements exhibit certain approximation properties:

Assumption 3.7. *There exists a J -minimising solution u^\dagger of (1) that satisfies the source condition (7) with source element p^\dagger . Moreover, for $n, N \in \mathbb{N}$ there exist $b_{n,N} \in \mathbb{R}$ such that*

$$(17) \quad \text{err}_N(p^\dagger) := \left\| p^\dagger - \sum_{n=1}^N b_{n,N} \phi_n^* \right\| \rightarrow 0 \quad \text{and} \quad \sup_{N \in \mathbb{N}} \sum_{n=1}^N |b_{n,N}| < \infty.$$

Remark 3.2. i) Assumption 3.7 amounts to say that there exists a J -minimising solution u^\dagger that satisfies the source condition (7) with a source element p^\dagger that can be approximated sufficiently well by the used dictionary Φ . From (7) it becomes clear that we can always assume that $p^\dagger \in \text{ran}(K)$, such that the first condition in (17) is not very restrictive, in fact.

ii) Good estimates of approximation errors for non-orthogonal dictionaries Φ are hard to come up with in general. Examples of non-orthogonal dictionaries where such estimates are available are wavelet- [28] and curvelet- [18] frames.

iii) It is important to note that, given prior information on the true solution u^\dagger , the conditions in Assumption 3.7 may indicate whether a given dictionary is well suited for the reconstruction of u^\dagger or not. As we will see in Section 4, a priori information on the smoothness of u^\dagger can typically be employed.

Theorem 3.8. *Let the requirements of Theorem 3.6 be satisfied and assume further that Assumption 3.7 holds with $g \in \overline{\text{span}}\Phi$. If $\eta_k := \max(\zeta_k, \text{err}_{N_k}(p^\dagger)) \rightarrow \infty$, then*

$$(18) \quad \limsup_{k \rightarrow \infty} \frac{D_J^{K^*} p^\dagger(\hat{u}_k, u^\dagger)}{\eta_k} < \infty \quad \text{and} \quad \limsup_{k \rightarrow \infty} \sup_{1 \leq n \leq N_k} \frac{|\langle \phi_n^*, K\hat{u}_k - Ku^\dagger \rangle|}{\eta_k} < \infty \quad a.s.$$

Remark 3.3. The convergence rate result in Theorem 3.8 is rather general, in the sense that the rate function η_k in (18) has to be determined for each choice of K , J and Φ separately. We outline a general procedure how this can be done in practice: assume that u^\dagger is a J -minimising solution of (1) that satisfies Assumption 3.7 with a source element p^\dagger .

(i) The sequence $\{-\inf_{1 \leq n \leq N} \lambda_N(\|\phi_n\|)\}_{N \in \mathbb{N}}$ is positive according to (9). Hence

$$N_k := \inf \left\{ N \in \mathbb{N} : \text{err}_N(p^\dagger) \leq -\sigma_k \inf_{1 \leq n \leq N} \lambda_N(\|\phi_n\|) \right\}$$

is well-defined and since $\{\sigma_k\}_{k \in \mathbb{N}}$ is non-increasing one has $N_k \leq N_{k+1}$ and $N_k \rightarrow \infty$ as $k \rightarrow \infty$.

(ii) After setting $\eta_k = -\sigma_k \inf_{1 \leq n \leq N_k} \lambda_{N_k}(\|\phi_n\|)$ it remains to check that the sequence of test-levels $\alpha_k = \exp\left(-(\kappa \eta_k / \sigma_k)^2\right)$ is summable (for some constant $\kappa > 0$).

For the so constructed sequences N_k , η_k and α_k , the assertions of Theorem 3.8.

4. APPLICATIONS AND EXAMPLES

In Section 3 we developed a general method for estimation of J -minimising solutions of linear and ill-posed operator equations from noisy data. Our estimation scheme thereby employed the MR-statistic T_N (cf. Definition 3.1). In this section we will study particular instances of MR-statistics covered by the general theory in Section 3:

- We study the case where T_N constitutes the extreme-value statistic of the coefficients w.r.t. an orthonormal dictionary Φ (Section 4.1). We show how Assumption 3.7 in this case reduces to the requirement that the true solution u^\dagger lies in a Sobolev-ellipsoid w.r.t. the system Φ . Moreover, it will turn out that for the case when Φ denotes the eigensystem of a compact operator, SMR estimation can be considered as soft-thresholding.

- In Section 4.2 we skip the assumption of orthonormality and examine general SMR-estimation w.r.t. (non-orthonormal) dictionaries that satisfy certain entropy conditions. In particular, we will consider the case when $U = V = L^2([0, 1]^d)$ and when Φ consists of indicator functions w.r.t. a redundant system of subcubes in $[0, 1]^d$. As indicated in Example 1.1, the main application we here have in mind is locally adaptive imaging.
- Finally, we study the case when the penalty functional J is chosen to be the total-variation semi-norm on $U = L^2(\Omega)$ in Section 4.3. We shed some light on the meaning behind the source-condition (7) and the Bregman-divergence for total-variation regularisation, complementing the examples in Appendix A. Additionally, we highlight the implications of our general convergence rate results for image deconvolution (cf. Example 1.2).

Throughout this section we assume that Assumptions 2.1 and 3.4 hold. Moreover, we shall agree upon $\{\sigma_k\}_{k \in \mathbb{N}}$ being a sequence of noise levels such that $\sigma_k \rightarrow 0^+$ and that for $k \in \mathbb{N}$ there are $\alpha_k \in (0, 1)$ and $N_k \in \{N_0, N_0 + 1, \dots\}$ such that (13) holds.

4.1. Introductory Example: Gaussian Sequence Model. In this section we shall consider the case where the dictionary $\Phi = \{\phi_1, \phi_2, \dots\}$ constitutes an orthonormal basis of $\overline{\text{ran}(K)}$. Evaluation of Equation (2) at the elements ϕ_n hence yields

$$y_n = \theta_n + \sigma \varepsilon_n,$$

where $Y(\phi_n) = y_n$, $\theta_n = \langle Ku, \phi_n \rangle$ and $\varepsilon_n = \varepsilon(\phi_n)$. We define the MR-statistic T_N by setting $t_N(s, r) = s - \sqrt{2 \log N}$ in Definition 3.1. In other words, we consider the maximum of the coefficients w.r.t. to the dictionary Φ , that is

$$(19) \quad T_N(v) = \sup_{1 \leq n \leq N} |\langle v, \phi_n \rangle| - \sqrt{2 \log N}.$$

Since $\{\phi_1, \phi_2, \dots\}$ are linearly independent and normalised, it follows that the random variables $\varepsilon_1, \varepsilon_2, \dots$ are independent and standard normally distributed. This implies that $\sup_{N \in \mathbb{N}} T_N(\varepsilon) < \infty$ holds almost surely.

In what follows, we will apply Theorems 3.6 and 3.8 to the present case. To this end, we observe that for $\sigma > 0$ and $N \in \mathbb{N}$ it follows that

$$-\sigma \inf_{1 \leq n \leq N} \lambda_N(\|\phi_n\|) = \sigma \sqrt{2 \log N}.$$

With the above preparations, we are able to reformulate the consistency result in Theorem 3.6.

Corollary 4.1. Let $u^\dagger \in U$ be a J -minimising solution of (1) where $g \in \overline{\text{span}\Phi}$. Moreover, assume that $\sigma_k^2 \max(\log N_k, -\log \alpha_k) \rightarrow 0$. Then, the SMRE $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$ almost surely satisfies (15) and (16).

In order to apply the convergence rate result in Theorem 3.8, Assumption 3.7 has to be verified. We set $b_{n,N} \equiv \langle p^\dagger, \phi_n \rangle$ in Assumption 3.7. Note that the expression $\text{err}_N(p)$ denotes the approximation error of the N -th partial Fourier-series w.r.t. Φ . Thus, Assumption 3.7 is linked to absolute summability of the Fourier-coefficients w.r.t. the basis Φ , i.e.

$$(20) \quad \sum_{n=1}^{\infty} |\langle p^\dagger, \phi_n \rangle| < \infty$$

The *Bernstein-Stechkin criterion* is a classical method for testing for absolute summability. We present a version suitable for our purpose in the following

Proposition 4.2. Let $p^\dagger \in V$. Then, (20) is satisfied if $\sum_{N=1}^{\infty} \text{err}_N(p^\dagger) / \sqrt{N} < \infty$.

Proof. The classical version of the Bernstein-Stechkin Theorem [see e.g. 64, Thm. 7.4] states that for each $f \in L^2([0, 1])$ and each ON-basis $\underline{v} = \{v_1, v_2, \dots\}$ of $L^2([0, 1])$, the Fourier-coefficients of f are absolutely summable, if $\sum_{N=1}^{\infty} \text{err}_N(p^\dagger)/\sqrt{N} < \infty$. Since each separable Hilbert space is isometrically isomorph to $L^2([0, 1])$, the assertion finally follows. \square

Following the procedure outlined in Remark 3.3 (Section 3) we define

$$(21) \quad N_k := \inf \left\{ N \in \mathbb{N} : \text{err}_N(p^\dagger) \leq \sigma_k \sqrt{2 \log N} \right\} \quad \text{and} \quad \eta_k := \sigma_k \sqrt{2 \log N_k}.$$

Corollary 4.3. Let $g \in V$ be attainable and $u^\dagger \in U$ be a J -minimising solution of (1) that satisfies the source condition with a source element p^\dagger such that (4.2) holds. Moreover, let N_k and η_k be defined as in (21). If

$$\alpha_k := e^{-\left(\frac{\kappa \eta_k}{\sigma_k}\right)^2} = N_k^{-2\kappa^2} \in \ell^1(0, 1)$$

for a constant $\kappa > 0$, then the SMRE $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$ almost surely satisfies (18).

The problem of characterising those elements $p^\dagger \in V$ that satisfy the assumption of Proposition (4.2) is a classical issue in Fourier-analysis and approximation theory. Sufficient condition are usually formalised by characterising the decay properties of the Fourier-coefficients. In a function space setting, this leads to particular smoothness classes of functions and in the general situation can be given in terms of *Sobolev ellipsoids*: for a constants $\beta, Q > 0$ we define $\Theta(\beta, Q)$ as the infinite-dimensional ellipsoid

$$(22) \quad \Theta(\beta, Q) = \left\{ \theta \in \ell^2 : \sum_{n \in \mathbb{N}} n^{2\beta} \theta_n^2 \leq Q^2 \right\}.$$

The *Sobolev class* $W(\beta, Q) \subset V$ is then defined to consists of all $v \in V$ such that $\{\langle v, \phi_n \rangle\}_{n \in \mathbb{N}} \subset \Theta(\beta, Q)$ [see 75, Sec.1.10.1]. For $v \in W(\beta, Q)$ we have that Proposition 4.2 is applicable if $\beta > 1/2$.

Example 4.4. Assume that $J(u) = \frac{1}{2} \|u\|^2$ and let K be a compact operator with singular value decomposition (SVD) $\{(\psi_n, \phi_n, s_n)\}_{n \in \mathbb{N}}$: $\{\psi_n\}_{n \in \mathbb{N}}$ is an orthonormal basis (ONB) of $\ker(K)^\perp$, $\{\phi_n\}_{n \in \mathbb{N}}$ is an ONB of $\text{ran}(K)$ and the singular values $\{s_n\}_{n \in \mathbb{N}}$ are positive and $s_n \rightarrow 0$ as $n \rightarrow \infty$. Moreover

$$(23) \quad K\psi_n = s_n\phi_n \quad \text{and} \quad K^*\phi_n = s_n\psi_n,$$

for all $n \in \mathbb{N}$. For $N \in \mathbb{N}$ and $\alpha \in (0, 1]$ it turns out (e.g. by applying the method of Lagrangian multipliers) that the SMRE $\hat{u}_N(\alpha)$ with T_N as in (19) is a *shrinkage estimator* given by

$$\hat{u}_N(\alpha) = \sum_{n=1}^N s_n^{-1} y_n \left(1 - \frac{q_N(\alpha) + \sqrt{2 \log N}}{|y_n|} \right)_+ \psi_n.$$

We note that $\hat{u}_N(\alpha)$ is a particular instance of a soft thresholding estimator.

Now, let $u^\dagger \in U$ be a minimum-norm solution of (1) that satisfies the source condition $K^*p^\dagger = u^\dagger$ (cf. Example A.1) with source element $p^\dagger \in W(\beta, Q)$ for $Q > 0$ and $\beta > 1/2$. Then, $\text{err}_N(p^\dagger) \leq QN^{-\beta}$ and it follows from (21) that

$$N_k \sim \left(\frac{Q}{\sigma_k} \right)^2 \quad \text{and} \quad \eta_k \sim \sigma_k \sqrt{-\log \sigma_k}.$$

If σ_k has polynomial decay, we can choose a constant $\kappa > 0$ such that $\alpha_k = \exp(-(\kappa \eta_k / \sigma_k)^2) = \sigma_k^{\kappa^2}$ is summable and it follows from Corollary 4.3 and Example A.1 that

$$\limsup_{k \rightarrow \infty} \frac{1}{\sigma_k \sqrt{-\log \sigma_k}} \|u^\dagger - \hat{u}_{N_k}(\alpha_k)\|^2 < \infty \quad \text{a.s.}$$

This corresponds to the choice $\gamma_k = \sigma_k \sqrt{-\log \sigma_k}$ in [6].

As mentioned above, sufficient conditions for the Bernstein-Stechkin criterion (cf. Proposition 4.2) in a function space setting are usually formalised in characterising smoothness properties. The following example shows how this applies to Hölder-continuity.

Example 4.5. Let $V = L^2_{\text{per}}([0, 1])$ be the Hilbert space of all square-integrable and periodic functions on the unit interval. Moreover, we assume that $\overline{\text{ran}(K)} = L^2([0, 1])$ and consider the *trigonometric basis*

$$\phi_{2n} = \sqrt{2} \cos(n\pi x) \quad \text{and} \quad \phi_{2n+1} = \sqrt{2} \sin(n\pi x).$$

Assume that $p^\dagger \in \mathcal{H}_\beta([0, 1]) \cap V$ (cf. Definition B.4) with $\beta > 1/2$. Then we have that $\text{err}_N(p^\dagger) \leq QN^{-\beta} \log N$ for a suitable constant $Q > 0$ and therefore it follows from Proposition 4.2 that (20) holds.

Hence, if u^\dagger is a J -minimising solution of (1) that satisfies the source condition (7) with source element $p^\dagger \in \mathcal{H}_\beta([0, 1])$ and if the sequences N_k, η_k and α_k are chosen as in Example 4.4, then $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$ almost surely satisfy (18).

Remark 4.1. i) The assertions of Example 4.5 still hold if the trigonometric basis is replaced by any other orthonormal basis $\{\phi_n\}_{n \in \mathbb{N}}$ of $\overline{\text{ran}(K)}$ such that the Bernstein-Stechkin criterion 4.2 is satisfied. This holds for example for a vast class of orthonormal wavelet bases of $L^2([0, 1])$ as studied in [24].

ii) For the trigonometric basis in Example 4.5, the Bernstein-Stechkin criterion 4.2 can be replaced by the requirement that $p^\dagger \in \mathcal{H}_\beta([0, 1])$ for any $\beta > 0$ is additionally of bounded variation [see 82, Vol.1 Thm.3.6].

4.2. Non-orthogonal Models. In contrast to Section 4.1, where we considered orthonormal dictionaries, we will now focus on more general (non-orthogonal) systems. In other words, we consider sequences

$$\Phi = \{\phi_1, \phi_2, \dots\} \subset \overline{\text{ran}(K)} \setminus \{0\}$$

and assume that $\|\phi_n\| \leq 1$ for all $n \in \mathbb{N}$. Moreover, we will make use of the MR-statistic T_N (cf. Definition 3.1) defined by

$$(24) \quad t_N(s, r) = s - \sqrt{-2\gamma \log r}, \quad (s, r) \in \mathbb{R}^+ \times (0, 1]$$

where $\gamma > 0$ is a constant. As outlined in Example 3.2, one verifies that $t_N(s, r)$ satisfies the assumptions of Definition 3.1. In particular, we find that $\lambda_N(r) = -\sqrt{-2\gamma \log r} > -\infty$ for all $r \in (0, 1]$.

The parameter γ that appears in (24) has to be chosen appropriately in dependence on Φ in order to guarantee that the MR-statistic $T_N(\varepsilon)$ is bounded almost surely. A sufficient condition on γ has for example been given in [35, Thm 7.1]

Proposition 4.6. *If there exists constants $A, B > 0$ such that*

$$(25) \quad D(u\delta, \{\phi \in \Phi : \|\phi\| \leq \delta\}) \leq Au^{-B} \delta^{-\gamma}, \quad \text{for all } u, \delta \in (0, 1]$$

then almost surely $\sup_{N \in \mathbb{N}} T_N(\varepsilon) < \infty$. Here D denotes the capacity number (cf. Definition B.6).

Corollary 4.7. Let $u^\dagger \in U$ be a J -minimising solution of (1) where $g \in \overline{\text{span}\Phi}$ and $\gamma > 0$ be chosen such that the assumption of Proposition 4.6 is satisfied. Moreover, assume that

$$\sigma_k^2 \min_{1 \leq n \leq N_k} (\log(\|\phi_n\|), \log \alpha_k) \rightarrow 0.$$

Then, the SMRE $\hat{u}_k = \hat{u}_k(\alpha_k)$ almost surely satisfies (15).

In order to apply the convergence rate results in Theorem 3.8, it is necessary that a J -minimising solution u^\dagger of (1) satisfies the source condition (7) with a source element p^\dagger that can be approximated by the dictionary Φ sufficiently well (cf. Assumption 3.7). We illustrate the assertion of Theorem 3.8 when $U = V = L^2([0, 1]^d)$ ($d \geq 1$) and when Φ consists of a countable selection of indicator functions on cubes in $[0, 1]^d$ (cf. Example 1.1).

First, we shall examine when Proposition 4.6 holds. To this end, we will focus first on the (uncountable) collection Φ_s of indicator functions on cubes in $[0, 1]^d$. Then, according to Proposition B.8, the assumptions of Proposition 4.6 are satisfied for $\Phi = \Phi_s$ and $\gamma = d$. Particularly, it follows that the assertion of Proposition 4.6 also holds for arbitrary (countable) sub-systems $\Phi \subset \Phi_s$, that is the statistic

$$T_N(\varepsilon) = \sup_{1 \leq n \leq N} |\varepsilon(\chi_{Q_n})| - \sqrt{-d \log(\lambda_d(Q_n))} \quad \text{where} \quad \chi_{Q_n} \in \Phi$$

stays bounded a.s. as $N \rightarrow \infty$ (note here, that $\|\chi_{Q_n}\| = \sqrt{\lambda_d(Q)}$).

Next, we study Assumption 3.7 in the present setting. Let $\mathcal{P} = \{Q_1, Q_2, \dots\}$ be a countable system of cubes and set $\Phi = \{\chi_{Q_n} : n \in \mathbb{N}\}$. We shall assume that \mathcal{P} satisfies the conditions of Lemma B.5 (where $X = [0, 1]^d$ and $A_i = Q_i$ for $i \in \mathbb{N}$). Let $\{n_l\}_{l \in \mathbb{N}}$ and $\{\delta_l\}_{l \in \mathbb{N}}$ be defined accordingly. Moreover, we define

$$\varepsilon_l = \inf_{n_l < j \leq n_{l+1}} \sqrt{\lambda_d(Q_j)} = \inf_{n_l < j \leq n_{l+1}} \|\chi_{Q_j}\|,$$

where we assume that $\{\varepsilon_l\}_{l \in \mathbb{N}}$ is non-increasing. This means that we decompose the set $[0, 1]^d$ into sub-cubes $\{I_j\}_{n_l < j \leq n_{l+1}}$ whose size (or *scale*) is bounded by $[\varepsilon_l, \delta_l]$. It is more natural to formulate convergence rate results in terms of the total number m of used scales rather than in the total number of sub-cubes $N = N(m) = n_{m+1}$. Following Remark 3.3 and applying Lemma B.5 we therefore define for a given continuous function $p^\dagger : [0, 1]^d \rightarrow \mathbb{R}$

$$(26) \quad m_k := \inf \left\{ m \in \mathbb{N} : \frac{m+1}{\sum_{v=0}^m \omega^{-2}(\delta_v, p^\dagger)} \leq -2\sigma_k^2 \log \varepsilon_m \right\} \quad \text{and} \quad \eta_k := \sigma_k \sqrt{-2 \log \varepsilon_{m_k}}.$$

Here $\omega(\cdot, p^\dagger)$ denotes the modulus of continuity of p^\dagger (cf. Definition B.4). With this and the general convergence rate result in Theorem 3.8 we immediately obtain

Corollary 4.8. Let $u^\dagger \in L^2([0, 1]^d)$ be a J -minimising solution of (1) where $g \in \overline{\text{span}\Phi}$ and that satisfies the source condition (7) with source element $p^\dagger \in C([0, 1]^d)$. Moreover, let m_k and η_k be defined as in (26). If

$$\lim_{k \rightarrow \infty} \eta_k = 0 \quad \text{and} \quad \alpha_k := e^{-\left(\frac{\kappa \eta_k}{\sigma_k}\right)^2} = \varepsilon_{m_k}^{-2\kappa^2} \in \ell^1(0, 1)$$

for a constant $\kappa > 0$, then the SMRE $\hat{u}_k = \hat{u}_{N(m_k)}(\alpha_k)$ almost surely satisfy (18).

Example 4.9. We consider the system of all dyadic partitions $\mathcal{P} = \mathcal{P}_2$ of $[0, 1]^d$ as in Example B.9. In particular, we note that the assumptions of Lemma B.5 are fulfilled with $n_l = (2^{d(l+1)} - 1)/(2^d - 1)$, $\delta_l = 2^{-l} \sqrt{d}$ and $\varepsilon_l = 2^{-ld/2}$.

If $p^\dagger \in \mathcal{H}_\beta([0, 1]^d)$ for $0 < \beta \leq 1$, then there exists a constant $Q = Q(p^\dagger) > 0$ such that $\omega(\delta_l, p^\dagger) \leq Q \delta_l^\beta$. This shows that

$$\frac{m+1}{\sum_{v=0}^m \omega^{-2}(\delta_v, p^\dagger)} \leq Q^2 d^\beta (2^{2\beta} - 1) \frac{m+1}{2^{2\beta(m+1)} - 1}$$

for $m \in \mathbb{N}$ large enough. From this and (26) it is easy to see, that

$$m_k + 1 \sim \frac{1}{2\beta \log 2} \log \left(\frac{Q^2 d^2 (2^{2\beta} - 1)}{d \log 2 \sigma_k^2} + 1 \right) \quad \text{and} \quad \eta_k \sim \sigma_k \sqrt{-\log \sigma_k}.$$

Thus, if there exists a constant $\kappa > 0$ such that

$$\alpha_k = e^{-\left(\frac{\kappa \eta_k}{\sigma_k}\right)^2} = \sigma_k^{\kappa^2}$$

is summable and if the true J -minimising solution u^\dagger satisfies the source condition (7) with source element $p^\dagger \in \mathcal{H}_\beta([0, 1])$, then it follows that the SMRE $\hat{u}_k = \hat{u}_{N(m_k)}(\alpha_k)$ almost surely satisfy (18) with $\eta_k = \sigma_k \sqrt{-\log \sigma_k}$.

4.3. TV-Regularisation. In this section we will study SMR-estimation for the special case where J denotes the *total-variation semi-norm* of measurable, bi-variate functions. This has a particular appeal for linear inverse problems arising in imaging (such as deconvolution), since discontinuities along curves (edges, that is) are not smoothed by minimising J .

Over the last years regularisation of (inverse) regression problems in a single space dimension invoking the total-variation semi-norm has been studied intensively and efficient numerical methods, such as the *taut-string algorithm* in [29], have been proposed (see e.g. [29, 30, 60] and references therein). In two or more space dimensions, however, the situation is much more involved and a generalisation is difficult [see e.g. 49]. We study here an extension to the case of space dimension 2 as well as to deconvolution by applying the results in Section 3 to the following setting:

We assume henceforth that $\Omega \subset \mathbb{R}^2$ is an open and bounded domain with Lipschitz-boundary $\partial\Omega$ and outer unit normal ν . Moreover, we set $U = L^2(\Omega)$ and define $\text{BV}(\Omega)$ to be the collection of $u \in U$ whose derivative Du (in the sense of distributions) is a signed \mathbb{R}^2 -valued Radon-measure with finite total-variation $|Du|$, that is

$$|Du|(\Omega) = \sup_{\substack{\psi \in C_0^1(\Omega, \mathbb{R}^2) \\ |\psi| \leq 1}} \int_{\Omega} \text{div}(\psi) u \, dx < \infty.$$

We note that the norm $\|u\|_{\text{BV}} := \|u\|_{L^1} + |Du|(\Omega)$ turns $\text{BV}(\Omega)$ into a Banach-space and that with this norm $\text{BV}(\Omega)$ is continuously embedded into $L^2(\Omega)$. The embedding is even compact if $L^2(\Omega)$ is replaced by $L^p(\Omega)$ with $p < 2$ (a proof of these embedding results can be found in [1, Thm. 2.5]. For an exhaustive treatment of $\text{BV}(\Omega)$ see [38, 81]). With this, we define

$$J(u) = \begin{cases} |Du|(\Omega) & \text{if } u \in \text{BV}(\Omega) \\ +\infty & \text{else.} \end{cases}$$

The functional J is convex and proper and, as it was shown e.g. in [1, Thm. 2.3], J is lower semi-continuous on $L^2(\Omega)$. This shows, that J satisfies Assumption 2.1 (ii). Next, we examine Assumption 3.4:

Lemma 4.10. If there exists $n_0 \in \mathbb{N}$ such that $|\langle K\mathbf{1}, \phi_{n_0} \rangle| > 0$ then Assumption 3.4 holds. Here, $\mathbf{1}$ denotes the constant 1-function on Ω .

Proof. Let $c \in \mathbb{R}$ and $\{u_k\}_{k \in \mathbb{N}} \subset \Lambda(c)$. Then in particular it follows that $\sup_{k \in \mathbb{N}} J(u_{k_n}) \leq c < \infty$ and thus we find with Poincaré's inequality [see 81, Thm. 5.11.1]

$$\|u_k - \bar{u}_k\|_{L^2} \leq c_1 J(u_k) \leq c_2 < \infty$$

for suitable constants $c_1, c_2 \in \mathbb{R}$, where $\bar{u}_k = \lambda_2(\Omega)^{-1} \int_{\Omega} u_k(\tau) d\tau$. Now choose $\phi \in \{\phi_1, \dots, \phi_N\}$ and observe that

$$\begin{aligned} \frac{|\bar{u}_k| |\langle \phi, K\mathbf{1} \rangle|}{\|\phi\|} &= \frac{|\langle \phi, K\bar{u}_k \rangle|}{\|\phi\|} \leq \frac{|\langle \phi, K(\bar{u}_k - u_k) \rangle|}{\|\phi\|} + \frac{|\langle \phi, Ku_k \rangle|}{\|\phi\|} \\ &\leq \|K\| \|u_k - \bar{u}_k\|_{L^2} + \sup_{1 \leq n \leq N} \frac{|\langle Ku_k, \phi_n \rangle|}{\|\phi_n\|} \leq \|K\| c_2 + c. \end{aligned}$$

Let $1 \leq n_0 \leq N$ be such that $|\langle K\mathbf{1}, \phi_{n_0} \rangle| =: \gamma > 0$. Then, $|\bar{u}_n| \leq (\|K\| c_2 + c) \|\phi_{n_0}\| / \gamma =: c_3$ and we find

$$\|u_n\|_{L^2} \leq (\|u_n - \bar{u}_n\|_{L^2} + \|\bar{u}_n\|_{L^2}) \leq c_2 + c_3 \lambda_2(\Omega).$$

□

We note that the assumptions in Lemma 4.10 already imply the weak compactness of the sets (8) and thus guarantee existence of a J -minimising solution of (1). From the above cited embedding properties of the space $BV(\Omega)$ it is easy to derive an improved version of the consistency result in Theorem 3.6.

Corollary 4.11. Let $g \in \overline{\text{span}\Phi}$ and assume that $u^\dagger \in BV(\Omega)$ is the unique J -minimising solution of (1). Moreover, let $\{\alpha_k\}_{k \in \mathbb{N}}$ and $\{N_k\}_{k \in \mathbb{N}}$ be as in Theorem 3.6 and define $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$. Then, additionally to the assertions in Theorem 3.6 we have that

$$\lim_{k \rightarrow \infty} \|\hat{u}_k - u^\dagger\|_{L^p} = 0 \quad \text{a.s.}$$

for every $1 \leq p < 2$.

Proof. From Theorem 3.6 it follows that $\{\hat{u}_k\}_{k \in \mathbb{N}}$ is bounded a.s. in $L^2(\Omega)$ and that each weak cluster point is a J -minimising solution of (1). Since we assumed that u^\dagger is the unique J -minimising solution of (1), it follows that $\hat{u}_k \rightharpoonup u^\dagger$ in $L^2(\Omega)$ a.s. and therefore also in $L^p(\Omega)$ for each $1 \leq p < 2$.

Since Ω is assumed to be bounded, it follows that $L^2(\Omega)$ is continuously embedded into $L^1(\Omega)$. Thus, it follows from Theorem 3.6 that almost surely $\sup_{k \in \mathbb{N}} \|\hat{u}_k\|_{BV} < \infty$. From the compact embedding $BV(\Omega) \hookrightarrow L^p(\Omega)$ for $1 \leq p < 2$, it hence follows that $\{\hat{u}_k\}_{k \in \mathbb{N}}$ is compact in $L^p(\Omega)$. Thus, the assertion follows, since weak and strong limits coincide. □

Unfortunately, the above embedding technique can not be used in order to improve the convergence rate result in Theorem 3.8 to strong L^p -convergence and thus we have to settle for the general results in Theorem 3.8. Therefore, we aim for an interpretation of convergence w.r.t. the Bregman-divergence in (18). We summarise:

Lemma 4.12. One has $\xi \in \partial J(u)$ if and only if there exists $z \in L^\infty(\Omega, \mathbb{R}^2)$ with $\|z\|_{L^\infty} \leq 1$ such that $\langle z, v \rangle = 0$ on $\partial\Omega$,

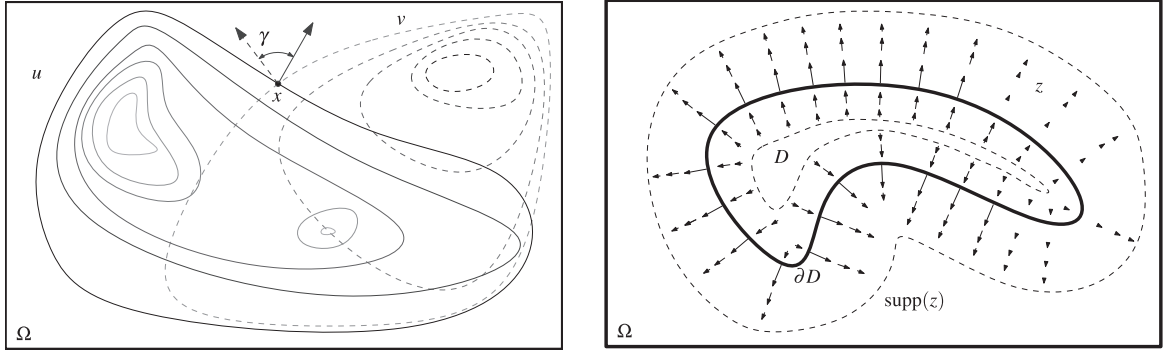
$$\text{div}(z) = \xi \quad \text{and} \quad \int_{\Omega} \xi u dx = |Du|(\Omega).$$

If $\xi \in \partial J(u)$, then $D_J^\xi(v, u) = |Dv|(\Omega) - \int_{\Omega} \xi v dx$.

Proof. Assertion (ii) directly follows from the definition of the Bregman-divergence and (i). The equivalence relation in (i) was proved e.g. in [39, Thm. 4.4.2]. □

Remark 4.2. The result in Lemma 4.12 (ii) allows a geometrical interpretation of the Bregman-divergence w.r.t. the functional J . As it was worked out in [16, Sec. 5.1], one can show that

$$D_J^\xi(v, u) = \int_{\Omega} (1 - \cos(\gamma(v, u, x))) d|Dv|(x)$$



(a) Angle $\gamma = \gamma(v, u, x)$ between the unit normals of the level lines of u (solid) and v (dashed) at a point $x \in \Omega$.

(b) Indicator function $u = \chi_D$ on a compact set D with smooth boundary ∂D and corresponding vector field z with compact support satisfying $\operatorname{div}(z) \in \partial J(u)$

FIGURE 2. TV-Regularisation.

where $\gamma(v, u, x)$ denotes the angle between the unit normals of the sub-levelsets of u and v at the point $x \in \Omega$ (cf. Figure 2(a)).

We recall that a function $u \in \operatorname{BV}(\Omega)$ satisfies the source condition, if there exists $\xi \in \operatorname{ran}(K^*)$ such that $\xi \in \partial J(u)$. It is important to note, that in many applications the elements in $\operatorname{ran}(K^*)$ exhibit high regularity such as continuity or smoothness. Thus it is of particular interest, if such regular elements in $\partial J(u)$ exist. If u is itself a smooth function, application of Green's Formula and Lemma 4.12 yield [see also 72, Lem.3.71].

Lemma 4.13. Let $u \in C_0^1(\Omega)$ and set $E[u] = \{x \in \Omega : \nabla u(x) \neq 0\}$. Assume that there exists $z \in C_0^1(\Omega, \mathbb{R}^2)$ with $|z| \leq 1$ and

$$z(x) = -\frac{\nabla u(x)}{|\nabla u(x)|} \quad \text{for } x \in E[u].$$

Then, $\xi := \operatorname{div}(z) \in \partial J(u)$.

In many applications (such as imaging) the true solution $u \in \operatorname{BV}(\Omega)$ is not continuous, as e.g. if u is the indicator function of a smooth set $D \subset \Omega$. The following examples shows that in this case we still have $\partial J(u) \cap C_0^\infty(\Omega) \neq \emptyset$. For the analytical details we refer to [72, Ex. 3.74]

Example 4.14. Assume that $D \subset \Omega$ is a closed and bounded set with C^∞ -boundary ∂D and set $u = \chi_D$. The outward unit-normal n of D then can be extended to a compactly supported C^∞ -vector field z with $|z| \leq 1$ (cf. Figure 2(b)). Independent of the choice of the extension, we then have $\xi := \operatorname{div}(z) \in \partial J(u)$ and $\xi \in C_c^\infty(\Omega)$.

Example 4.15. We consider $\Omega = [0, 1]^2$ and $V = L^2(\Omega)$. Moreover, we assume that \mathcal{P}_2 denotes the set of all dyadic partitions of Ω (cf. Example B.9) and that Φ is the collection of indicator functions w.r.t. elements in \mathcal{P}_2 .

For a function $k : \mathbb{R}^2 \rightarrow \mathbb{R}$, we consider the *convolution operator* on U defined by

$$(Ku)(x) = \int_{\mathbb{R}^2} k(x-y)\bar{u}(y) dx \quad \text{for } x \in \Omega$$

where \bar{u} denotes the extension of u on \mathbb{R}^2 by zero-padding. Assume further that u^\dagger is the indicator function on a closed and bounded set $D \subset \Omega$ with C^∞ -boundary ∂D and that $\xi \in \partial J(u^\dagger)$ is as in

Example 4.14. If the Fourier-transform $\mathcal{F}(k) =: \hat{k}$ of k is non-zero a.e. in \mathbb{R}^2 and if there exists $\beta \in (1, 2]$ such that

$$(1 + |\cdot|^2)^{-\beta/2} \left(\hat{\xi} / \hat{k} \right) \in L^2(\mathbb{R}^2) \quad \text{and} \quad \text{supp} \left(p^\dagger := \mathcal{F}^{-1} \left(\hat{\xi} / \hat{k} \right) \right) \subset \Omega,$$

then Assumption 3.7 is satisfied. To be more precise, we have that $p^\dagger \in \mathcal{H}_{\beta-1}(\Omega)$ [see 2, Thm. 7.63] and if there exists a constant $\kappa > 0$ such that $\alpha_k := \sigma_k^{2\kappa}$ is summable it follows from Example 4.9 and Lemma 4.12 that

$$\limsup_{k \rightarrow \infty} \frac{|\mathbf{D}\hat{u}_k|(\Omega) - \int_{\Omega} \hat{\xi} \hat{u}_k \, dx}{\sigma_k \sqrt{-\log \sigma_k}} = \limsup_{k \rightarrow \infty} \frac{\int_{\Omega} 1 - \cos(\gamma(\hat{u}_k, u^\dagger, x)) \, d|\mathbf{D}\hat{u}_k|(x)}{\sigma_k \sqrt{-\log \sigma_k}} < \infty \quad \text{a.s.}$$

for the SMRE $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$ (where N_k is as in Example 4.9).

ACKNOWLEDGEMENT

K.F. is supported by the DFG-SNF Research Group FOR916 *Statistical Regularization* (Z-Project). P.M. and A.M. are supported by the BMBF project 03MUPAH6 *INVERS* and by the SFB755 *Photonic Imaging on the Nanoscale*. A.M. is supported by the SFB803 *Functionality Controlled by Organization in and between Membranes*. The authors would like to thank S. Hell, A. Egner and A. Schoenle (Department of NanoBiophotonics, Max Planck Institute for Biophysical Chemistry, Göttingen) for providing the microscopy data and L. Dümbgen (University of Bern) for stimulating discussions.

APPENDIX A. SOURCE-CONDITION AND BREGMAN-DIVERGENCE: SOME EXAMPLES

The notions of source-condition and Bregman-divergence are very common in the field of inverse problems but are rather seldom found in the statistical literature. For this reason, we will summarise the meaning of the source-condition (7) and the Bregman-divergence for some frequently used regularisation functionals J . We also note that in Section 4.3 the more complex example where J is the total-variation of a measurable function on a domain Ω is studied in more detail.

Example A.1. Let $J(u) = \frac{1}{2} \|u\|^2$. Then, J is differentiable on U and for all $u \in U$ the set $\partial J(u)$ consists of the single element $\{u\}$. We have that $J'(v)(w) = \langle v, w \rangle$ and consequently

$$D_J(v, u) = D_J^\xi(v, u) = \frac{1}{2} \|v - u\|^2 \quad \text{for } \xi = u \in \partial J(u).$$

Moreover, the source condition (7) can be rewritten to

$$u^\dagger \in \text{ran}(K^*).$$

Since $\text{ran}(K^*) = \text{ran}(K^*K)^{1/2}$, this shows that the source condition (7) corresponds to the *Hölder-source condition* $u^\dagger \in \text{ran}(K^*K)^\beta$ for $\beta = 1/2$ [see 37]. In [7, Sec. 5.3], the Hölder-source condition w.r.t. a *smoothing operator* K on Hilbert-scales has been discussed. To be more precise, assume that $\{H_\mu\}_{\mu \in \mathbb{R}}$ is a scale of Hilbert spaces and that K is a -times smoothing, i.e. $K : H_{\mu-a} \rightarrow H_\mu$ is continuous with continuous inverse. Then the condition $u^\dagger = (K^*K)^\beta p^\dagger$ implies that $u^\dagger \in H_{2a\beta}$. A prototype for Hilbert scales are Sobolev spaces. Here the index μ corresponds to the Sobolev index.

Example A.2. Let $\{\psi_n\}_{n \in \mathbb{N}}$ be a ONB of U and define

$$J(u) = \|u\|_1 := \sum_{j \in \mathbb{N}} |\langle u, \psi_n \rangle|.$$

In applications this functional promotes *sparse solutions*, that is solutions that have only few non-zero coefficients w.r.t the basis $\{\psi_n\}_{n \in \mathbb{N}}$. As it was argued in [45, Rem. 17] the source-condition (7) holds if and only if there exist constants $a, b, \gamma > 0$ such that $\|u^\dagger\|_1 < a$ and

$$\|u\|_1 - \|u^\dagger\|_1 \geq -\gamma \|K(u - u^\dagger)\|$$

for all $u \in U$ such that $\|u\|_1 < a$ and $\|K(u - u^\dagger)\| < b$. If additionally for every finite set $J \subset \mathbb{N}$ the restriction of K to the set $\{\psi_n : n \in J\}$ is injective, there exist constants $\beta_1, \beta_2 > 0$ such that

$$\|u - u^\dagger\|_1 \leq \beta_1 D_J^{K^* p^\dagger}(u, u^\dagger) + \beta_2 \|K(u - u^\dagger)\|$$

for all $u \in U$ (see the proof of [45, Thm. 15] and [40, Thm 6.4]).

Example A.3. Assume that $U = L^2(\Omega)$ for an open and bounded set $\Omega \subset \mathbb{R}^n$ with Lipschitz boundary $\partial\Omega$ and outer unit-normal ν and let $H^\beta(\Omega)$ denote the Sobolev-space of order $\beta \in \mathbb{R}$. We define

$$J(u) = \begin{cases} \int_\Omega |\nabla u|^2 \, dx & \text{if } u \in H^1(\Omega) \\ +\infty & \text{else.} \end{cases}$$

Then [see 3, pp.63], the set $D(\partial J)$ consists of all elements $u \in H^2(\Omega)$ that have vanishing normal derivative $\langle \nabla u, \nu \rangle$ on $\partial\Omega$ and if $u \in D(\partial J)$, then $\partial J(u) = \{-\Delta u\}$. With this, it follows that $J'(v)(w) = \langle \nabla v, \nabla w \rangle$ and

$$D_J(v, u) = D_J^\xi(v, u) = \frac{1}{2} \|\nabla(v - u)\|^2 \quad \text{for } \xi = -\Delta u \in \partial J(u).$$

Moreover, u^\dagger satisfies the source condition (7) with source element $p^\dagger \in V$ if and only if

$$\begin{aligned} -(K^* p^\dagger)(x) &= \Delta u^\dagger(x) \quad \text{in } \Omega \\ \nabla u^\dagger \cdot \nu &= 0 \quad \mathcal{H}^{n-1}\text{-a.e. on } \partial\Omega \end{aligned}$$

(here \mathcal{H}^{n-1} stands for the $(n-1)$ -dimensional Hausdorff-measure on $\partial\Omega$).

Example A.4. Let U be as in Example A.3 and define the *negentropy* by

$$J(u) = \begin{cases} -\int_\Omega u \log u \, dx & \text{if } u \geq 0 \text{ a.e. and } u \log u \in L^1(\Omega) \\ +\infty & \text{else.} \end{cases}$$

Then [see 4, Chap. 2 Prop 2.7], the set $D(\partial J)$ consists of all non-negative functions in $L^\infty(\Omega)$ that are bounded away from zero. One has $J'(v)(w) = \langle 1 + \log v, w \rangle$ and if $u \in D(\partial J)$, then $\partial J(u) = \{1 + \log u\}$. After some re-arrangements we find

$$D_J(v, u) = D_J^\xi(v, u) = \int_\Omega \left(v \log \left(\frac{v}{u} \right) - v + u \right) dx,$$

that is, the Bregman-divergence coincides in this particular case with the *Kullback-Leiber-divergence*. It was proved in [11, Lem. 2.2] that

$$\|v - u\|_{L^1}^2 \leq \left(\frac{2}{3} \|v\|_{L^1} + \frac{4}{3} \|u\|_{L^1} \right) D_J(v, u).$$

In other words, Bregman-consistency (or convergence rates) w.r.t. the negentropy yields strong consistency (convergence rates) in $L^1(\Omega)$. Finally, we note that $u^\dagger \in D(\partial J)$ satisfies the source condition (7) with source element $p^\dagger \in V$ if and only if

$$e^{(K^* p^\dagger)(x)-1} = u^\dagger(x) \quad \text{for a.e. } x \in \Omega.$$

APPENDIX B. PROOFS

B.1. Proofs of the main results. In this section the proofs of the main results, that is existence, consistency and convergence rates for SMRE, are collected. We start with a basic estimate for the quantile function $q_N(\cdot)$ of the MR-statistic as defined in (11). We shall assume that Assumptions 2.1 and 3.4 hold.

Lemma B.1. Assume that T_N is an MR-statistic and let $\alpha \in (0, 1)$ and $N \in \mathbb{N}$. Then,

$$q_N(\alpha) \leq \text{med}(T_N(\varepsilon)) + L\sqrt{-2\log(2\alpha)}.$$

Proof. First, we introduce the function $f(x_1, \dots, x_N) = \sup_{1 \leq n \leq N} t_N(x_n, \|\phi_n\|)$. Then, f is Lipschitz continuous with $\|f\|_{\text{Lip}} \leq L$. Next, define for $1 \leq n \leq N$ the random variables $\varepsilon_n := \varepsilon(\phi_n^*)$. Then, $(\varepsilon_1, \dots, \varepsilon_N) \sim \mathcal{N}(0, \Sigma)$ for a symmetric and positive matrix $\Sigma \in \mathbb{R}^{N \times N}$ with $\|\Sigma\|_2 = 1$. Hence

$$T_N(\varepsilon) = \sup_{1 \leq n \leq N} t_N(\varepsilon(\phi_n^*), \|\phi_n\|) = f(\varepsilon_1, \dots, \varepsilon_N) = f(\Sigma^{1/2}Z),$$

where Z is an N -dimensional random vector with independent standard normal components. In other words, the statistic $T_N(\varepsilon)$ can be written as the image of Z under the Lipschitz function $f(\Sigma^{1/2}\cdot)$. Applying Borel's inequality [see 76, Lem. A.2.2] we find that $2\mathbb{P}(T_N(\varepsilon) - \text{med}(T_N(\varepsilon)) > L\eta) \leq \exp(-\eta^2/2)$ for all $\eta \in \mathbb{R}$. Now let $\alpha \in (0, 1)$ and set $\eta = q_N(\alpha)L$. Then,

$$\alpha \leq \mathbb{P}(T_N(\varepsilon) > q_N(\alpha)) \leq \frac{1}{2} \exp\left(-\frac{1}{2} \left(\frac{q_N(\alpha) - \text{med}(T_N(\varepsilon))}{L}\right)^2\right).$$

Rearranging the above inequality yields the desired estimate. \square

We proceed with the proof of the existence result in Theorem 3.5. To this end we use a standard compactness argument from convex optimisation. For the sake of completeness, however, we will present the proof.

Proof of Theorem 3.5. Let $N \geq N_0$ and $y \in V$ be arbitrary. Due to Assumption 2.1 (ii), $D(J) \subset U$ is dense and hence there exists for all given $\delta > 0$ an element $u_0 \in \overline{D(J)}$ such that $\|Ku_0 - \tilde{y}\| \leq \delta$, where \tilde{y} denotes the orthonormal projection of y onto $\text{ran}(K)$. Since $\phi_n \in \text{ran}(K)$ and $\|\phi_n^*\| = 1$ for all $n \in \mathbb{N}$, this implies that $|\langle Ku_0 - y, \phi_n^* \rangle| = |\langle Ku_0 - \tilde{y}, \phi_n^* \rangle| \leq \delta$ for all $n \in \mathbb{N}$.

Now let $\sigma > 0$ and $\alpha \in (0, 1)$. Since T_N is an MR-statistic (cf. Definition 3.1) we find that $t_N(0, r) < 0$ for all $r \in (0, 1]$. Thus, according to the reasoning above, there exists $u_0 \in D(J)$ such that for $1 \leq n \leq N$

$$(27) \quad L\sigma^{-1} |y_n - \langle Ku_0, \phi_n^* \rangle| \leq q_N(\alpha) - \sup_{1 \leq n \leq N} \lambda_N(\|\phi_n\|),$$

if the right-hand side is positive. To see this, assume that $q_N(\alpha) \leq \sup_{1 \leq n \leq N} \lambda_N(\|\phi_n\|)$. Since for $1 \leq n \leq N$ we have that $t_N(|\varepsilon(\phi_n^*)|, \|\phi_n\|) \geq \lambda_N(\|\phi_n\|)$ almost surely according to (10), it then follows that

$$\mathbb{P}(T_N(\varepsilon) \geq q_N(\alpha)) \geq \mathbb{P}\left(T_N(\varepsilon) \geq \sup_{1 \leq n \leq N} \lambda_N(\|\phi_n\|)\right) = 1.$$

This is a contradiction to the definition of $q_N(\alpha)$ in (11) and thus $u_0 \in D(J)$ as in (27) can be chosen. Since $s \mapsto t_N(s, r)$ is Lipschitz-continuous with constant L and increasing for all $r \in (0, 1]$, we find

$t_N(\sigma^{-1}|y_n - \langle Ku_0, \phi_n^* \rangle|, \|\phi_n\|) \leq t_N(0, \|\phi_n\|) + L\sigma^{-1}|y_n - \langle Ku_0, \phi_n^* \rangle| \leq q_N(\alpha)$ for $1 \leq n \leq N$. In other words, there exists at least one element $u_0 \in D(J)$ such that

$$u_0 \in S := \left\{ u \in U : \sup_{1 \leq n \leq N} t_N(\sigma^{-1}|y_n - \langle Ku, \phi_n^* \rangle|, \|\phi_n\|) \leq q_N(\alpha) \right\}.$$

Now, choose a sequence $\{u_k\}_{k \in \mathbb{N}} \subset S$ such that $J(u_k) \rightarrow \inf_{u \in S} J(u)$. This shows that $\sup_{k \in \mathbb{N}} J(u_k) =: a < \infty$. Moreover, we find from (10), that there exist constants $c_1, c_2 > 0$ such that for all $1 \leq n \leq N$

$$\begin{aligned} c_1 \sigma^{-1} |y_n - \langle Ku_k, \phi_n^* \rangle| + c_2 t_N(|y_n - \langle Ku_k, \phi_n^* \rangle|, \|\phi_n\|) \\ \leq t_N(\sigma^{-1} |y_n - \langle Ku_k, \phi_n^* \rangle|, \|\phi_n\|) \leq q_N(\alpha). \end{aligned}$$

Together with (9), this shows $c_1 \sigma^{-1} |y_n - \langle Ku_k, \phi_n^* \rangle| + c_2 \lambda_N(\|\phi_n\|) \leq q_N(\alpha)$. Rearranging the inequality above yields

$$\sup_{1 \leq n \leq N} |\langle Ku_k, \phi_n^* \rangle| \leq \sup_{1 \leq n \leq N} |y_n| + \frac{\sigma}{c_1} \left(q_N(\alpha) - c_2 \inf_{1 \leq n \leq N} \lambda_N(\|\phi_n\|) \right) =: b < \infty.$$

Summarising, we find that $u_k \in \Lambda(a + b)$ for all $k \in \mathbb{N}$, as a consequence of which we can drop a weakly convergent sub-sequence (indexed by $\rho(k)$ say) with weak limit \hat{u} . Since we assumed that $t_N(\cdot, r)$ is convex for all $r \in (0, 1]$, it follows that the admissible region S is convex and closed and therefore weakly closed. This shows that $\hat{u} \in S$. Moreover, the weak lower semi-continuity of J (cf. Assumption 2.1 (ii)) implies

$$J(\hat{u}) \leq \liminf_{k \rightarrow \infty} J(u_{\rho(k)}) = \inf_{u \in S} J(u)$$

and the assertion follows with $\hat{u}_N(\alpha) = \hat{u}$ □

In order to prove Bregman-consistency of SMR-estimation in Theorem 3.6, we first establish a basic estimate for the data error.

Lemma B.2. Let $N \geq N_0$ and $\alpha \in (0, 1)$. Moreover, assume that u^\dagger is a solution of (1) and that $\hat{u}_N(\alpha)$ is an SMRE. Then, for $1 \leq n \leq N$

$$c_1 \sigma^{-1} |\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^* \rangle| \leq T_N(\varepsilon) - 2c_2 \lambda_N(\|\phi_n\|) + \text{med}(T_N(\varepsilon)) + L\sqrt{-2\log(2\alpha)}.$$

Proof. From Definition 3.3 it follows that $t_N(\sigma^{-1} |\langle Ku^\dagger - K\hat{u}_N(\alpha) + \sigma\varepsilon, \phi_n^* \rangle|, \|\phi_n\|) \leq q_N(\alpha)$ for $1 \leq n \leq N$. The convexity of t_N hence implies that

$$\begin{aligned} t_N((2\sigma)^{-1} |\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^* \rangle|, \|\phi_n\|) \\ \leq \frac{1}{2} (t_N(\sigma^{-1} |\langle Y - K\hat{u}_N(\alpha), \phi_n^* \rangle|, \|\phi_n\|) + t_N(|\varepsilon(\phi_n^*)|, \|\phi_n\|)) \leq \frac{1}{2} (q_N(\alpha) + T_N(\varepsilon)). \end{aligned}$$

By setting $v = (2\sigma)^{-1} |\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^* \rangle|$ and $r = \|\phi_n\|$ in (10), the above estimate shows that

$$c_1 (2\sigma)^{-1} |\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^* \rangle| + c_2 t_N \left(\frac{1}{2} |\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^* \rangle|, \|\phi_n^*\| \right) \leq \frac{q_N(\alpha) + T_N(\varepsilon)}{2}.$$

Since $t_N(v, r) \geq \lambda_N(r)$ for all $v \in \mathbb{R}^+$ and $r \in (0, 1]$ (cf. (9)) this implies $c_1 \sigma^{-1} |\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^* \rangle| \leq q_N(\alpha) + T_N(\varepsilon) - 2c_2 \lambda_N(\|\phi_n\|)$ for $1 \leq n \leq N$. Finally, the assertion follows from Lemma B.1. □

With these preparations, we are now able to prove Bregman-consistency.

Proof of Theorem 3.6. By the definition of the SMRE $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$, it follows that

$$\mathbb{P}(J(\hat{u}_k) > J(u^\dagger)) \leq \mathbb{P}(T_{N_k}(\sigma_k^{-1}(Y - Ku^\dagger)) > q_{N_k}(\alpha_k)) = \mathbb{P}(T_{N_k}(\varepsilon) > q_{N_k}(\alpha_k)) \leq \alpha_k$$

for all $k \in \mathbb{N}$. Since $\sum_{k=1}^{\infty} \alpha_k < \infty$, it follows from the Borel-Cantelli Lemma [see 73, p 255] that $\mathbb{P}(J(\hat{u}_k) > J(u^\dagger) \text{ i.o.}) \leq \mathbb{P}(T_{N_k}(\varepsilon) > q_{N_k}(\alpha_k) \text{ i.o.}) = 0$, or in other words

$$(28) \quad \mathbb{P}(\exists k_0 \in \mathbb{N} : J(\hat{u}_k) \leq J(u^\dagger) \text{ for all } k \geq k_0) = 1.$$

In particular, it follows that $\sup_{k \in \mathbb{N}} J(\hat{u}_k) =: a < \infty$ a.s.

Next, we note that $\sup_{N \in \mathbb{N}} T_N(\varepsilon) < \infty$ a.s. implies that $\sup_{N \in \mathbb{N}} \text{med}(T_N(\varepsilon)) < \infty$. Hence, it follows from Lemma B.2 and (14) that $\sup_{1 \leq n \leq N_k} |\langle Ku^\dagger - K\hat{u}_k, \phi_n^* \rangle| = \mathcal{O}(\zeta_k)$ almost surely. as $k \rightarrow \infty$ which proves (16). In particular, (16) and the fact that $N_k > N_0$ imply $\sup_{k \in \mathbb{N}} \sup_{1 \leq n \leq N_0} |\langle K\hat{u}_k, \phi_n^* \rangle| =: b < \infty$ a.s. Summarising, we find that $\hat{u}_k \in \Lambda(a + b)$ which is sequentially weakly precompact according to Assumption 3.4 (ii). Choose a sub-sequence indexed by $\rho(k)$ with weak limit $\hat{u} \in U$. Since $N_k \rightarrow \infty$ as $k \rightarrow \infty$ it follows from (16) and (14) that

$$|\langle g - K\hat{u}, \phi_n^* \rangle| = \lim_{k \rightarrow \infty} |\langle Ku^\dagger - K\hat{u}_{\rho(k)}, \phi_n^* \rangle| = 0 \quad \text{for all } n \in \mathbb{N}.$$

Since we assumed that $g \in \overline{\text{span}\Phi}$ this shows that $K\hat{u} = g$. Furthermore, according to (28) there exists (almost surely) an index k_0 such that $J(\hat{u}_{\rho(k)})$ does not exceed $J(u^\dagger)$ for all $k \geq k_0$. Together with the weak lower semi-continuity of J this shows $J(\hat{u}) \leq \liminf_{k \rightarrow \infty} J(\hat{u}_{\rho(k)}) \leq \limsup_{k \rightarrow \infty} J(\hat{u}_{\rho(k)}) \leq J(u^\dagger)$. Since u^\dagger is a J -minimising solution of (1) we conclude that the same holds for \hat{u} and that $J(\hat{u}) = J(u^\dagger) = \lim_{k \rightarrow \infty} J(\hat{u}_{\rho(k)})$. In particular, for each sub-sequence $\{J(u_k)\}_{k \in \mathbb{N}}$ there exists a further sub-sequence that converges to $J(u^\dagger)$. This already shows that $\lim_{k \rightarrow \infty} J(\hat{u}_k) = J(u^\dagger)$ a.s.

We next prove that $D_J(u^\dagger, \hat{u}_k) \rightarrow 0$. To this end, recall that there almost surely exists an index k_0 such that for $k \geq k_0$ one has $T_{N_k}(\varepsilon) \leq q_{N_k}(\alpha_k)$. In order to exploit strong duality arguments, however, we have to make sure that the interior of the admissible region is non-empty (Slater's constraint qualification). But since we assumed that $s \mapsto t_N(s, r)$ is (strictly) increasing for each fixed $r \in (0, 1]$ it follows that $\mathbb{P}(t_{N_k}(|\varepsilon(\phi_n^*)|, \|\phi_n^*\|) = q_{N_k}(\alpha_k)) = 0$ for all $n \in \mathbb{N}$ and thus

$$(29) \quad \mathbb{P}(\exists k_0 : T_{N_k}(\varepsilon) < q_{N_k}(\alpha_k) \text{ for all } k \geq k_0) = 1.$$

By introducing the functional

$$G_k(v) = \begin{cases} 0 & \text{if } T_{N_k}(\sigma_k^{-1}(Y - v)) \leq q_{N_k}(\alpha_k) \\ +\infty & \text{else,} \end{cases}$$

we can rewrite (12) into $\hat{u}_k \in \text{argmin}_{u \in U} J(u) + G_k(Ku)$. From (29) it follows that u^\dagger lies in the interior of the admissible set of the convex problem (12). In other words, the functionals G_k are continuous at Ku^\dagger for k large enough. Therefore we can apply [36, Chap. II Prop. 4.1] (cf. also Chapter II, Remark 4.2 therein) and choose an element $\xi_k \in V$ such that $K^* \xi_k \in \partial J(\hat{u}_k)$ and $-\xi_k \in \partial G_k(K\hat{u}_k)$. The second inclusion and the definition of the sub-gradient show that $G_k(Ku) \geq G_k(\hat{u}_k) - \langle \xi_k, Ku - K\hat{u}_k \rangle = \langle K^* \xi_k, \hat{u}_k - u \rangle$ for all $u \in U$. In particular, u^\dagger satisfies $T_{N_k}(\sigma_k^{-1}(Y - Ku^\dagger)) = T_{N_k}(\varepsilon) < q_{N_k}(\alpha_k)$ and thus $G_k(Ku^\dagger) = 0$. This shows $0 \geq \langle K^* \xi_k, \hat{u}_k - u^\dagger \rangle$. Since $J(\hat{u}_k) \rightarrow J(u^\dagger)$ we find

$$\begin{aligned} 0 &\leq \limsup_{k \rightarrow \infty} D_J(u^\dagger, \hat{u}_k) \leq \limsup_{k \rightarrow \infty} D_J^{K^* \xi_k}(u^\dagger, \hat{u}_k) \\ &= \limsup_{k \rightarrow \infty} J(u^\dagger) - J(\hat{u}_k) - \langle K^* \xi_k, u^\dagger - \hat{u}_k \rangle \leq \limsup_{k \rightarrow \infty} J(u^\dagger) - J(\hat{u}_k) = 0. \end{aligned}$$

This proves (15). □

It remains to prove the convergence rate results in Theorem 3.8. To this end additional regularity of the true J -minimising solutions u^\dagger of (1) has to be taken into account. This is formulated in Assumption 3.7. With this we get the following basic estimate.

Lemma B.3. Assume that Assumption 3.7 holds and let $N \geq N_0$ and $\alpha \in (0, 1)$. Then,

$$\begin{aligned} |\langle K^* p^\dagger, \hat{u}_N(\alpha) - u^\dagger \rangle| &\leq \frac{\sigma}{c_1} \left(\tilde{T}_N(\varepsilon) - 2c_2 \inf_{1 \leq n \leq N} \lambda_N(\|\phi_n\|) + L\sqrt{-2\log(2\alpha)} \right) \sum_{n=1}^N |b_{n,N}| \\ &\quad + \rho_N \|K\hat{u}_N(\alpha) - Ku^\dagger\|, \end{aligned}$$

where $\tilde{T}_N(\varepsilon) = T_N(\varepsilon) + \text{med}(T_N(\varepsilon))$.

Proof. From Assumption 3.7 we find that

$$\begin{aligned} |\langle K^* p^\dagger, \hat{u}_N(\alpha) - u^\dagger \rangle| &= |\langle p^\dagger, K\hat{u}_N(\alpha) - Ku^\dagger \rangle| \\ &\leq \left| \left\langle \sum_{n=1}^N b_{n,N} \phi_n^*, K\hat{u}_N(\alpha) - Ku^\dagger \right\rangle \right| + \rho_N \|K\hat{u}_N(\alpha) - Ku^\dagger\| \\ &\leq \sum_{n=1}^N |b_{n,N}| \sup_{1 \leq n \leq N} |\langle \phi_n^*, K\hat{u}_N(\alpha) - Ku^\dagger \rangle| + \rho_N \|K\hat{u}_N(\alpha) - Ku^\dagger\|. \end{aligned}$$

From Lemma B.2 it follows that

$$\sup_{1 \leq n \leq N} |\langle \phi_n^*, K\hat{u}_N(\alpha) - Ku^\dagger \rangle| \leq \frac{\sigma}{c_1} \left(\tilde{T}_N(\varepsilon) - 2c_2 \inf_{1 \leq n \leq N} \lambda_N(\|\phi_n\|) + L\sqrt{-2\log(2\alpha)} \right)$$

which shows the assertion. \square

Combination of the auxiliary result in Lemma B.3 with Theorem 3.6 paves the way to the proof of Theorem 3.8.

Proof of Theorem 3.8. First, observe that Assumption 3.7 and the definition of η_k imply (14), that is, all assumptions in Theorem 3.6 are satisfied. Therefore $\{\hat{u}_k\}_{k \in \mathbb{N}}$ is bounded almost surely and due to the continuity of K we find that $\sup_{k \in \mathbb{N}} \|K\hat{u}_k - Ku^\dagger\| < \infty$ a.s. After setting $B := \sup_{N \in \mathbb{N}} \sum_{n=1}^N |b_{n,N}|$, which is finite according to Assumption 3.7, it follows from Lemma B.3 and the definition of η_k that

$$(30) \quad |\langle K^* p^\dagger, \hat{u}_k - u^\dagger \rangle| \leq \frac{B\sigma_k}{c_1} \tilde{T}_{N_k}(\varepsilon) + C\eta_k$$

for a suitably chosen constant $C > 0$. Since $\sup_{N \in \mathbb{N}} T_N(\varepsilon) < \infty$ almost surely, it follows that also $\sup_{N \in \mathbb{N}} \tilde{T}_N(\varepsilon) = \sup_{N \in \mathbb{N}} (T_N(\varepsilon) + \text{med}(T_N(\varepsilon))) < \infty$ a.s. Combining this with (30) shows

$$|\langle K^* p^\dagger, \hat{u}_k - u^\dagger \rangle| = \mathcal{O}(\eta_k) \quad \text{a.s.}$$

Next, recall from (28) in the proof of Theorem 3.6 that almost surely an index k_0 can be chosen such that for all $k \geq k_0$ one has $J(\hat{u}_k) \leq J(u^\dagger)$. This shows that

$$D_J^{K^* p^\dagger}(\hat{u}_k, u^\dagger) = J(\hat{u}_k) - J(u^\dagger) - \langle K^* p^\dagger, \hat{u}_k - u^\dagger \rangle \leq |\langle K^* p^\dagger, \hat{u}_k - u^\dagger \rangle| = \mathcal{O}(\eta_k)$$

for $k \geq k_0$. This proves the first estimate in (18). The second estimate follows directly from Lemma B.2. \square

B.2. Approximation of continuous functions and entropy estimates. In this section we collect some results on the approximation properties and entropy estimates for systems of piecewise constant functions defined on a convex and compact set $X \subset \mathbb{R}^d$ ($d \geq 1$). We start with the following basic

Definition B.4. Let $X \subset \mathbb{R}^d$ be compact and convex.

(i) For a function $g : X \rightarrow \mathbb{R}$, the *modulus of continuity* is defined by

$$\omega(\delta, g) = \sup_{\substack{s, t \in X \\ \|s-t\|_2 \leq \delta}} |g(s) - g(t)| \quad \text{for } \delta > 0.$$

(ii) A function $g : X \rightarrow \mathbb{R}$ is called *Hölder-continuous with exponent* $\beta \in (0, 1]$ if $\omega(\delta, g) = \mathcal{O}(\delta^\beta)$. The collection of all functions on X that are Hölder-continuous with exponent β is denoted by $\mathcal{H}_\beta(X)$.

The following lemma provides an error estimate for the approximation of a continuous $g : X \subset \mathbb{R}^d \rightarrow \mathbb{R}$ by piecewise constant functions in terms of the modulus of continuity.

Lemma B.5. Let $X \subset \mathbb{R}^d$ be a compact and convex set and $\{A_1, A_2, \dots\}$ be a collection of measurable sub-sets of X . Assume that there exists an increasing sequence $\{n_l\}_{l \in \mathbb{N}} \subset \mathbb{N}$ with $n_0 = 0$ such that

- (i) for all $n_l + 1 \leq i < j \leq n_{l+1}$ one has $\lambda_d(A_i \cap A_j) = 0$,
- (ii) and $X = A_{n_l+1} \cup \dots \cup A_{n_{l+1}}$

for all $l \in \mathbb{N}$ (λ_d denotes the d -dimensional Lebesgue measure). Then, for all continuous $g : X \rightarrow \mathbb{R}$ there exist coefficients $b_{j,l}^m$ such that

$$\sup_{m \in \mathbb{N}} \sum_{l=0}^m \sum_{j=n_l+1}^{n_{l+1}} |b_{j,l}^m| \leq \|g\|_\infty \quad \text{and} \quad \left\| g - \sum_{l=0}^m \sum_{j=n_l+1}^{n_{l+1}} b_{j,l}^m \chi_{A_j} \right\|^2 \leq \frac{m+1}{\sum_{l=0}^m \omega^{-2}(\delta_l, g)},$$

where $\delta_l := \max_{n_l < j \leq n_{l+1}} \text{diam}(A_j)$.

Proof. Let $g : X \rightarrow \mathbb{R}$ be continuous. For $l \in \mathbb{N}$ we define

$$g_l = \sum_{j=n_l+1}^{n_{l+1}} \lambda_d(A_j)^{-1} \int_{A_j} g(\tau) \, d\tau \cdot \chi_{A_j}.$$

Next, we introduce $a_{lm} = (\omega^{-2}(\delta_l, g)) / (\sum_{v=0}^m \omega^{-2}(\delta_v, g))$ for $m \in \mathbb{N}$ and $1 \leq l \leq m$. Note, that $a_{lm} \in (0, 1)$ and $\sum_{0 \leq l \leq m} a_{lm} = 1$. With this, we define for $0 \leq l \leq m$ and $n_l < j \leq n_{l+1}$ the coefficients $b_{j,l}^m = (a_{lm} \int_{A_j} g(\tau) \, d\lambda_d(\tau)) / \lambda_d(A_j)$. Since we assumed that g is continuous on the compact set X , it follows that $|b_{j,l}^m| \leq \|g\|_\infty a_{lm}$ and hence $\sum_{l=0}^m \sum_{j=n_l+1}^{n_{l+1}} |b_{j,l}^m| \leq \|g\|_\infty$ for all $m \in \mathbb{N}$. Moreover, we have for all $s \in X$ that

$$\left| \sum_{l=0}^m a_{lm} g_l(s) - g(s) \right| \leq \sum_{l=0}^m a_{lm} \left(\sum_{j=n_l+1}^{n_{l+1}} \frac{1}{\lambda_d(A_j)} \int_{A_j} |g(\tau) - g(s)| \, d\tau \cdot \chi_{A_j}(s) \right).$$

After applying Jensen's inequality and keeping in mind that $|s - t| \leq \delta_l$ for $s, t \in A_j$ and $n_l < j \leq n_{l+1}$ it follows that

$$\begin{aligned} \int_X \left| \sum_{l=0}^m a_{lm} g_l(s) - g(s) \right|^2 ds &\leq \sum_{l=0}^m a_{lm} \int_X \left(\sum_{j=n_l+1}^{n_{l+1}} \frac{1}{\lambda_d(A_j)} \int_{A_j} |g(\tau) - g(s)|^2 d\tau \cdot \chi_{A_j}(s) \right) ds \\ &= \sum_{l=0}^m a_{lm} \sum_{j=n_l+1}^{n_{l+1}} \int_{A_j} \frac{1}{\lambda_d(A_j)} \int_{A_j} |g(\tau) - g(s)|^2 d\tau ds \\ &\leq \sum_{l=0}^m a_{lm} \omega^2(\delta_l, g) \sum_{j=n_l+1}^{n_{l+1}} \lambda_d(A_j). \end{aligned}$$

Assumptions (i) and (ii) together with the definition of the coefficients a_{lm} eventually yield

$$\int_X \left| \sum_{l=0}^m a_{lm} g_l(s) - g(s) \right|^2 ds \leq \frac{m+1}{\sum_{v=0}^m \omega^{-2}(\delta_v, g)}.$$

□

For the remainder of this section we collect some results concerning the capacity number of (subsystems of) the set Φ_d of indicator functions on convex and closed sets in $[0, 1]^d$ with $d \geq 1$. We first recall the basic definition

Definition B.6. Let (T, d) be a semi-metric space, $T' \subset T$ and $\varepsilon > 0$. The *capacity number* is defined by

$$D(\varepsilon, T') := \sup_{T'' \subset T'} (\#\{T'' : d(a, b) \geq \varepsilon \text{ for all } a \neq b \in T''\}).$$

From a practical point of view, it is often more convenient to express (25) in terms of the ε -covering number $N(\varepsilon, T')$ of T' which is defined as the smallest number of ε -balls in T needed to cover T' (the center points need not to be elements of T' , though). It is common knowledge [see 76, p.98] that for all $\varepsilon > 0$

$$(31) \quad N(\varepsilon, T) \leq D(\varepsilon, T) \leq N(\varepsilon/2, T).$$

We consider $\Phi_d \subset L^2([0, 1]^d)$ as a metric space with the induced L^2 -metric, i.e. for $\chi_P, \chi_Q \in \Phi_d$ we have

$$d(\chi_Q, \chi_P)^2 = \|\chi_P - \chi_Q\|^2 = \int_{[0, 1]^d} (\chi_Q - \chi_P)^2 d\lambda_d = \lambda_d(Q \Delta P).$$

The entire set Φ_d is too large in order to render the test-statistic T_N in (24) finite: it was shown in [13] [see also 33, Chap. 8.4]) that the ε -covering number of Φ_d of all nonempty, closed and convex sets contained in the unit ball $\{x \in \mathbb{R}^d : |x| \leq 1\}$ is of the same order as $\exp(\varepsilon^{(1-d)/2})$ (for $d \geq 2$) as $\varepsilon \rightarrow 0^+$. This proves that there cannot exist any constants A, B and γ such that (25) holds with $\Phi = \Phi_d$.

For particular classes of convex sets, however, entropy estimates as in (25) are at hand. The collection Φ_r of indicator functions on d -dimensional rectangles in $[0, 1]^d$ constitutes such an example:

Proposition B.7. *There exists a constant $A = A(d) > 0$ such that*

$$D(u\delta, \{\phi \in \Phi_r : \|\phi\| \leq \delta\}) \leq A(u\delta)^{-4d}$$

for all $u, \delta \in (0, 1]$.

Proof. From [76, Thm. 2.6.7] it follows that the ε -covering number of Φ_r can be estimated by $A\varepsilon^{-2(V-1)}$ where V denotes the VC-index of the set of subgraphs $\{(x, t) : t < \phi(x)\}$ for $\phi \in \Phi_r$. This in turn is equal to the VC-index of the collections of all rectangles in $[0, 1]^d$ which is $2d + 1$ [see 76, Ex. 2.6.1]. \square

For certain subsets of Φ_r better estimates can be derived. We close this section with results for the system Φ_s and Φ_2 of indicator functions on all squares and dyadic partitions in $[0, 1]^d$ respectively. We skip the proofs, for they are elementary but rather tedious.

Proposition B.8. *There exists a constant $A = A(d) > 0$ such that*

$$D(u\delta, \{\phi \in \Phi_s : \|\phi\| \leq \delta\}) \leq Au^{-2(d+1)}\delta^{-d}, \quad \text{for all } u, \delta \in (0, 1].$$

Proposition B.9. *Let $d \geq 2$ and consider the system of all dyadic partitions in $[0, 1]^d$, that is*

$$\mathcal{P}_2 := \left\{ Q \subset [0, 1]^d : Q = 2^{-k}(i + [0, 1]^d), k \in \mathbb{N}, i = (i_1, \dots, i_d) \in \mathbb{N}^d \right\}.$$

Let Φ_2 the set of all indicator functions on elements in \mathcal{P}_2 . Then, there exists a constant $A = A(d) > 0$ such that

$$A^{-1}u^{-2}\delta^{-2} \leq D(u\delta, \{\phi \in \Phi_2 : \|\phi\| \leq \delta\}) \leq Au^{-2}\delta^{-2}, \quad \text{for all } u, \delta \in (0, 1].$$

REFERENCES

- [1] R. Acar and C. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10:1217–1229, 1994.
- [2] R. A. Adams. *Sobolev Spaces*, volume 65 of *Pure and Applied Mathematics*. Academic Press, New York - London, 1975.
- [3] V. Barbu. *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Editura Academiei Republicii Socialiste Romănia, Bucharest, 1976.
- [4] V. Barbu and T. Precupanu. *Convexity and Optimization in Banach Spaces*. Editura Academiei, Bucharest, revised edition, 1978. Translated from the Romanian.
- [5] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IoP, 1998.
- [6] N. Bissantz, T. Hohage, and A. Munk. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20(6):1773–1789, 2004.
- [7] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636, 2007.
- [8] N. Bissantz and A. Munk. New statistical goodness-of-fit techniques in noisy inhomogeneous regression problems with an application to the problem of recovering of the luminosity density of the Milky Way from surface brightness data. In Feigelson, E. D. & Babu, G. J., editor, *Statistical Challenges in Astronomy*, pages 399–400. Springer-Verlag, 2003.
- [9] K. T. Block, M. Uecker, and J. Frahm. Undersampled radial mri with multiple coils. iterative image reconstruction using a total variation constraint. *Magnetic Resonance in Medicine*, 57(6):1086–1098, 2007.
- [10] S. M. Block. Making light work with optical tweezers. *Nature*, 360:493–495, Dec. 1992.
- [11] J. M. Borwein and A. S. Lewis. Convergence of best entropy estimates. *SIAM J. Optim.*, 1(2):191–205, 1991.
- [12] L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

- [13] E. M. Bronštejn. ε -entropy of convex sets and functions. *Sibirsk. Mat. Ž.*, 17(3):508–514, 715, 1976.
- [14] L. D. Brown, T. Cai, and H. H. Zhou. Nonparametric regression in exponential families. *Ann. Stat.*, 38(4):2005–2046, 2010.
- [15] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Stat.*, 24(6):2384–2398, 1996.
- [16] M. Burger, K. Frick, S. Osher, and O. Scherzer. Inverse total variation flow. *Multiscale Model. Simul.*, 6(2):365–395 (electronic), 2007.
- [17] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004.
- [18] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise c_2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, 2004.
- [19] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002. Dedicated to the memory of Lucien Le Cam.
- [20] L. Cavalier and A. Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002.
- [21] J. Cavanagh, W. J. Fairbrother, A. G. Palmer, N. J. Skelton, and M. Rance. *Protein NMR Spectroscopy, Second Edition: Principles and Practice*. Academic Press, 2 edition, 2006.
- [22] J. Ching, A. C. To, and S. D. Glaser. Microseismic source deconvolution: Wiener filter versus minimax, fourier versus wavelets, and linear versus nonlinear. *The Journal of the Acoustical Society of America*, 115(6):3048–3058, 2004.
- [23] P.-L. Chow, I. A. Ibragimov, and R. Z. Khasminskii. Statistical approach to some ill-posed problems for linear partial differential equations. *Probab. Theory Related Fields*, 113(3):421–441, 1999.
- [24] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.
- [25] A. Cohen, M. Hoffmann, and M. Reiß. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42(4):1479–1501 (electronic), 2004.
- [26] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. *Mach. Learn*, 48(48):253–285, 2002.
- [27] I. Csizár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, 19(4):2032–2066, 1991.
- [28] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.
- [29] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29(1):1–65, 2001. With discussion and rejoinder by the authors.
- [30] P. L. Davies, A. Kovac, and M. Meise. Nonparametric regression, confidence regions and regularization. *Ann. Statist.*, 37(5B):2597–2625, 2009.
- [31] D. L. Donoho. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Different Perspectives on Wavelets*, volume 47 of *Proc. Sympos. Appl. Math.*, pages 173–205, Providence, RI, 1993. Amer. Math. Soc.
- [32] D. L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126, 1995.
- [33] R. M. Dudley. *Uniform Central Limit Theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.

- [34] L. Dümbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152, 2001.
- [35] L. Dümbgen and G. Walther. Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785, 2008.
- [36] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*, volume 1 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-Oxford, 1976.
- [37] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1996.
- [38] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.
- [39] K. Frick. *The Augmented Lagrangian Method and Related Evolution Equations*. Phd-thesis, University of Innsbruck, 2008.
- [40] K. Frick, D. A. Lorenz, and E. Resmerita. Morozov’s principle for the augmented lagrangian method applied to linear inverse problems, 2010. arXiv:1010.5181v1.
- [41] K. Frick, P. Marnitz, and A. Munk. Statistical multiresolution estimation in imaging: Fundamental concepts and algorithmic framework, 2011. arXiv:1101.4373v1.
- [42] K. Frick and O. Scherzer. Regularization of ill-posed linear equations by the non-stationary Augmented Lagrangian Method. *J. Integral Equations Appl.*, 22(2):217–257, 2010.
- [43] A. Goldenshluger and S. V. Pereverzev. On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli*, 9(5):783–807, 2003.
- [44] I. Grama and M. Nussbaum. Asymptotic equivalence for nonparametric generalized linear models. *Probability Theory and Related Fields*, 111:167–214, 1998.
- [45] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with l^q penalty term. *Inverse Problems*, 24(5):055020, 2008.
- [46] R. J. Hanisch and R. L. White, editors. *The restoration of HST images and spectra - II*, 1994.
- [47] S. W. Hell. Far-Field Optical Nanoscopy. *Science*, 316(5828):1153–1158, 2007.
- [48] S. W. Hell. Microscopy and its focal switch. *Nature Methods*, 6(1):24–32, 2008.
- [49] W. Hinterberger, M. Hintermüller, K. Kunisch, M. von Oehsen, and O. Scherzer. Tube methods for BV regularization. *J. Math. Imaging Vision*, 19(3):219–235, 2003.
- [50] M. Hoffmann and M. Reiss. Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Stat.*, 36(1):310–336, 2008.
- [51] I. M. Johnstone. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica*, 9(1):51–83, 1999.
- [52] I. M. Johnstone, G. Kerkycharian, D. Picard, and M. Raimondo. Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(3):1467–9868, 2004.
- [53] I. M. Johnstone and B. W. Silverman. Discretization effects in statistical inverse problems. *Journal of Complexity*, 7(1):1–34, 1991.
- [54] G. Kerkycharian, G. Kyriazis, E. L. Pennec, P. Petrushev, and D. Picard. Inversion of noisy radon transform by svd based needlelets. *Applied and Computational Harmonic Analysis*, 28(1):24–45, 2010.
- [55] S. A. Kim, K. G. Heinze, and P. Schwille. Fluorescence correlation spectroscopy in living cells. *Nature Methods*, 4(11):963–973, Oct. 2007.
- [56] J. Lafferty, S. Pietra, and V. Pietra. Statistical learning algorithms based on bregman distances. In *Proceedings of the Canadian Workshop on Information Theory*, pages 77–80, Toronto, Canada, June 1997.
- [57] R. Liu, W. Strawderman, and C.-H. Zhang. *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*. Institute of Mathematical Statistics Lecture Notes—Monograph

- Series, 54. Institute of Mathematical Statistics, Hayward, CA, 2007.
- [58] J.-M. Loubes and C. Ludeña. Adaptive complexity regularization for linear inverse problems. *Electronic Journal of Statistics*, 2:661–677, 2008.
- [59] B. A. Mair and F. H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5):1424–1444, 1996.
- [60] E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997.
- [61] P. Mathé and S. V. Pereverzev. Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. *SIAM J. Numer. Anal.*, 38(6):1999–2021, 2001.
- [62] P. Mathé and S. V. Pereverzev. Discretization strategy for linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(6):1263–1277, 2003.
- [63] P. Mathé and S. V. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789–803, 2003.
- [64] J. R. McLaughlin. Absolute convergence of series of Fourier coefficients. *Trans. Amer. Math. Soc.*, 184:291–316, 1973.
- [65] A. Meister. Asymptotic equivalence of functional linear regression and a white noise inverse problem, 2011. to appear.
- [66] M. Nussbaum and S. Pereverzev. The degree of ill-posedness in stochastic and deterministic noise models. Technical Report 509, WIAS, 1999. Preprint.
- [67] F. O’Sullivan. A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527, 1986. With comments and a rejoinder by the author.
- [68] J. B. Pawley. *Handbook of Biological Confocal Microscopy*. Springer, 2006.
- [69] M. Popovic and A. Taflove. Two-Dimensional FDTD Inverse-Scattering Scheme for Determination of Near-Surface Material Properties at Microwave Frequencies. *IEEE Transactions on Antennas and Propagation*, 52:2366–2373, Sept. 2004.
- [70] M. Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *ArXiv Mathematics e-prints*, July 2006.
- [71] E. Resmerita. On total convexity, Bregman projections and stability in Banach spaces. *J. Convex Anal.*, 11(1):1–16, 2004.
- [72] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.
- [73] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1996. Translated from the first (1980) Russian edition by R. P. Boas.
- [74] G. Siuzdak. *Mass Spectrometry for Biotechnology*. Academic Press, 1996.
- [75] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.
- [76] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.
- [77] Y. Vardi. Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *Journal of the American Statistical Association*, 91(433):365–377, March 1996.
- [78] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4):651–667, 1977.
- [79] W. Walz, editor. *Patch-Clamp Analysis: Advanced Techniques*, volume 38 of *Neuromethods*. Humana Press, New Jersey, 2007.
- [80] T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

- [81] W. P. Ziemer. *Weakly Differentiable Functions*. Springer Verlag, New York, 1989.
- [82] A. Zygmund. *Trigonometric Series. Vol. I, II*. Cambridge University Press, Cambridge, 1977.
Reprinting of the 1968 version of the second edition with Volumes I and II bound together.