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

Preprint FOR916 10-15

Updated version
former title "Locally Adaptive Regularization of Linear Statistical Inverse Problems"

Preprint-Series of the Research Group FOR916

FOR916/10-15 2010

SHAPE CONSTRAINED REGULARISATION BY STATISTICAL MULTIRESOLUTION FOR INVERSE PROBLEMS

KLAUS FRICK, PHILIPP MARNITZ AND AXEL MUNK

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 sub-spaces 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 local adaptive behaviour.

1. Introduction

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

$$Ku = g, \tag{1}$$

where $K : U \to V$ 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). *(Statistical) regularisation methods* are one approach for computing stable approximations of true solutions u from (statistically) perturbed data g.

To be more precise, assume that $u \in U$ is a solution of (1) and that we are given the observation

$$Y = Ku + \sigma\varepsilon. \tag{2}$$

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

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

Model (2) is very common in the theory of statistical inverse problems (see e.g. [5, 16, 17, 19, 44]) and covers numerous models arising in many applications (see [5] for various examples).

The literature on statistical regularisation methods is vast and we only give a few, selective references: Penalized least-squares estimation (that includes Tikohonov-Philipps and maximum entropy regularisation) [6, 45, 53], wavelet methods [24, 25, 36], estimation in Hilbert-scales [5, 34, 38, 40–42] and regularisation by projection [15, 16, 19, 40] to name but

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

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

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

$$J(u) \to \inf! \quad \text{subject to} \quad T(Y, Ku) \leq q. \tag{3}$$

Here, $T(v,w)$ denotes some notion of *distance* on the image space V that measures the deviation of the data Y and the estimated image Ku, $q \in \mathbb{R}$ is a *threshold value* and J denotes a measure of *complexity* for candidate estimators $u \in U$. In other words, regularisation methods of type (3) pick among all estimators u for which the distance T of Ku and the data Y does not exceed a given threshold value q one with smallest complexity J.

Whereas much 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. The most common choice in a Hilbert-space setting is the squared-norm fidelity

$$T(Y, Ku) = \|Y - Ku\|^2 .$$

and mostly the choice of J is considered to be more relevant for proper reconstruction of u. We claim, however, that from a statistical perspective the choice of T is of equal importance. To this end, we will introduce a particular family of distance functions T, the so called *multi-resolution (MR)-statistics* within the framework of statistical inverse problems. In their simplest form, 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 $1 \leq n \leq N$), that is

$$T(Y, Ku) = \sigma^{-1} \sup_{1 \leq n \leq N} |\langle Y - Ku, \phi_n \rangle| .$$

Under the hypothesis that u is the true solution of (1), we have that $\langle Y - Ku, \phi_n\rangle \sim \mathcal{N}(0, \sigma^2)$ for $1 \leq n \leq N$ and $T(Y, Ku)$ does not exceed a (yet to be defined) threshold with high probability. If, however, the residual $Y - Ku$ contains a non-random signal and for some $1 \leq n_0 \leq N$

$$\mathbf{E}\left(\langle Y - Ku, \phi_{n_0}\rangle\right) \neq 0 \tag{4}$$

the statistic $T(Y, Ku)$ becomes relatively large and u happens to lie outside the admissible domain of the optimisation problem (3). Hence, the multi-resolution constraint in (3) protects against too parsimonious reconstructions due to minimising J.

The choice of the test-elements $\phi_1, \ldots, \phi_N$ is subtle, since they should not miss any non-random information in the residual, if present. In principle, T 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 span$\{\phi_1, \ldots, \phi_N\}$. This approach, however, turns (3) into an optimisation problem with a huge number of constraints which is hard to tackle numerically. Besides these numerical difficulties, there is also a statistical limitation: If the entropy of the system $\{\phi_n\}_{n\in\mathbb{N}}$ becomes too large, the asymptotic distribution of T will be degenerated and hence useless for our purposes. Instead, it is necessary and possible to incorporate a-priori knowledge on the true solution of (1) in order to come up with *sparse* and efficient systems of test-elements.

The study of MR-statistics has attracted much attention recently. In [50] MR-statistics (called *scanning-statistics* there) are studied in order to detect a signal against a noisy background on multi-dimensional spatial regions, where the background is modelled as independent observation of exponential family distribution. In [27, 28] MR-statistics are used in order to test qualitative hypothesis (as monotonicity or concavity) in non-parametric regression problems. MR-statistics in non-parametric regression problems are also studied in [22] where the focus is put on controlling local extremes. In [7] the authors use MR-statistics in order to formulate a stopping criterion for the EM-Algorithm for Positron-Emission-Tomography.

The regularisation scheme (3) with MR-statistic T was first studied in [23] for non-parametric regression in one space dimension. The authors focused on the total-variation semi-norm as complexity measure J. We will extend these results in several ways: First, our analysis allows for *indirectly observed data*, that is linear inverse problems of the form (1). We mention that this approach can be combined with many choices of J. To illustrate our idea, we discuss total variation penalisation for two (and possibly higher) dimensions, which is of particular interest for image processing tasks.

Furthermore, we present very general consistency and convergence rates results for SMR-estimation and discuss their impact on particular applications. To our best knowledge, results of this type have never been obtained before. We note, that in the situation of inverse problems it is necessary to assume additional regularity of the true solution of (1) in order to come up with convergence rates results. This is usually done by formulating so-called *source conditions* that determine smoothness classes of solutions for (1) that allow fast reconstruction. In this work we study the standard source conditions used in the framework of Bregman-divergences that yield for each penalty functional J in (3) *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.

We mentioned that for large values of N, the optimisation problem (3) in general is hard to be solved numerically. In particular, our experience shows that standard algorithms such as interior point or conjugate gradient methods are far from being satisfactory. Currently, we develop a feasible numerical scheme based on a combination of the Augmented Lagrangian Method and Dykstra's projection algorithm [9]. For the sake of brevity, we will not discuss the details here and refer to upcoming work.

This paper is organized 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 (3) where we employ multi-resolution statistics 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 complexity 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 complexity function J is chosen to be the total-variation semi-norm, which has a particular appeal for imaging problems. Finally, the proofs of the main results and some auxiliary lemmata are collected in the Appendix A.

2. Basic Definitions

In this section we summarize some relevant definitions and assumptions needed throughout the paper. We start with the basic

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 $J : U \to \overline{\mathbb{R}}$ 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) $K : U \to V$ is a linear and bounded operator.

In the course of this paper we will frequently make use of tools from convex analysis. For a standard reference see [29].

- The sub-differential (or generalized derivative) $\partial J(u)$ of J at u is the set of all elements $p \in U$ satisfying
$$J(v) - J(u) - \langle p, v - u\rangle \geq 0 \qquad \text{for all } v \in U$$
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
$$J'(v)(w) = \lim_{h\to 0^+} \frac{J(v + hw) - J(v)}{h}.$$
and is well defined for convex functions (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
$$0 \leq D_J(v, u) \leq D_J^\xi(v, u), \qquad \text{for all } \xi \in \partial J(u).$$

Remark 2.2. 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 (3), 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 [46] 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 2.4-2.6 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 [10] and has recently attracted much attention e.g. in the inverse problems community (cf. [12, 14, 21, 33, 47]) or in statistical and machine learning ([20, 37, 54]).

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

Definition 2.3. (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

$$K^* p \in \partial J(u). \tag{5}$$

Then, we say that u *satisfies the source condition* (5).

It is well-known in the theory of inverse problems with deterministic noise (cf. [30]) that the source condition (5) 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, the source condition (5) characterises *one particular* smoothness-class of solutions for (1) for which fast reconstruction is guaranteed. We note, that for standard choices for J (as e.g. in Example 2.4) there exist more sophisticated source conditions than (5) that allow for improved convergence rate results (as e.g. Hölder source conditions). Such extensions, however, are not straightforward to generalise and are—to a large extent—not well understood so far.

We clarify the notions *Bregman-divergence* and *source condition* by some examples.

Example 2.4. 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

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

Moreover, the source condition (5) can be rewritten to

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

Since $\operatorname{ran}(K^*) = \operatorname{ran}(K^*K)^{1/2}$, this shows that the source condition (5) corresponds to the *Hölder-source condition* $u^\dagger \in \operatorname{ran}(K^*K)^{\beta}$ for $\beta = 1/2$ (cf. [30]).

Example 2.5. Assume that $U = \mathrm{L}^2(\Omega)$ for an open and bounded set $\Omega \subset \mathbb{R}^n$ with Lipschitz boundary $\partial\Omega$ and outer unit-normal ν and let $\mathrm{H}^{\beta}(\Omega)$ denote the Sobolev-space of order $\beta \in \mathbb{R}$. We define

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

Then (cf. [3, pp.63]), the set $D(\partial J)$ consists of all elements $u \in \mathrm{H}^2(\Omega)$ 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

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

Moreover, $u^\dagger$ satisfies the source condition (5) 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 2.6. Let U be as in Example 2.5 and define the *negentropy* by

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

Then (cf. [4, Chap. 2 Prop 2.7]), the set $D(\partial J)$ consists of all non-negative functions in $\mathrm{L}^\infty(\Omega)$ 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

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

that is, the Bregman-divergence coincides in this particular case with the *Kullback-Leiber-divergence.* It was proven in [8, Lem. 2.2] that

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

In other words, Bregman-consistency (or convergence rates) w.r.t. the negentropy yields strong consistency (convergence rates) in $\mathrm{L}^1(\Omega)$. Finally, we note that $u^\dagger \in D(\partial J)$ satisfies the source condition (5) with source element $p^\dagger \in V$ if and only if

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

In Section 4.3 we will study a more complex example in more detail, where J is the total-variation of a measurable function on a domain Ω.

Under fairly general conditions existence of J minimising solution can be guaranteed. We formalize these conditions in the following result, however, we omit the proof since it is standard in convex analysis (cf. [29, Chap. II Prop. 2.1]).

Proposition 2.7. *Let $g \in V$ be attainable and assume that for all $c \in \mathbb{R}$ the sets*

$$\{u \in U \ : \ \|Ku\| + J(u) \leq c\} \tag{6}$$

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 2.3) 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. **Multi-resolution 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

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

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

(1) 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

$$\sup_{\substack{r\in(0,1]\\ N\in\mathbb{N}}} L_{Nr} =: L < \infty \tag{7}$$

and

$$0 > \lambda_N(r) := \inf_{s\in\mathbb{R}^+} t_N(s,r) > -\infty. \tag{8}$$

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

$$t_N(s,r) \geq c_1 s + c_2 t_N(\sigma s, r) \quad \text{for } (s,r) \in \mathbb{R}^+ \times (0,1] \text{ and } N \in \mathbb{N}. \tag{9}$$

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

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

is called a *multi-resolution (MR)- statistic.*

Remark 3.2. Let $\varepsilon : V \to \mathrm{L}^2(\Omega, \mathfrak{A}, \mathbb{P})$ be a white noise process and consider the random variables

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

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

$$q_N(\alpha) := \inf\left\{q \in \mathbb{R} \;:\; \mathbb{P}\left(T_N(\varepsilon) \leq q\right) \geq 1-\alpha\right\} \tag{10}$$

Definition 3.1 allows for a vast class of MR-statistics and the conditions in (7)-(9) 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 allows for more involved models, as e.g. introduced in [28].

Example 3.3. Assume that $\{f_N : (0,1] \to \mathbb{R}\}_{N\in\mathbb{N}}$ 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 3.4. 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

$$J(u) \to \inf! \quad \text{s.t.} \quad T_N(\sigma^{-1}(Y - Ku)) \leq q_N(\alpha) \tag{11}$$

is called a *statistical multi-resolution 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

$$\mathcal{A}_N(\alpha) = \left\{u \in U \ : \ T_N(\sigma^{-1}(Y - Ku)) \leq q_N(\alpha)\right\}$$

of the optimisation problem (11). From construction it follows that the exact solution(s) of (1) lie within $\mathcal{A}_N(\alpha)$ with a probability of at least $1-\alpha$. Thus, $\mathcal{A}_N(\alpha)$ serves as a $1-\alpha$ confidence region for $\hat{u}_N(\alpha)$ and therefore α exhibits an intrinsic statistical meaning (see also [23]). This stands 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.

In order to guarantee existence of a solution of the convex problem in Definition 3.4, that is existence of an SMRE, it is necessary to impose further (standard) assumptions:

Assumption 3.5. *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 3.5 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 2.4 or 2.6) then Assumption 3.5 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 3.5.

Application of standard arguments from convex optimisation yields

Proposition 3.6. *Assume that Assumption 3.5 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 3.5 already implies the requirements in Proposition 2.7 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 argumentation following Definition 3.4, 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 (11) 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 3.5 hold and that $\{\sigma_k\}_{k\in\mathbb{N}}$ is a sequence of positive noise-levels in (2) such that $\sigma_k \to 0^+$ as $k \to \infty$. 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

$$\sum_{k=1}^{\infty} \alpha_k < \infty \quad \text{and} \quad \lim_{k\to\infty} N_k = \infty. \tag{12}$$

Theorem 3.7. *Let $u^\dagger$ be a J-minimising solution of* (1) *where $g \in \overline{\mathrm{span}\Phi}$ and assume that*

$$\sup_{N\in\mathbb{N}} T_N(\varepsilon) < \infty \qquad a.s. \tag{13}$$

and that

$$\eta_k := \sigma_k \max\left(\inf_{1\le n\le N_k} \lambda_{N_k}(\|\phi_n\|), \sqrt{-\log\alpha_k}\right) \to 0. \tag{14}$$

Then, for $\hat{u}_k := \hat{u}_{N_k}(\alpha_k)$ as in (11) *one has*

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

as well as

$$\limsup_{k\to\infty} \sup_{1\le n\le N_k} \frac{|\langle \phi_n^*, K\hat{u}_k - Ku^\dagger\rangle|}{\eta_k} < \infty \qquad a.s. \tag{16}$$

Theorem 3.7 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 MR-statistic $T_N(\varepsilon)$ in (13) 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 2.3 (iii) are known to constitute sufficient regularity conditions in deterministic noise models (cf. [30]). In our situation we additionally have to assume that the source elements exhibit certain approximation properties:

Assumption 3.8. *There exists a J-minimising solution $u^\dagger$ of* (1) *that satisfies the source condition* (5) *with source element $p^\dagger$. Moreover, for $n, N \in \mathbb{N}$ there exist $b_{n,N} \in \mathbb{R}$ such that*

$$\lim_{N\to\infty} \left\| p^\dagger - \sum_{n=1}^N b_{n,N}\phi_n^* \right\| = 0 \quad \text{and} \quad \sup_{N\in\mathbb{N}} \sum_{n=1}^N |b_{n,N}| < \infty. \tag{17}$$

Remark 3.9. i) Assumption 3.8 amounts to say that there exists a J-minimising solution $u^\dagger$ that satisfies the source condition (5) with a source element $p^\dagger$ that can be approximated sufficiently well by the used system of test-functions Φ. From (5) it becomes clear that we can always assume that $p^\dagger \in \overline{\mathrm{ran}(K)}$, such that the first condition in (17) is not very restricitve, in fact.

For the sake of convenience, we introduce an abbreviation for the approximation error w.r.t. the dictionary Φ

$$\mathrm{err}_N(p^\dagger) := \left\| p^\dagger - \sum_{n=1}^N b_{n,N}\phi_n^* \right\|. \tag{18}$$

ii) It is important to note that, given prior knowledge of the true solution $u^\dagger$, the conditions in Assumption 3.8 indicate how to choose an efficient system of test-functions, which will become apparent for particular applications.

Theorem 3.10 (Convergence rates for SMRE)**.** *Let the requirements of Theorem 3.7 be satisfied and assume further that Assumption 3.8 holds with $g \in \overline{\operatorname{span}\Phi}$. If*

$$\eta_k := \max\left(-\sigma_k \inf_{1\le n\le N_k} \lambda_{N_k}(\|\phi_n\|),\ \operatorname{err}_{N_k}(p^\dagger),\ \sigma_k\sqrt{-\log\alpha_k}\right) \to 0 \tag{19}$$

as $k \to \infty$, then (16) *holds and additionally*

$$\limsup_{k\to\infty} \frac{D_J^{K^*p^\dagger}(\hat u_k, u^\dagger)}{\eta_k} < \infty \quad a.s. \tag{20}$$

Remark 3.11. The convergence rate result in Theorem 3.10 is rather general, in the sense that the rate function η_k in (20) 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 3.8 with a source element $p^\dagger$.

(1) The sequence $\{-\inf_{1\le n\le N}\lambda_N(\|\phi_n\|)\}_{N\in\mathbb{N}}$ is positive according to (8). Hence

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

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

(2) After setting $\eta_k = -\sigma_k \inf_{1\le n\le N_k}\lambda_{N_k}(\|\phi_n\|)$ it remains to check that the sequence of test-levels defined by

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

is summable (for a constant $\kappa > 0$).

For the above construction of N_k, η_k and α_k, it follows from Theorem 3.10 that for the SMR-estimators $\hat u_k = \hat u_{N_k}(\alpha_k)$ the estimate in (20) holds.

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 first study the case where T_N constitutes the extreme-value statistic of the coefficients w.r.t. an orthonormal systems of test-functions Φ (Section 4.1). In Section 4.2 we skip the assumption of orthonormality and examine general SMR-estimation w.r.t. (non-orthonormal) systems of test-functions that satisfy certain entropy conditions. Finally, we study the case when the penalty functional J is chosen to be the total-variation semi-norm on $U = \mathrm{L}^2(\Omega)$ in Section 4.3.

Throughout this section we assume that Assumptions 2.1 and 3.5 hold. Moreover we shall agree upon $\{\sigma_k\}_{k\in\mathbb{N}}$ being a sequence of noise levels such that $\sigma_k \to 0^+$ and that for $k \in \mathbb{N}$ there are $\alpha_k \in (0,1)$ and $N_k \in \{N_0, N_0+1, \ldots\}$ such that (12) holds.

4.1. Introductory Example: Gaussian Sequence Model.

In this section we shall consider the case where the dictionary $\Phi = \{\phi_1, \phi_2, \ldots\}$ constitutes an orthonormal basis of $\overline{\operatorname{ran}(K)}$. 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 multi-resolution 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 system Φ, that is

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

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

$$\lim_{N\to\infty} T_N(\varepsilon) = \lim_{N\to\infty} \left(\sup_{1\leq n\leq N} |\varepsilon_n| - \sqrt{2\log N} \right) = 0 \qquad \text{a.s.}$$

In what follows, we will apply Theorems 3.7 and 3.10 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 3.7.

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

$$\lim_{k\to\infty} \sigma_k^2 \max(\log N_k, -\log\alpha_k) = 0$$

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 3.10, Assumption 3.8 has to be verified. We set $b_{nN} \equiv \left\langle p^\dagger, \phi_n \right\rangle$ in Assumption 3.8. Note that for $p \in V$ the expression $\text{err}_N(p)$ (as defined (18)) denotes the approximation error of the N-th partial Fourier-series w.r.t. Φ. Thus, Assumption 3.8 is linked to absolute summability of the Fouerier-coefficients w.r.t. the basis Φ, i.e.

$$\sum_{n=1}^{\infty} \left| \left\langle p^\dagger, \phi_n \right\rangle \right| < \infty \tag{21}$$

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,* (21) *is satisfied if*

$$\sum_{N=1}^{\infty} \frac{\text{err}_N(p^\dagger)}{\sqrt{N}} < \infty. \tag{22}$$

Proof. The classical version of the Bernstin-Stechkin Theorem (see e.g. [43, Thm. 7.4]) states that for each $f \in \mathrm{L}^2(0,1)$ and each ON-basis $\underline{v} = \{v_1, v_2, \ldots\}$ of $\mathrm{L}^2(0,1)$, the Fourier-coefficients of f are absolutely summable, if (22) holds. Since each seperabel Hilbert space is isometrical isomorph to $\mathrm{L}^2(0,1)$, the assertion finally follows. □

Following the procedure outlined in Remark 3.11, Section 3, we define

$$N_k := \inf\left\{ N \in \mathbb{N} \;:\; \text{err}_N(p^\dagger) \leq \sigma_k\sqrt{2\log N} \right\} \qquad \text{and} \qquad \eta_k := \sigma_k\sqrt{2\log N_k}. \tag{23}$$

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 (23). 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 (20).

The problem of characterising those elements $p^\dagger \in V$ that satisfy (22) is a classical issue in Fourier-analysis and approximation theory. Sufficient conditions for (22) to hold are usually formalized 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 elliposids*: for a constants $\beta, Q > 0$ we define $\Theta(\beta, Q)$ as the infinite-dimensional ellipsoid

$$\Theta(\beta, Q) = \left\{ \theta \in \ell^2 \ : \ \sum_{n\in\mathbb{N}} n^{2\beta}\theta_n^2 \le Q^2 \right\}. \tag{24}$$

The *Sobolev class* $W(\beta, Q) \subset V$ is then defined as the collection of elements $v \in V$ such that $\{\langle v, \phi_n\rangle\}_{n\in\mathbb{N}} \subset \Theta(\beta, Q)$ (cf. [51, Sec.1.10.1]). For $v \in W(\beta, Q)$ we have that (22) holds 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 ONB of $\ker(K)^\perp$, $\{\phi_n\}_{n\in\mathbb{N}}$ is an ONB of $\overline{\operatorname{ran}(K)}$ and the singular values $\{s_n\}_{n\in\mathbb{N}}$ are positive and $s_n \to 0$ as $n \to \infty$. Moreover

$$K\psi_n = s_n\phi_n \quad \text{and} \quad K^*\phi_n = s_n\psi_n, \tag{25}$$

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 $\hat{u}_N(\alpha)$ 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 $\hat{u}_N(\alpha)$ resembles James-Stein type estimators, however uses $|y_n|$ in contrast to $|y_n|^2$.

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 2.4) with source element $p^\dagger \in W(\beta, Q)$ for $Q > 0$ and $\beta = 1/2 + \varepsilon$ (with $\varepsilon > 0$ small). Then, $\operatorname{err}_N(p^\dagger) \le QN^{-\beta}$ and it follows from (23) 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 2.4 that

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

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 (22) in a function space setting, are usually formalized in characterising smoothness properties. The following example shows how this applies to Hölder-continuity.

Example 4.5. Let $V = \mathrm{L}^2_{\mathrm{per}}([0,1])$ be the Hilbert space of all square-integrable and periodic functions on the unit interval. Moreover, we assume that $\overline{\mathrm{ran}(K)} = \mathrm{L}^2([0,1])$ 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 A.4) with $\beta = 1/2 + \varepsilon$. Then we have that $\mathrm{err}_N(p^\dagger)Q \leq N^{-\beta}\log N$ for a suitable constant $Q > 0$ and therefore it follows from Proposition 4.2 that (21) holds.

Hence, if $u^\dagger$ is a J-minimising solution of (1) that satisfies the source condition (5) 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 (20).

Remark 4.6. 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 $\overline{\mathrm{ran}(K)}$ such that (22) is satisfied. This holds for example for a vast class of orthonormal wavelet bases of $\mathrm{L}^2([0,1])$ as studied in [18].

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 $\beta > 0$ is of bounded variation (cf. [56, Vol.1 Thm.3.6]).

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

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

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

$$t_N(s,r) = s - \sqrt{-2\gamma\log r}, \quad (s,r) \in \mathbb{R}^+ \times (0,1] \tag{26}$$

where $\gamma > 0$ is a constant. As outlined in Example 3.3, 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 (26) has to be chosen appropriately in dependence on Φ in order to guarantee that the MR-statistic $T_N(\varepsilon)$ is bounded almost surely. Sufficient conditions on γ have been given in [27, 28] and we provide a brief summary of the respective results. To this end, we recall the following

Definition 4.7. Let (T,d) be a semi-metric space, $T' \subset T$ and $\varepsilon > 0$. The *capacity number* is defined by

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

With this preparation we can apply [28, Thm 7.1] and find

Proposition 4.8. *If there exists constants $A, B > 0$ such that*

$$D(u\delta, \{\phi \in \Phi \ : \ \|\phi\| \leq \delta\}) \leq Au^{-B}\delta^{-\gamma}, \qquad \text{for all } u, \delta \in (0,1] \tag{27}$$

then

$$\sup_{N\in\mathbb{N}} T_N(\varepsilon) < \infty \qquad a.s.$$

Corollary 4.9. Let $u^\dagger \in U$ be a J-minimising solution of (1) where $g \in \overline{\operatorname{span}\Phi}$ and $\gamma > 0$ be chosen such that the assumption of Proposition 4.8 is satisfied. Moreover, assume that

$$\lim_{k\to\infty} \sigma_k^2 \min(\min_{1\leq n\leq N_k} \log\left(\|\phi_n\|\right), \log \alpha_k) = 0.$$

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 3.10, it is necessary that a J-minimising solution $u^\dagger$ of (1) satisfies the source condition (5) with a source element $p^\dagger$ that can be approximated by the system of test-functions Φ sufficiently well (cf. Assumption 3.8). Good estimates of approximation errors for general systems Φ are hard to come up with in general and will not be treated in this work. Instead, we illustrate the assertion of Theorem 3.10 when $U = V = \mathrm{L}^2([0,1]^d)$ $(d \geq 1)$ and when Φ consists of a countable selection of indicator functions on cubes in $[0,1]^d$.

First, we shall examine when the MR-statistic $T_N(\varepsilon)$ (almost surely) stays finite, a sufficient condition of which is formulated in Proposition 4.8. 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 A.8, the assumptions of Proposition 4.8 are satisfied for $\Phi = \Phi_{\mathrm{s}}$ and $\gamma = d$. Particularly, it follows that the assertion of Proposition 4.8 also holds for arbitrary (countable) sub-systems $\Phi \subset \Phi_s$, that is the statistic

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

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

Next, we study Assumption 3.8 in the present setting. Let $\mathcal{P} = \{Q_1, Q_2, \ldots\}$ 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 A.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

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

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 3.11 and applying Lemma A.5 we therefore define for a given continuous function $p^\dagger : [0,1]^d \to \mathbb{R}$

$$m_k := \inf\left\{ m \in \mathbb{N} \ : \ \frac{m+1}{\sum_{\nu=0}^m \omega^{-2}(\delta_\nu, p^\dagger)} \leq -2\sigma_k^2 \log \varepsilon_m \right\} \quad \text{and} \quad \eta_k := \sigma_k\sqrt{-2\log\varepsilon_{m_k}}. \tag{28}$$

Here $\omega(\cdot, p^\dagger)$ denotes the modulus of continuity of $p^\dagger$ (cf. Definition A.4). With this and the general convergence rate result in Theorem 3.10 we immediately obtain

Corollary 4.10. Let $u^\dagger \in \mathrm{L}^2([0,1]^d)$ be a J-minimising solution of (1) where $g \in \overline{\mathrm{span}\Phi}$ and that satisfies the source condition (5) with source element $p^\dagger \in C([0,1]^d)$. Moreover, let m_k and η_k be defined as in (28). If

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

for a constant $\kappa > 0$, then the SMRE $\hat{u}_k = \hat{u}_{N(m_k)}(\alpha_k)$ almost surely satisfy (20).

Example 4.11. We consider the system of all dyadic partitions $\mathcal{P} = \mathcal{P}_2$ of $[0,1]^d$ as in Example A.9. In particular, we note that the assumptions of Lemma A.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

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

for $m \in \mathbb{N}$ large enough. From this and (28) 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 (5) 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 (20) 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 [22], have been proposed (see e.g. [22, 23, 39] and references therein). In two or more space dimensions, however, the situation is much more involved and a generalisation of numerical methods is usually not straightforward (see e.g. [35]). In [23] SMR-estimation with total variation penalty was studied for the case of pure regression problems ($K = \mathrm{Id}$) in one space dimension. We study here an extension 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 $U = \mathrm{L}^2(\Omega)$ and define $\mathrm{BV}(\Omega)$ to be the collection of $u \in U$ whose derivative $\mathrm{D}u$ (in the sense of distributions) is a signed $\mathbb{R}^2$-valued Radon-measure with finite total-variation $|\mathrm{D}u|$, that is

$$|\mathrm{D}u|(\Omega) = \sup_{\substack{\psi \in C_0^1(\Omega,\mathbb{R}^2) \\ |\psi| \leq 1}} \int_\Omega \mathrm{div}(\psi)\, u \,\mathrm{d}x < \infty.$$

We note that the norm

$$\|u\|_{\mathrm{BV}} := \|u\|_{\mathrm{L}^1} + |\mathrm{D}u|(\Omega)$$

turns $\mathrm{BV}(\Omega)$ into a Banach-space and that with this norm $\mathrm{BV}(\Omega)$ is continuously embedded into $\mathrm{L}^2(\Omega)$. The embedding is even compact if $\mathrm{L}^2(\Omega)$ is replaced by $\mathrm{L}^p(\Omega)$ with $p < 2$ (a proof of these embedding results can be found in [1, Thm. 2.5]. For an exhaustive treatment of $\mathrm{BV}(\Omega)$ see [31, 55]). With this, we define

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

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 $\mathrm{L}^2(\Omega)$. This shows, that J satisfies Assumption 2.1 (ii). Next, we examine Assumption 3.5:

Lemma 4.12. If there exists $n_0 \in \mathbb{N}$ such that

$$|\langle K\mathbf{1}, \phi_{n_0}\rangle| > 0,$$

then Assumption 3.5 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 $\sup_{k\in\mathbb{N}} J(u_{k_n}) \le c < \infty$ and thus we find with Poincaré's inequality (cf. [55, Thm. 5.11.1])

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

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

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

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

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

□

We note that the assumptions in Lemma 4.12 already imply the weak compactness of the sets (6) and thus guarantee existence of a J-minimising solution of (1). From the above cited embedding properties of the space $\mathrm{BV}(\Omega)$ it is easy to derive an improved version of the consistency result in Theorem 3.7.

Corollary 4.13. Let $g \in \overline{\mathrm{span}\Phi}$ and assume that $u^\dagger \in \mathrm{BV}(\Omega)$ 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 3.7 and define $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$. Then, additionally to the assertions in Theorem 3.7 we have that

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

for every $1 \le p < 2$.

Proof. From Theorem 3.7 it follows that $\{\hat{u}_k\}_{k\in\mathbb{N}}$ is bounded a.s. in $\mathrm{L}^2(\Omega)$ 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 $\mathrm{L}^2(\Omega)$ a.s. and therefore also in $\mathrm{L}^p(\Omega)$ for each $1 \le p < 2$.

Since Ω is assumed to be bounded, it follows that $\mathrm{L}^2(\Omega)$ is continuously embedded into $\mathrm{L}^1(\Omega)$. Thus, it follows from Theorem 3.7 that

$$\sup_{k\in\mathbb{N}} \|\hat{u}_k\|_{\mathrm{BV}} < \infty \qquad \text{a.s.}$$

From the compact embedding $\mathrm{BV}(\Omega) \hookrightarrow \mathrm{L}^p(\Omega)$ for $1 \le p < 2$, it hence follows that $\{\hat{u}_k\}_{k\in\mathbb{N}}$ is compact in $\mathrm{L}^p(\Omega)$. 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 3.10 to strong L^p-convergence and thus we have to settle for the general results in Theorem 3.10. Therefore, we aim for an interpretation of convergence w.r.t. the Bregman-divergence in (20). We summarize:

Lemma 4.14. (i) One has $\xi \in \partial J(u)$ if and only if there exists $z \in \mathrm{L}^\infty(\Omega, \mathbb{R}^2)$ with $\|z\|_{\mathrm{L}^\infty} \le 1$ such that $\langle z, \nu\rangle = 0$ on $\partial\Omega$,

$$\operatorname{div}(z) = \xi \qquad \text{and} \qquad \int_\Omega \xi u \,\mathrm{d}x = |\mathrm{D}u|(\Omega).$$

(ii) Let $\xi \in \partial J(u)$. Then,

$$D_J^\xi(v,u) = |\mathrm{D}v|(\Omega) - \int_\Omega \xi v \,\mathrm{d}x.$$

Proof. Assertion (ii) directly follows from the definition of the Bregman-divergence and (i). The equivalence relation in (i) was proven e.g. in [32, Thm. 4.4.2]. □

Remark 4.15. The result in Lemma 4.14 (ii) allows a geometrical interpretation of the Bregman-divergence w.r.t. the functional J. As it was worked out in [13, Sec. 5.1], one can show that

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

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 1(a)).

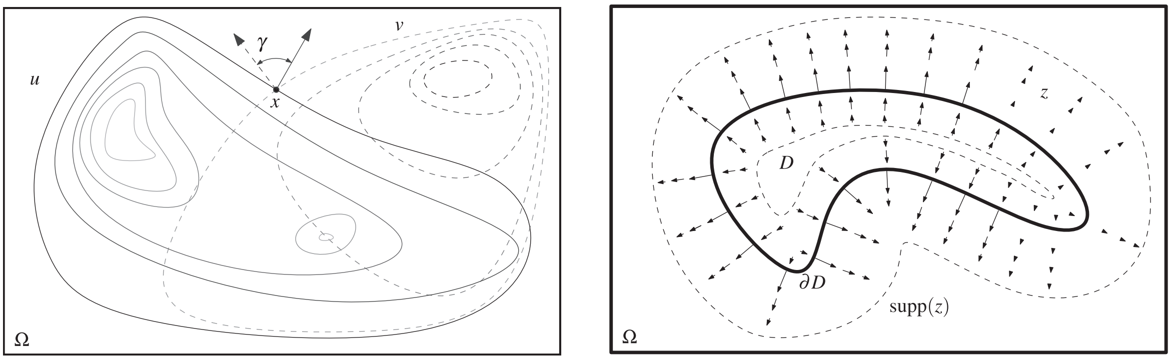
We recall that a function $u \in \mathrm{BV}(\Omega)$ satisfies the source condition, if there exists $\xi \in \operatorname{ran}(K^*)$ such that $\xi \in \partial J(u)$. It is important to note, that in many applications the elements in $\operatorname{ran}(K^*)$ exhibit high regularity such as continuity 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 4.14 yield (see also [48, Lem.3.71])

Lemma 4.16. 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 $|z| \le 1$ and

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

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



(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 $\mathrm{div}\,(z) \in \partial J(u)$

Figure 1. TV-Regularisation.

In many applications (such as imaging) the true solution $u \in \mathrm{BV}(\Omega)$ 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 [48, Ex. 3.74]

Example 4.17. 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 1(b)). Independent of the choice of the extension, we then have $\xi := \mathrm{div}\,(z) \in \partial J(u)$ and $\xi \in C_c^\infty(\Omega)$.

Example 4.18 (continue Example 4.11)**.** We consider $\Omega = [0,1]^2$ and $V = \mathrm{L}^2(\Omega)$. Moreover, we assume that $\mathcal{P}_2$ denotes the set of all dyadic partitions of Ω (cf. Example A.9) and that Φ is the collection of indicator functions w.r.t. elements in $\mathcal{P}_2$.

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

$$(Ku)(x) = \int_{\mathbb{R}^2} k(x-y)\bar{u}(y)\,\mathrm{d}x \qquad \text{for } x \in \Omega$$

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 4.17. 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

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

then Assumption 3.8 is satisfied. To be more precise, we have that $p^\dagger \in \mathcal{H}_{\beta-1}(\Omega)$ (cf. [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 4.11 and Lemma 4.14 that

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

for the SMRE $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$ (where N_k is as in Example 4.11).

Appendix A. Proofs

A.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 (10). Unless otherwise stated, we shall assume that Assumptions 2.1 and 3.5 hold.

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

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

Proof. First, we introduce the function $f : \mathbb{R}^N \to \mathbb{R}$ as

$$f(x_1, \dots, x_N) = \sup_{1 \leq n \leq N} t_N(x_n, \|\phi_n\|)$$

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 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 (c.f. [52, Lem. A.2.2]) we find that for all $\eta \in \mathbb{R}$

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

Now let $\alpha \in (0,1)$. Then,

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

Rearranging the above inequality yields the desired estimate. □

We proceed with the proof of the existence result in Theorem 3.6. 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 3.6. Let $N \geq N_0$ and $y \in V$ be arbitrary. Due tu Assumption 2.1 (ii), $D(J) \subset U$ is dense and hence there exists for all given $\delta > 0$ an element $u_0 \in D(J)$ such that $\|Ku_0 - \tilde{y}\| \leq \delta$, where $\tilde{y}$ denotes the orthonormal projection of y onto $\overline{\operatorname{ran}(K)}$. Since $\phi_n \in \overline{\operatorname{ran}(K)}$ and $\|\phi_n^*\| = 1$ for all $n \in \mathbb{N}$, this implies that

$$|\langle Ku_0 - y, \phi_n^* \rangle| = |\langle Ku_0 - y, \phi_n^* \rangle| \leq \delta$$

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 according to the reasoning above, there exists $u_0 \in D(J)$ such that for $1 \leq n \leq N$

$$L\sigma^{-1} |y_n - \langle Ku_0, \phi_n^* \rangle| \leq q_N(\alpha) - \sup_{1 \leq n \leq N} \lambda_N(\|\phi_n\|), \tag{29}$$

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 (9), it then follows that

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

This is a contradiction to the definition of $q_N(\alpha)$ in (10) and thus $u_0 \in D(J)$ as in (29) 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 for $1 \leq n \leq N$

$$t_N(\sigma^{-1}|y_n - \langle Ku_0, \phi_n^*\rangle|, \|\phi_n\|) \leq t_N(0, \|\phi_n\|) + L\sigma^{-1}|y_n - \langle Ku_0, \phi_n^*\rangle| \leq q_N(\alpha).$$

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

$$\lim_{k\to\infty} J(u_k) = \inf_{u\in S} J(u).$$

This shows that $\sup_{k\in\mathbb{N}} J(u_k) =: a < \infty$. Moreover, we find from (9), 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 (8), 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

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

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

In order to prove Bregman-consistency of SMR-estimation in Theorem 3.7, we first establish a basic estimate for the data error.

Lemma A.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 $1 \leq n \leq N$

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

Proof. From Definition 3.4 it becomes clear that

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

for $1 \leq n \leq N$. The convexity of t_N hence implies that

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

By setting $v = (2\sigma)^{-1}\left|\langle Ku^\dagger - K\hat{u}_N(\alpha), \phi_n^*\rangle\right|$ and $r = \|\phi_n\|$ in (9), the above estimate shows that

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

Since $t_N(v, r) \geq \lambda_N(r)$ for all $v \in \mathbb{R}^+$ and $r \in (0, 1]$ (cf. (8)) this implies

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

for $1 \leq n \leq N$. Finally, the assertion follows from Lemma A.1. □

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

Proof of Theorem 3.7. By the definition of the SMRE $\hat{u}_k = \hat{u}_{N_k}(\alpha_k)$, it follows that

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

for all $k \in \mathbb{N}$. Since $\sum_{k=1}^\infty \alpha_k < \infty$, it follows from the Borel-Cantelli Lemma (cf. [49, p 255]) that

$$\mathbb{P}\left(J(\hat{u}_k) > J(u^\dagger) \text{ i.o.}\right) \leq \mathbb{P}\left(T_{N_k}(\varepsilon) > q_{N_k}(\alpha_k) \text{ i.o.}\right) = 0, \tag{30}$$

or in other words

$$\mathbb{P}\left(\exists k_0 \in \mathbb{N} : \ J(\hat{u}_k) \leq J(u^\dagger) \text{ for all } k \geq k_0\right) = 1. \tag{31}$$

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 $\sup_{N\in\mathbb{N}} \operatorname{med}(T_N(\varepsilon)) < \infty$. Hence, it follows from Lemma A.2 and (14) that

$$\sup_{1\leq n\leq N_k}\left|\left\langle Ku^\dagger - K\hat{u}_k, \phi_n^*\right\rangle\right| = \mathcal{O}(\eta_k) \quad \text{a.s.} \tag{32}$$

as $k \to \infty$ which proves (16). In particular, (32) and the fact hat $N_k > N_0$ imply

$$\sup_{k\in\mathbb{N}}\sup_{1\leq n\leq N_0}\left|\langle K\hat{u}_k, \phi_n^*\rangle\right| =: b < \infty \quad \text{a.s.}$$

Summarising, we find that $\hat{u}_k \in \Lambda(a + b)$ which is sequentially weakly pre-compact according to Assumption 3.5 (ii). Choose a sub-sequence indexed by $\rho(k)$ with weak limit $\hat{u} \in U$. Since $N_k \to \infty$ as $k \to \infty$ it follows from (32) and (14) that

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

Since we assumed that $g \in \overline{\mathrm{span}\Phi}$ this shows that $K\hat{u} = g$. Furthermore, according to (31) 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

$$J(\hat{u}) \leq \liminf_{k\to\infty} J(\hat{u}_{\rho(k)}) \leq \limsup_{k\to\infty} J(\hat{u}_{\rho(k)}) \leq J(u^\dagger).$$

Since $u^\dagger$ is a J-minimising solution of (1) we conclude that the same holds for $\hat{u}$ and that

$$J(\hat{u}) = J(u^\dagger) = \lim_{k\to\infty} J(\hat{u}_{\rho(k)}).$$

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

$$\lim_{k\to\infty} J(\hat{u}_k) = J(u^\dagger) \quad \text{a.s.} \tag{33}$$

We next prove that

$$\lim_{k\to\infty} D_J(u^\dagger, \hat{u}_k) = 0 \quad \text{a.s.}$$

To this end, recall (30), i.e. almost surely there 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

$$\mathbb{P}\left(t_{N_k}(\left|\varepsilon(\phi_n^*)\right|, \|\phi_n^*\|) = q_{N_k}(\alpha_k)\right) = 0$$

for all $n \in \mathbb{N}$ and thus

$$\mathbb{P}\left(\exists k_0 : \ T_{N_k}(\varepsilon) < q_{N_k}(\alpha_k) \text{ for all } k \geq k_0\right) = 1. \tag{34}$$

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 (11) to

$$\hat{u}_k \in \operatorname*{argmin}_{u\in U} J(u) + G_k(Ku).$$

From (34) it follows that $u^\dagger$ lies in the interior of the admissible set of the convex problem (11). In other words, the functionals G_k are continuous at $Ku^\dagger$ for k large enough. Therefore we can apply [29, Chap. II Prop. 4.1] (cf. also Chapter II, Remark 4.2 therein) and choose an element $\xi_k \in V$ such that

$$K^*\xi_k \in \partial J(\hat{u}_k) \quad \text{ and } \quad -\xi_k \in \partial G_k(K\hat{u}_k).$$

The second inclusion and the definition of the sub-gradient show that for all $u \in U$

$$G_k(Ku) \geq G_k(\hat{u}_k) - \langle \xi_k, Ku - K\hat{u}_k \rangle = \langle K^*\xi_k, \hat{u}_k - u \rangle .$$

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

$$0 \geq \left\langle K^*\xi_k, \hat{u}_k - u^\dagger \right\rangle .$$

Since $J(\hat{u}_k) \to J(u^\dagger)$ a.s. (cf. (33)) as $k \to \infty$ we find

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

This proves (15). □

It remains to prove the convergence rate results in Theorem 3.10. 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 3.8. With this we get the following basic estimate.

Lemma A.3. Assume that Assumption 3.8 holds and let $N \geq N_0$ and $\alpha \in (0,1)$. Then,

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

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

Proof. From Assumption 3.8 we find that

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

From Lemma A.2 it follows that

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

which shows the assertion. □

Combination of the auxiliary result in Lemma A.3 with Theorem 3.7 paves the way to the proof of Theorem 3.10.

Proof of Theorem 3.10. First, observe that Assumption 3.8 and (19) imply (14), that is, all assumptions in Theorem 3.7 are satisfied. Therefore $\{\hat{u}_k\}_{k\in\mathbb{N}}$ is bounded almost surely and due to the continuity of K we find

$$
\sup_{k\in\mathbb{N}} \left\|K\hat{u}_k - Ku^\dagger\right\| < \infty \quad \text{a.s.}
$$

After setting $B := \sup_{N\in\mathbb{N}} \sum_{n=1}^{N} |b_{n,N}|$, which is finite according to Assumption 3.8, it follows from Lemma A.3 and (19) that

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

for a suitably chosen constant $C > 0$. Since $\sup_{N\in\mathbb{N}} T_N(\varepsilon) < \infty$ almost surely, it follows that

$$\sup_{N\in\mathbb{N}} \tilde{T}_N(\varepsilon) = \sup_{N\in\mathbb{N}} \left(T_N(\varepsilon) + \operatorname{med}(T_N(\varepsilon))\right) < \infty \qquad \text{a.s.}$$

Combining the previous two estimates shows

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

Next, recall from (31) in the proof of Theorem 3.7 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

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

for $k \geq k_0$. This proves the first estimate in (20). The second estimate follows directly from Lemma A.2. □

A.2. **Technical results.** 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 A.4. Let $X \subset \mathbb{R}^d$ be compact and convex.

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

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

(2) A function $g : X \to \mathbb{R}$ 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 $g : X \subset \mathbb{R}^d \to \mathbb{R}$ by piecewise constant functions in terms of the modulus of continuity.

Lemma A.5. Let $X \subset \mathbb{R}^d$ be a compact and convex set and $\{A_1, A_2, \ldots\}$ 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 $n_l < i < j \leq n_{l+1}$ one has $\lambda_d(A_i \cap A_j) = 0$.
(ii) and

$$X = \bigcup_{j=n_l+1}^{n_{l+1}} A_j$$

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

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

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

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

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

Next, we set for $m \in \mathbb{N}$ and $1 \leq l \leq m$

$$a_{lm} = \frac{\omega^{-2}(\delta_l, g)}{\sum_{\nu=0}^m \omega^{-2}(\delta_\nu, g)} \in (0,1).$$

Note, that for all $m \in \mathbb{N}$ one has $\sum_{0\leq l\leq m} a_{lm} = 1$. With this, we define for $0 \leq l \leq m$ and $n_l < j \leq n_{l+1}$ the coefficients $b^m_{j,l} = (a_{lm} \int_{A_j} g(\tau)\,\mathrm{d}\lambda_d(\tau))/\lambda_d(A_j)$. Since we assumed that g is continuous on the compact set X, it follows that $\left|b^m_{j,l}\right| \leq \|g\|_\infty a_{lm}$ and hence

$$\sum_{l=0}^{m} \sum_{j=n_l+1}^{n_{l+1}} \left|b^m_{j,l}\right| \leq \|g\|_\infty \quad \text{for all } m \in \mathbb{N}.$$

Moreover, we have for all $s \in X$ that

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

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

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

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

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

□

For the remainder of this section we collect some results concerning the capacity (cf. Definition 4.7) of (subsystems of) the set Φ_d of indicator functions on convex and closed sets in $[0,1]^d$ with $d \geq 1$.

Remark A.6. From a practical point of view, it is often more convenient to express (27) 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 (cf. [52, p.98]) that for all $\varepsilon > 0$

$$N(\varepsilon, T) \leq D(\varepsilon, T) \leq N(\varepsilon/2, T). \tag{35}$$

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

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

The entire set Φ_d is too large in order to render the test-statistic T_N in (26) finite: it was shown in [11] (see also [26, Chap. 8.4]) that the ε-covering number of Φ_d of all nonempty, closed and convex sets contained in the unit ball $\left\{x \in \mathbb{R}^d \ : \ |x| \leq 1\right\}$ is of the same order as $\exp(\varepsilon^{(1-d)/2})$ (for $d \geq 2$) as $\varepsilon \to 0^+$. This proves that there cannot exist any constants A, B and γ such that (27) holds with $\Phi = \Phi_d$.

For particular classes of convex sets, however, entropy estimates as in (27) are at hand. The collection Φ_r of indicator functions on d-dimensional rectangles in $[0,1]^d$ constitutes such an example:

Proposition A.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 [52, 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 Vapnik-Červonenkis (VC)-index of the collection 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 $2d+1$ (cf. [52, Ex. 2.6.1]). □

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

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

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

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

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

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

$$A^{-1}u^{-2}\delta^{-2} \leq D(u\delta, \{\phi \in \Phi_2 \ : \ \|\phi\| \leq \delta\}) \leq Au^{-2}\delta^{-2}$$

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

References

[1] Robert Acar and Curtis Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10:1217–1229, 1994.

[2] Robert A. Adams. *Sobolev Spaces*, volume 65 of *Pure and Applied Mathematics*. Academic Press, New York - London, 1975.

[3] Viorel Barbu. *Nonlinear semigroups and differential equations in Banach spaces*. Editura Academiei Republicii Socialiste Romnia, Bucharest, 1976.

[4] Viorel Barbu and Teodor Precupanu. *Convexity and optimization in Banach spaces*. Editura Academiei, Bucharest, revised edition, 1978. Translated from the Romanian.

[5] 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.

[6] Nicolai Bissantz, Thorsten Hohage, and Axel Munk. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20(6):1773–1789, 2004.

[7] Nicolai Bissantz, Bernard Mair, and Axel Munk. A statistical stopping rule for mlem reconstructions in pet. *IEEE Nucl. Sci. Symp. Conf. Rec.*, 8:4198–4200, 2008.

[8] J. M. Borwein and A. S. Lewis. Convergence of best entropy estimates. *SIAM J. Optim.*, 1(2):191–205, 1991.

[9] James P. Boyle and Richard L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in order restricted statistical inference (Iowa City, Iowa, 1985)*, volume 37 of *Lecture Notes in Statist.*, pages 28–47. Springer, Berlin, 1986.

[10] 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.

[11] E. M. Bronšteĭn. ε-entropy of convex sets and functions. *Sibirsk. Mat. Ž.*, 17(3):508–514, 715, 1976.

[12] M. Burger, E. Resmerita, and L. He. Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2-3):109–135, 2007.

[13] Martin Burger, Klaus Frick, Stanley Osher, and Otmar Scherzer. Inverse total variation flow. *Multiscale Model. Simul.*, 6(2):365–395 (electronic), 2007.

[14] Martin Burger and Stanley Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004.

[15] 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.

[16] Laurent Cavalier and Alexandre Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002.

[17] Pao-Liu Chow, Ildar A. Ibragimov, and Rafail Z. Khasminskii. Statistical approach to some ill-posed problems for linear partial differential equations. *Probab. Theory Related Fields*, 113(3):421–441, 1999.

[18] Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.

[19] Albert Cohen, Marc Hoffmann, and Markus Reiß. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42(4):1479–1501 (electronic), 2004.

[20] M Collins, R E Schapire, and Y Singer. Logistic regression, adaboost and bregman distances. *Mach. Learn*, 48(48):253–285, 2002.

[21] Imre Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, 19(4):2032–2066, 1991.

[22] 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.

[23] P. L. Davies, A. Kovac, and M. Meise. Nonparametric regression, confidence regions and regularization. *Ann. Statist.*, 37(5B):2597–2625, 2009.

[24] David L. Donoho. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Different perspectives on wavelets (San Antonio, TX, 1993)*, volume 47 of *Proc. Sympos. Appl. Math.*, pages 173–205. Amer. Math. Soc., Providence, RI, 1993.

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

[26] R. M. Dudley. *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.

[27] Lutz Dümbgen and Vladimir G. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152, 2001.

[28] Lutz Dümbgen and Günther Walther. Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785, 2008.

[29] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, volume 1 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-Oxford, 1976.

[30] Heinz Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1996.

[31] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[32] Klaus Frick. *The Augmented Lagrangian Method and Related Evolution Equations*. Phd-thesis, University of Innsbruck, 2008.

[33] Klaus Frick and Otmar Scherzer. Regularization of ill-posed linear equations by the nonstationary Augmented Lagrangian Method. *J. Integral Equations Appl.*, 22(2):217–258, 2010.

[34] Alexander Goldenshluger and Sergei V. Pereverzev. On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli*, 9(5):783–807, 2003.

[35] Walter Hinterberger, Michael Hintermüller, Karl Kunisch, Markus von Oehsen, and Otmar Scherzer. Tube methods for BV regularization. *J. Math. Imaging Vision*, 19(3):219–235, 2003.

[36] Iain M. Johnstone. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica*, 9(1):51–83, 1999.

[37] J. Lafferty, S. Pietra, and V. Pietra. Statistical learning algorithms based on bregman distances. In *Proceedings of 1997 Canadian workshop on information theory*, pages 77–80, 1997.

[38] Bernard A. Mair and Frits H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5):1424–1444, 1996.

[39] Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997.

[40] Peter Mathé and Sergei 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.

[41] Peter Mathé and Sergei V. Pereverzev. Discretization strategy for linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(6):1263–1277, 2003.

[42] Peter Mathé and Sergei V. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789–803, 2003.

[43] James R. McLaughlin. Absolute convergence of series of Fourier coefficients. *Trans. Amer. Math. Soc.*, 184:291–316, 1973.

[44] Michael Nussbaum and Sergei Pereverzev. The degree of ill-posedness in stochastic and deterministic noise models. Technical Report 509, WIAS, 1999. Preprint.

[45] Finbarr 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.

[46] Elena Resmerita. On total convexity, Bregman projections and stability in Banach spaces. *J. Convex Anal.*, 11(1):1–16, 2004.

[47] Elena Resmerita. Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Problems*, 21(4):1303–1314, 2005.

[48] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, and Frank Lenzen. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[49] 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.

[50] David Siegmund and Benjamin Yakir. Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213, 2000.

[51] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

[52] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

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

[54] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

[55] William P. Ziemer. *Weakly differentiable functions*. Springer Verlag, New York, 1989.

[56] 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.

Institut für Mathematische Stochastik, University of Göttingen