ÖAW
Austrian Academy
of Sciences

Doctoral Program
Computational Mathematics
Numerical Analysis and Symbolic Computation

JKU
JOHANNES KEPLER
UNIVERSITY LINZ

A stochastic convergence analysis for Tikhonov regularization with sparsity constraints

Daniel Gerth Ronny Ramlau

DK-Report No. 2013-08 October 2013

A–4040 LINZ, ALTENBERGERSTRASSE 69, AUSTRIA

Supported by

Austrian Science Fund (FWF) Upper Austria

FWF
Der Wissenschaftsfonds.

LAND
OBERÖSTERREICH

DK sponsors:

- **Johannes Kepler University Linz** (JKU)
- **Austrian Science Fund** (FWF)
- **Upper Austria**

A Stochastic Convergence Analysis for Tikhonov Regularization with Sparsity Constraints

Daniel Gerth and Ronny Ramlau

Abstract

In this paper we investigate convergence properties of Tikhonov regularization for linear ill-posed problems under a stochastic error model. Namely, we assume that we are given a finite amount of measurements, each contaminated by Gaussian noise with zero mean and known finite variance. Using Besov-space penalty terms to promote sparse solutions with respect to a preassigned wavelet basis, the Ky-Fan metric allows us to lift deterministic convergence results into the stochastic setting. In particular, we formulate a general convergence theorem and propose a formula to directly calculate a suitable regularization parameter. This immediately leads to convergence rates. Numerical examples are presented to verify the theoretical results.

1 Introduction

We study the solution of the linear ill-posed problem

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{1}$$

with $\mathbf{A} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are (in general infinite dimensional) Hilbert spaces. In practice $\mathbf{y}$ is only available via a finite amount of measurements which are additionally corrupted by measurement noise. Since computers can handle only finite dimensional quantities, (1) has to be approximated by an equation

$$Ax + \epsilon = y^\delta \tag{2}$$

where $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, $x \in \mathbb{R}^n$, $y^\delta \in \mathbb{R}^m$ and ϵ denotes the measurement error. Here $m \in \mathbb{N}$ and $m \in \mathbb{N}$ denote the level of discretization of measurements $\mathbf{y}$ and unknown $\mathbf{x}$. Due to the ill-posedness of the problem, already small errors in the data may lead to computed solutions far off the correct one, rendering straight-forward approaches to solve (2) useless. To deal with this phenomenon, regularization methods have to be applied to ensure stability of the solutions x with respect to the data y^δ. In order to design these methods appropriately, it is crucial to use available information about the noise ϵ. A deterministic assumption is the worst case estimate between the true and measured data $||y - y^\delta|| = ||\epsilon|| \leq \delta$, $\delta > 0$. Problems of type (1) and (2) under this error bound have been studied intensively in literature, see for example [9], [13] or [18]. It is also possible to use a more explicit model for ϵ, e.g. assuming a certain random distribution of ϵ. For an overview on this stochastic setting we refer to [14]. When dealing with stochastic noise, we will always denote the perturbed data by y^σ instead of y^δ. In recent years, sparse regularization emerged as a powerful tool to find a solution of (2). In this framework, it is assumed that the unknown x can be approximated well with only few coefficients of its expansion with respect to a preassigned basis or frame in $\mathcal{X}$. Pushed by the seminal paper [6], where the authors studied the so-called iterative soft-shrinking algorithm to calculate the minimizer of a special Tikhonov-functional under the deterministic error assumption, sparsity has become a widely used regularization strategy with generalizations to, for example, Banach spaces [3] or nonlinear operators [22]. Sparse regularization also gained attention in the stochastic setting, c.f. [15, 16, 23]. However, research in the two fields seems to develop rather independently. In this paper we seek to bridge this gap and connect deterministic and stochastic results in the linear Hilbert space setting. For classical Tikhonov regularization this has been done before in [20], where the authors considered a Gaussian error model. The paper is organized as follows. A more precise statement of the problem (1) and the discrete model (2) is given in Section 2. There we build up the connection between the maximum a-priori solution for stochastic, so called Bayesian, inversion with Besov space priors and the deterministic Tikhonov-type regularization with a Besov space penalty. These Besov spaces priors, which stabilize the reconstructions, are discussed in more detail in Section 3. The Ky Fan metric which we use to measure convergence of the random variables is introduced in Section 4. A general convergence theorem is then given in Section 5. Convergence rates connected with a particular parameter choice rule are given in

Section 6. Numerical examples are presented in Section 7 to exemplify the theory.

2 Statement of the problem

We start again with equation (1) and introduce a linear orthogonal projection operator $P_m : \mathcal{Y} \to \mathbb{R}^m$, modelling the mapping of the (possibly infinite dimensional) object $\mathbf{y}$ on an m-dimensional vector y. The projection depends on the actual measurement device. One might for example think of measured function values or coefficients with respect to certain basis functions in $\mathcal{Y}$. As before $\epsilon \in \mathbb{R}^m$ denotes the typically unavoidable measurement noise. We thus have the *practical measurement model*

$$P_m\mathbf{y} = P_m\mathbf{A}\mathbf{x} + \epsilon.$$

Throughout this paper we assume that each component of the error is normally distributed with zero mean and variance σ^2, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1 \dots m$, $\sigma > 0$. This especially means that arbitrarily large errors will be allowed, but with low probability. Let $\{\psi_\lambda : \lambda \in \Lambda\}$ be an orthonormal basis in $\mathcal{X}$, where Λ is an appropriate index set. In order to characterize the unknown $\mathbf{x}$ by its coefficients with respect to $\{\psi_\lambda : \lambda \in \Lambda\}$, we introduce a second operator

$$T : \mathcal{X} \to \ell_2 \quad \text{via} \quad \mathbf{x} \mapsto \{\langle \mathbf{x}, \psi_\lambda \rangle\}_{\lambda \in \Lambda}. \tag{3}$$

T and its adjoint T^*,

$$T^* : \ell_2 \to \mathcal{X} \quad \text{via} \quad \mathbf{g} \mapsto \sum_{\lambda \in \Lambda} \mathbf{g}_\lambda \psi_\lambda \tag{4}$$

allow us to switch between function $\mathbf{x}$ and the coefficients $\mathbf{x}_\lambda := \langle \mathbf{x}, \psi_\lambda \rangle$. Here $\langle \cdot, \cdot \rangle$ denotes the L_2-inner product. Computations on a computer require a finite dimensional representation of $\mathbf{x}$. Therefore we restrict the index set Λ to a finite set Λ_n, where $n \in \mathbb{N}$ is the number of basis functions used for the discretization. The truncated projectors with respect to Λ_n are defined analogously to (3) and (4), respectively. We have

$$\begin{aligned} T_n : \mathcal{X} \to \ell_2, &\quad \mathbf{x} \mapsto \{\langle \mathbf{x}, \psi_\lambda \rangle\}_{\lambda \in \Lambda_n}, \\ T_n^* : \ell_2 \to \mathcal{X}, &\quad \mathbf{g} \mapsto \sum_{\lambda \in \Lambda_n} \mathbf{g}_\lambda \psi_\lambda. \end{aligned} \tag{5}$$

Although our theory can be expanded to frames instead of a basis in $\mathcal{X}$, we restrict ourselves to the latter case for simplicity. Thus we arrive at the *computational model*

$$P_m\mathbf{y} = P_m\mathbf{A}T_n^*T_n\mathbf{x} + \epsilon. \tag{6}$$

Due to the stochastic noise assumption we chose a Bayesian approach for the solution of (6). For a detailed introduction to Bayesian inversion theory see for example [14]. In this framework, all occurring quantities are treated as random variables, even if some of them might be deterministic. To simplify the notation we denote $y := P_m\mathbf{y}$, $y^\sigma := y + \epsilon$, $x := T_n\mathbf{x}$, $A := P_m\mathbf{A}T_n^*$ and obtain the linear model

$$y^\sigma = Ax + \epsilon \tag{7}$$

where the variables x, y^σ and ϵ are realizations of the corresponding random variables in the equation

$$Y^\sigma = AX + \mathcal{E}. \tag{8}$$

Here $X(x,\omega), Y(y,\omega)$ and $\mathcal{E}(y,\omega)$ are random functions from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively and $\omega \in \Omega$. In this notation, the sample space Ω is a set of outcomes of the stochastic process, $\mathcal{F}$ the corresponding σ-algebra and $\mathbb{P}$ a probability measure, $\mathbb{P} : (\Omega, \mathcal{F}) \to [0,1]$. In the Bayesian framework the solution of the inverse problem is given as a distribution of the random variable of interest, the *posterior distribution* π_{post}, determined by Bayes formula

$$\pi_{post}(x|y^\sigma) = \frac{\pi_{pr}(x)\pi_\epsilon(y^\sigma|x)}{\pi_{y^\sigma}(y^\sigma)}. \tag{9}$$

However, for practical reasons one is usually more interested in finding a single representation as solution instead of the distribution itself. Popular choices are the *conditional expectation* $\mathbb{E}(\pi_{post}(x|y^\sigma)) = \int x\pi_{post}(x|y^\sigma)dx$ and the *maximum a-posteriori* (MAP) solution

$$x^{\text{MAP}} = \underset{x\in\mathbb{R}^n}{\text{argmax}} \quad \pi_{post}(x|y^\sigma), \tag{10}$$

i.e. the most likely value for x. Both methods have certain advantages and disadvantages, see for example [14, 15, 16]. In this paper we will only consider the maximum a-posteriori approach because it essentially leads to a Tikhonov-type minimization problem. Since neither π_{y^σ} nor the normalization constants of the distributions influence the position of the maximum in

(10) they can be neglected further on. The *likelihood function* $\pi_\epsilon(y^\sigma|x)$ in (9) represents the model for the measurement noise. Because of the normally distributed error we simply have

$$\pi_\epsilon \propto \exp\left\{-\frac{1}{2\sigma^2}||Ax - y^\sigma||^2\right\}.$$

Available a priori information about the unknown solution is expressed via the *prior distribution* π_{pr}. In classical Bayesian inversion theory (cf. [12, 14, 20]) a Gaussian prior distribution is assumed, leading to an explicit representation of the posterior distribution (see for example [14, Theorem 3.7]). However, we want to use a prior promoting solutions which are sparse with respect to the basis $\{\psi_\lambda : \lambda \in \Lambda\}$. In the deterministic case, where usually Tikhonov-type functionals

$$||\mathbf{A}\mathbf{x} - \mathbf{y}^\delta||^2_{\mathcal{Y}} + \hat{\alpha}\Phi_{\mathbf{w},p}(\mathbf{x}) \tag{11}$$

are minimized, it is known that penalties of the form

$$\Phi_{\mathbf{w},p}(\mathbf{x}) = \sum_{\lambda\in\Lambda} w_\lambda |\langle \mathbf{x}, \psi_\lambda\rangle|^p, \qquad 1 \le p < 2,$$

$\mathbf{w} = \{w_\lambda\}_{\lambda\in\Lambda}$ with $w_\lambda \geq c > 0\ \forall\lambda \in \Lambda$, indeed lead to sparse reconstructions, i.e. the amount of nonzero coefficients $\langle \mathbf{x}, \psi_\lambda\rangle$, $\lambda \in \Lambda$, is small [6, 21]. A particular choice for $\Phi_{\mathbf{w},p}$ are Besov space norms which have already been used as sparsity constraints in (11), see for example [6] or [21]. Additionally it has been shown in [15, 16] that Besov priors are *discretization invariant*, i.e. the representation of a priori knowledge remains the same for all discretization parameters n. As a counterexample, it has been shown in [17] that discrete (non-Gaussian) total variaton priors converge to a smooth Gaussian prior the more the level of discretization is refined. We recall that the Besov space $B^s_{p,q}(\mathbb{R}^d)$ is a function space on $\mathbb{R}^d$ consisting of, roughly spoken, functions which have s derivatives in $L_p(\mathbb{R}^d)$, where q provides some additional finetuning. For example, $B^s_{2,2}$ coincides with the Sobolev spaces H^s. In this paper, we will consider the case $p = q$ only and write $B^s_p(\mathbb{R}^d)$ instead of $B^s_{p,p}(\mathbb{R}^d)$. We refer to Section 3 for a more detailed characterization of Besov spaces and their representation via wavelet expansions, and to [5, 19] for deeper analysis. In the discrete setting (7) we may formally define the prior distribution

$$\pi_{pr}(x) \propto \exp\left(-\frac{\alpha}{2}||T^*_n x||^p_{B^s_p(\mathbb{R}^d)}\right), \tag{12}$$

where α is an additional tuning or regularization parameter. Bayes' formula (9) now yields

$$\pi_{post}(x|y^\sigma) \propto \exp\left(-\frac{1}{2\sigma^2}||Ax-y^\sigma||^2\right)\exp\left(-\frac{\alpha}{2}||T_n^* x||^p_{B^s_p(\mathbb{R}^d)}\right).$$

The maximum a posteriori solution is given by

$$x^{\mathrm{MAP}} = \operatorname*{argmax}_{x\in\mathbb{R}^n} \quad \exp\left(-\frac{1}{2\sigma^2}||Ax-y^\sigma||^2 - \frac{\alpha}{2}||T_n^* x||^p_{B^s_p(\mathbb{R}^d)}\right),$$

or equivalently

$$x^{\mathrm{MAP}} = \operatorname*{argmin}_{x\in\mathbb{R}^n} \quad \frac{1}{2\sigma^2}||Ax-y^\sigma||^2 + \frac{\alpha}{2}||T_n^* x||^p_{B^s_p(\mathbb{R}^d)}.$$

Setting $\hat{\alpha} = \alpha\sigma^2$ we arrive at

$$x^{\mathrm{MAP}}_{\hat{\alpha}} = \operatorname*{argmin}_{x\in\mathbb{R}^n} \quad ||Ax-y^\sigma||^2 + \hat{\alpha}||T_n^* x||^p_{B^s_p(\mathbb{R}^d)}, \tag{13}$$

which is exactly a discretized version of the Tikhonov functional (11) known from the deterministic setting. Consequently, the same techniques can be used to calculate the minimizer. However, since the deterministic theory is based on error bounds $||Ax - y^\delta|| \leq \delta$, convergence results do not immediately apply to our situation. To overcome this issue we will use a specific metric which essentially allows to combine deterministic results with the stochastic background. It will be important to carefully distinguish between the parameter α originating from the prior distribution (12) and the actual regularization parameter $\hat{\alpha} = \alpha\sigma^2$ in the Tikhonov functional (11).

3 Besov Spaces and Random Variables

3.1 Wavelet representation of Besov spaces

In Section 2 we formally introduced the Besov space prior

$$\pi_{pr}(x) \propto exp(-\frac{\alpha}{2}||T_n^* x||^p_{B^s_p(\mathbb{R}^d)})$$

where $1 \leq p \leq 2$ is an integrability parameter and $s \in \mathbb{R}$ describes the estimated smoothness of the solution. In order to characterize Besov spaces by the coefficients of the wavelet expansion of its functions, we follow Daubechies et. al [6] to construct a sufficiently smooth wavelet basis for $B^s_p(\mathbb{R})$. Let ϕ be a scaling function and ψ a compactly supported wavelet suitable for multi resolution analysis of smoothness $\tilde{s} > s$ in $L_2(\mathbb{R})$, i.e. $\phi \in C^{\tilde{s}}(\mathbb{R})$ and $\psi \in C^{\tilde{s}}(\mathbb{R})$. Define

$$\phi_{j,k}(t) = 2^{\frac{j}{2}}\phi(2^j t - k), \quad \psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k), \quad j,k \in \mathbb{Z}.$$

The functions ϕ and ψ are assumed to be chosen suitably, fulfilling

$$\overline{\mathrm{Span}\{\phi_{j,k} : k \in \mathbb{Z}\}} \oplus \overline{\mathrm{Span}\{\psi_{j,k} : k \in \mathbb{Z}\}} = \overline{\mathrm{Span}\{\phi_{j+1,k} : k \in \mathbb{Z}\}} \tag{14}$$

for all $j \in \mathbb{Z}$ and

$$\mathrm{Span}\{\phi_{0,k} : k \in \mathbb{Z}\} \oplus \bigoplus_{j \geq 0} \mathrm{Span}\{\psi_{j,k} : k \in \mathbb{Z}\} = L_2(\mathbb{R}). \tag{15}$$

Following Meyer [19] we expand this to a wavelet basis in $\mathbb{R}^d$. Let E denote the set of all $2^d - 1$ sequences $\nu = (\nu_1, \nu_2, \dots, \nu_d)$ with $\nu_j \in \{0,1\}$ for all $j = 1, \dots, d$ and $\sum_j \nu_j > 0$. For $\nu \in E$ and $j \in \mathbb{Z}$ we define the tensor-wavelets

$$\psi^{\nu}_{j,k}(t) := 2^{dj/2}\psi^{\nu_1}(2^j t_1 - k_1) \dots \psi^{\nu_d}(2^j t_d - k_d)$$

with the convention that $\psi^0 = \phi$, $\psi^1 = \psi$ and vectors $k = (k_1, k_2, \dots, k_d) \in \mathbb{Z}^d$. The $\psi^{\nu}_{j,k}$ constitute an orthonormal basis for $\mathbb{R}^d$. Let

$$\phi^d_{j,k}(t) := 2^{dj/2}\phi(2^j t_1 - k_1) \dots \phi(2^j t_d - k_d).$$

Then, analogously to (14) and (15),

$$\overline{\mathrm{Span}\{\phi^d_{j,k} : k \in \mathbb{Z}\}} \oplus \overline{\mathrm{Span}\{\psi^{\nu}_{j,k} : k \in \mathbb{Z}\}} = \overline{\mathrm{Span}\{\phi^d_{j+1,k} : k \in \mathbb{Z}\}}\ \forall j \in \mathbb{Z},$$
$$\mathrm{Span}\{\phi^d_{0,k} : k \in \mathbb{Z}\} \oplus \bigoplus_{j \geq 0} \mathrm{Span}\{\psi^{\nu}_{j,k} : k \in \mathbb{Z}, \nu \in E\} = L_2(\mathbb{R}^d). \tag{16}$$

With this the function $\mathbf{x}$ has the form

$$\mathbf{x} = \sum_{k \in \mathbb{Z}} \langle \mathbf{x}, \phi^d_{0,k} \rangle \phi^d_{0,k} + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}} \langle \mathbf{x}, \psi^{\nu}_{j,k} \rangle \psi^{\nu}_{j,k}. \tag{17}$$

To simplify the notation we again follow [6] and denote the set of all functions $\phi^d_{0,k}$ and $\psi^{\nu}_{j,k}$, $j = 1, 2, \dots$, $k \in \mathbb{Z}$, $\nu \in E$, in (17) by Ψ_λ. The $\Psi_\lambda = \{\psi_\lambda : \lambda \in$

$\Lambda\}$ are not only an orthonormal basis in $L_2(\mathbb{R}^d)$ but also a (Riesz) basis for other function spaces including the Besov spaces. Let $\varsigma = s + d\left(\frac{1}{2} - \frac{1}{p}\right) \geq 0$ to ensure $B^s_p(\mathbb{R}^d)$ is a subset of $L_2(\mathbb{R}^d)$. Using the convention $|\lambda| = j$ to denote the scale of the wavelets, the norm

$$||\mathbf{x}||_{B^s_p(\mathbb{R}^d)} = \left(\sum_{\lambda\in\Lambda} 2^{\varsigma p|\lambda|}|\langle \mathbf{x}, \Psi_\lambda\rangle|^p\right)^{\frac{1}{p}} \tag{18}$$

is equivalent the traditional Besov space norm [6]. However, in our framework we have to restrict the index set Λ such that on each scale we have only finitely many wavelets. This is guaranteed by the following assumptions on the wavelet expansion of $\mathbf{x}$:

- $\exists k^-_\phi, k^+_\phi \in \mathbb{Z}$, $k^-_\phi < k^+_\phi$: $\langle \mathbf{x}, \phi^d_{0,k}\rangle = 0$ for all $k < k^-_\phi$ and $k > k^+_\phi$. Define $\ell_\phi := k^+_\phi - k^-_\phi + 1$.
- On each scale $j \geq 0$, $\exists k^{j-}_\psi, k^{j+}_\psi \in \mathbb{Z}$, $k^{j-}_\psi \leq k^{j+}_\psi$: $\langle \mathbf{x}, \phi^d_{0,k}\rangle = 0$ for all $k < k^{j-}_\psi$ and $k > k^{j+}_\psi$. Define $\ell^j_\psi := k^{j+}_\psi - k^{j-}_\psi + 1$.
- There exists $\ell_\psi \in \mathbb{N}$ such that $\ell^j_\psi \leq 2^{jd}\ell_\psi$ for all $j \geq 0$.

These assumptions are for example satisfied for compactly supported functions or functions which are truly sparse, i.e. the number of nonzero inner products in (17) is finite. Thus (17) reads

$$\mathbf{x} = \sum_{k^-_\phi \leq k \leq k^+_\phi} \langle \mathbf{x}, \phi^d_{0,k}\rangle \phi^d_{0,k} + \sum_{j=0}^{\infty} \sum_{k^{j-}_\psi \leq k \leq k^{j+}_\psi} \langle \mathbf{x}, \psi^\nu_{j,k}\rangle \psi^\nu_{j,k}. \tag{19}$$

We denote the corresponding index set by Λ_f. Hence $\mathbf{x} = \sum_{\lambda\in\Lambda_f} \langle \mathbf{x}, \psi_\lambda\rangle \psi_\lambda$.

3.2 Random variables in Besov spaces

The wavelet basis Ψ_λ allows the following definition of random variables in Besov spaces, adapted from [16].

Definition 3.1. *Consider functions on $\mathbb{R}^d$, $d \in \mathbb{N}$. Let $1 \leq p < \infty$ and Λ_f as above. Take $s \in \mathbb{R}$ such that $\varsigma := s + d(\frac{1}{2} - \frac{1}{p}) > 0$. Let $(X^\alpha_\lambda)_{\lambda\in\Lambda_f}$ be independent identically distributed real-valued random variables with probability* *density function*

$$\pi_{X_\lambda^\alpha}(\tau) = c_p^\alpha \exp(-\frac{\alpha|\tau|^p}{2}), \quad \tau \in \mathbb{R}, \quad c_p^\alpha = \left(\frac{\alpha}{2}\right)^{\frac{1}{p}} \frac{p}{2\Gamma(\frac{1}{p})}. \tag{20}$$

Let $\mathbf{X}$ *be the random function*

$$\mathbf{X}(t) = \sum_{\lambda \in \Lambda_f} 2^{-\varsigma|\lambda|} X_\lambda^\alpha \psi_\lambda(t), \quad t \in \mathbb{R}^d.$$

Then we say $\mathbf{X}$ *is distributed according to a* B_p^s*-prior.*

The following Lemma characterizes the random variables $|X_\lambda^\alpha|^p$ on which the stochastic properties of $||\mathbf{X}||$ essentially depend.

Lemma 3.2. *Let* X_λ^α *be defined as in Definition 3.1. Then the random variables* $|X_\lambda^\alpha|^p$*,* $1 \leq p \leq 2$*, are distributed according to the probability density function*

$$\pi_{|X_\lambda^\alpha|^p}(\eta) = \left(\frac{\alpha}{2}\right)^{\frac{1}{p}} \frac{\eta^{\frac{1}{p}-1}}{\Gamma(\frac{1}{p})} \exp(-\frac{\alpha\eta}{2}), \quad \eta \geq 0 \tag{21}$$

and satisfy

$$\mathbb{E}\left(|X_\lambda^\alpha|^p\right) = \frac{2}{\alpha p}. \tag{22}$$

Proof. Let X_λ^α be defined as in (20). We are interested in the probability density of $Y := |X_\lambda^\alpha|^p$. Denote $\mathbb{F}_{X_\lambda^\alpha}(\tau)$ and $\mathbb{F}_Y(\eta)$ the cumulative distribution functions of X_λ^α and Y, respectively. Since $Y \geq 0$, also $\eta \geq 0$. Hence

$$\begin{aligned} \mathbb{F}_Y(\eta) &= \mathbb{P}(Y \leq \eta) = \mathbb{P}(|X_\lambda^\alpha|^p \leq \eta) = \mathbb{P}(-\sqrt[p]{\eta} \leq X_\lambda^\alpha \leq \sqrt[p]{\eta}) \\ &= \mathbb{F}_{X_\lambda^\alpha}(\sqrt[p]{\eta}) - \mathbb{F}_{X_\lambda^\alpha}(-\sqrt[p]{\eta}) \end{aligned}$$

and since $\sqrt[p]{\cdot}$ is continuously differentiable on $[0, \infty)$ for all $1 \leq p \leq 2$,

$$\begin{aligned} \pi_Y(\eta) &= \tfrac{d}{d\eta}\mathbb{F}_Y(\eta) = \tfrac{d}{d\eta}\left(\mathbb{F}_{X_\lambda^\alpha}(\sqrt[p]{\eta}) - \mathbb{F}_{X_\lambda^\alpha}(-\sqrt[p]{\eta})\right) \\ &= \pi_{X_\lambda^\alpha}(\sqrt[p]{\eta}) \cdot \left(\tfrac{1}{p}\eta^{1-\frac{1}{p}}\right) - \pi_{X_\lambda^\alpha}(-\sqrt[p]{\eta}) \cdot \left(-\tfrac{1}{p}\eta^{1-\frac{1}{p}}\right) \\ &= \left(\tfrac{\alpha}{2}\right)^{\frac{1}{p}} \tfrac{\eta^{\frac{1}{p}-1}}{\Gamma(\frac{1}{p})} \exp\left(-\tfrac{\alpha\eta}{2}\right), \quad \eta > 0. \end{aligned}$$

□

Since Λ_f contains infinitely many basis functions, a realization of such a Besov space random variable is an element of the space of definition with probability zero. To guarantee finiteness of the norm, the functions have to be defined in a Besov space which is smoother than the one where the realizations are measured. The following Lemma was adopted from [16, Lemma 2], but there the authors considered functions on a d-dimensional torus instead of $\mathbb{R}^d$.

Lemma 3.3. *Let $\mathbf{X}$ be defined in $B^r_p(\mathbb{R}^d)$ as in Definition 3.1 for some $r > 0$ and $2 < \alpha < \infty$. Then the following three conditions are equivalent:*

(i) $||\mathbf{X}||_{B^s_p(\mathbb{R}^d)} < \infty \quad \text{almost surely},$

(ii) $\mathbb{E}\exp\left(||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)}\right) < \infty,$

(iii) $s < r - \frac{d}{p}.$

Proof. Let $(X^\alpha_\lambda)_{\lambda\in\Lambda_f}$ be as in Definition 3.1. First consider the expectation of $||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)}$. Because of (22) we have

$$\begin{aligned}\mathbb{E}||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)} \quad &= \mathbb{E}\sum_{\lambda\in\Lambda_f} 2^{(s+d\left(\frac{1}{2}-\frac{1}{p}\right))p|\lambda|}\left|2^{-(r+d\left(\frac{1}{2}-\frac{1}{p}\right))|\lambda|}X^\alpha_\lambda\right|^p \\ &= \mathbb{E}\sum_{\lambda\in\Lambda_f} 2^{-(r-s)p|\lambda|}|X^\alpha_\lambda|^p = \sum_{\lambda\in\Lambda_f} 2^{-(r-s)p|\lambda|}\mathbb{E}|X^\alpha_\lambda|^p \\ &= \frac{2}{\alpha p}\sum_{\lambda\in\Lambda_f} 2^{-(r-s)p|\lambda|}. \end{aligned} \tag{23}$$

Because of the construction of Λ_f there are ℓ_ϕ scaling functions and ℓ_ψ wavelets on the coarsest scale. Additionally, on scale $j > 0$ we have at most $2^{jd}\ell_\psi$ wavelets. Hence, the summation in (23), which is actually a double some over all wavelets and scales, reduces to a simple sum and

$$\begin{aligned}\mathbb{E}||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)} &= \frac{2}{\alpha p}\left(\ell_\phi + \sum_{j=0}^\infty 2^{-(r-s)pj}\cdot 2^{jd}\ell_\psi\right) \\ &= \frac{2}{\alpha p}\left(\ell_\phi + \ell_\psi\sum_{j=0}^\infty 2^{-j((r-s)p-d)}\right)\end{aligned} \tag{24}$$

The sum converges if and only if $(r-s)p - d > 0$. Since finiteness of the expectation of a positive random variable implies almost sure finiteness of the random variable itself, $||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)} < \infty$ a.s. and also $||\mathbf{X}||_{B^s_p(\mathbb{R}^d)} < \infty$ a.s., hence $(i) \Leftrightarrow (iii)$. Now we turn to condition (ii). It is

$$\begin{aligned}\mathbb{E}\exp\left(||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)}\right) &= \mathbb{E}\exp\left(\sum_{\lambda\in\Lambda_f} 2^{-(r-s)p|\lambda|}|X^\alpha_\lambda|^p\right)\\ &= \prod_{\lambda\in\Lambda_f}\mathbb{E}\exp\left(2^{-(r-s)p|\lambda|}|X^\alpha_\lambda|^p\right)\\ &= \prod_{\lambda\in\Lambda_f}\left(1-\frac{2^{-(r-s)p|\lambda|+1}}{\alpha}\right)^{-1/p}\\ &= \left(1-\frac{2}{\alpha}\right)^{-\frac{\ell_\phi}{p}}\cdot\left(\prod_{j=0}^\infty\left(1-\frac{2^{-(r-s)p|\lambda|+1}}{\alpha}\right)^{2^{jd}\ell_\psi}\right)^{-\frac{1}{p}}\end{aligned} \tag{25}$$

where we used that the X^α_λ are independent and $\mathbb{E}\exp(c|X^\alpha_\lambda|^p) = (1-\frac{2c}{\alpha})^{-1/p}$ if $\alpha > 2c$ (which is why we have to require $\alpha > 2$). Since $\prod_{l=0}^\infty(1+a_l)$ converges if and only if $\sum_{l=0}^\infty \log(1+a_l)$ converges we find that $\mathbb{E}\exp\left(||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)}\right) < \infty$ if

$$\sum_{j=0}^\infty 2^{jd}\ell_\psi \log\left(1-\frac{2^{-(r-s)p|\lambda|+1}}{\alpha}\right) < \infty. \tag{26}$$

The root test yields

$$\lim_{j\to\infty}\left(2^{jd}\ell_\psi\log\left(1-\frac{2^{-(r-s)p|\lambda|+1}}{\alpha}\right)\right)^{\frac{1}{j}} = 2^{-(r-s)p+d}. \tag{27}$$

Hence the sum and by that (25) converges if (iii) holds. Since (ii) obviously implies (i) the proof is complete. □

The Lemma shows that, although we define the random variable in the Besov space $B^r_p(\mathbb{R}^d)$, its realizations will only be elements of the less smooth space $B^s_p(\mathbb{R}^d)$. If such a combination is used for spaces of definition and measurement of the random variables, finiteness of the norms in the latter space is ensured if condition *(iii)* is fulfilled. We will refer to this as the *infinite model* (MI). A second possibility is to consider a finite dimensional model: Let T_n, T_n^* be defined as in (5). Then for a function $\mathbf{x} \in L_2(\mathbb{R}^d)$ and arbitrary, but fixed $n \in \mathbb{N}$, $T_n^*T_n\mathbf{x} = \sum_{\lambda\in\Lambda_n} x_\lambda\psi_\lambda$ is an element of $B^s_p(\mathbb{R}^d)$ with probability one if the wavelet is smooth enough. This allows in particular to measure the realizations of random variables in the same norm as was used in the definition of the random function. This will be referred to as the *finite model* (MII). In order to derive convergence rates we need to calculate $\mathbb{P}(||X||_{B^s_p(\mathbb{R}^d)} \geq \varrho)$ for given $\varrho > 0$. The following Corollary shows how this can be done for model (MI) using Lemma 3.3.

Corollary 3.4. *Consider model (MI). Let* $\mathbf{X}$ *be defined in* $B^r_p(\mathbb{R}^d)$ *according to Definition 3.1 with* $2 < \alpha < \infty$. *Let* $s < r - \frac{d}{p}$ *and* $\varrho > 0$. *Then*

$$\mathbb{P}(||\mathbf{X}||_{B^s_p(\mathbb{R}^d)} > \varrho) \leq \frac{1}{\varrho}\left(\frac{2}{\alpha p}\left(\ell_\phi + \ell_\psi \sum_{j=0}^{\infty} 2^{-j((r-s)p-d)}\right)\right)^{\frac{1}{p}} \tag{28}$$

Proof. According to Chebyshev's inequality, for any nonnegative random variable ξ with $\mathbb{E}\xi < \infty$, $\mathbb{P}(\xi > \varrho) \leq \frac{1}{\varrho}\mathbb{E}\xi$. Since the mapping $z \mapsto z^p$ is bijective for $z \geq 0$ and $1 \leq p \leq 2$, we have for given $\varrho > 0$

$$\mathbb{P}(||\mathbf{X}||_{B^s_p(\mathbb{R}^d)} > \varrho) = \mathbb{P}(||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)} > \varrho^p) \leq \frac{1}{\varrho^p}\mathbb{E}||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)}.$$

The expectation of $||\mathbf{X}||^p_{B^s_p(\mathbb{R}^d)}$ is given by (24). □

Using the finite model, we get the following result.

Lemma 3.5. *Consider model (MII). Let* $\mathbf{X}$ *be defined as* $B^s_p(\mathbb{R}^d)$ *random function according to Definition 3.1,* T_n *as in (5) and take* $\varrho > 0$. *Denote* $X_n := T_n^* T_n \mathbf{X}$. *Then*

$$\mathbb{P}(||X_n||_{B^s_p(\mathbb{R}^d)} > \varrho) = \frac{\Gamma(\frac{n}{p}, \frac{\alpha\varrho^p}{2})}{\Gamma(\frac{n}{p})} \tag{29}$$

with the Gamma functions

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt, \qquad \Gamma(a,z) = \int_z^\infty t^{a-1}e^{-t}dt.$$

Proof. Let $\mathbf{X}$ be as in Definition 3.1. Then $X_n = \sum_{\lambda\in\Lambda_n} 2^{-\varsigma|\lambda|}X^\alpha_\lambda \psi_\lambda$ and $||X_n||^p_{B^s_p(\mathbb{R}^d)} = \sum_{\lambda\in\Lambda_n} |X^\alpha_\lambda|^p$ reduces to a sum of n i.i.d. random variables with density (21). The resulting density can be calculated using the moment generating function (c.f., e.g. [2]) of the X^α_λ which is just the Laplace transform $\mathcal{L}(\cdot)$ of the probability density function. The moment generating function of a sum of random variables is given by the product of the single moment generating functions [2]. With $\pi_{|X^\alpha_\lambda|^p}$ from (21) we get

$$\mathcal{L}[\pi_{|X^\alpha_\lambda|^p}](s) = \left(1 + \frac{2s}{\alpha}\right)^{-1/p}$$

and obtain the probability density function of $\pi_{\sum_{\lambda\in\Lambda_n}|X_\lambda^\alpha|^p}(\xi)$, $\xi \geq 0$, via the inverse Laplace transform $\mathcal{L}^{-1}$,

$$\pi_{\sum_{\lambda\in\Lambda_n}|X_\lambda^\alpha|^p}(\xi) = \mathcal{L}^{-1}\left[\left(1+\frac{2s}{\alpha}\right)^{-n/p}\right](\xi) = \frac{\xi^{\frac{n}{p}-1}}{\Gamma(\frac{n}{p})}\left(\frac{\alpha}{2}\right)^{\frac{n}{p}} e^{-\frac{\alpha}{2}\xi}. \tag{30}$$

Because $\sum_{\lambda\in\Lambda_n}|X_\lambda^\alpha|^p$ is non-negative, $\mathbb{P}(||X_n||_{B_p^s(\mathbb{R}^d)} > \varrho) = \mathbb{P}(||X_n||_{B_p^s(\mathbb{R}^d)}^p > \varrho^p)$. The claim follows by integrating (30) over ξ from ϱ^p to infinity. □

Remark 3.6. *Similar to the infinite dimensional setting, Chebyshev's inequality allows to estimate*

$$\mathbb{P}(||X_n||_{B_p^s(\mathbb{R}^d)} > \varrho) \leq \frac{1}{\varrho}\sqrt[p]{\frac{2n}{\alpha p}}. \tag{31}$$

This is indeed an upper bound for (29).

Remark 3.7. *The relation between the two models is best seen by comparing (28) and (31). Both probabilities solely differ in a term describing the wavelet structure. In model (MI) the term $\ell_\phi + \ell_\psi \sum_{j=0}^\infty 2^{-j((r-s)p-d)}$ ensures that $\mathbb{E}(||\mathbf{X}||)$ is bounded independently of n, whereas in the finite model the expectation $\mathbb{E}(||X_n||)$ grows unbounded. If ϱ is an a-priori estimate of $||\mathbf{X}||$ or $||X_n||$, respectively, it will have to be chosen differently for the two models, taking into account the different asymptotic behaviour of the respective random variables.*

4 The Ky Fan metric

In order to establish a convergence analysis for Inverse Problems in a stochastic setting, an appropriate metric for random variables is required. In this paper we consider the Ky Fan metric (cf. [10]) which is defined as follows.

Definition 4.1. *Let x_1 and x_2 be random variables in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a metric space (χ, d_χ). The distance between x_1 and x_2 in the Ky Fan metric is defined as*

$$\rho_K(x_1, x_2) := \inf_{\epsilon>0}\{\mathbb{P}(\{\omega\in\Omega : d_\chi(x_1(\omega), x_2(\omega)) > \epsilon\}) < \epsilon\}. \tag{32}$$

This metric essentially allows to lift results from a metric space to the space of random variables. In particular, if $\rho_K(x_1, x_2) \leq \epsilon$ for some $0 < \epsilon \leq 1$, $d_\chi(x_1, x_2) \leq \epsilon$ with probability $1 - \epsilon$. An immediate consequence of (32) is that $\rho_K(x_1, x_2) = 0$ if and only if $x_1 = x_2$ almost surely. Convergence in the Ky Fan metric is equivalent to convergence in probability, i.e. for a sequence $\{x_k\}_{k\in\mathbb{N}} \in \mathcal{X}$ and $x \in \mathcal{X}$,

$$\rho_K(x_k, x) \overset{k\to\infty}{\longrightarrow} 0 \quad \Leftrightarrow \quad \forall\epsilon > 0 \quad \mathbb{P}(||x_k - x||_{\mathcal{X}} > \epsilon) \overset{k\to\infty}{\longrightarrow} 0.$$

Hence convergence in the Ky Fan metric also leads to pointwise (almost sure) convergence of certain subsequences in the metric d_χ (cf. [7]), as formulated in the following Proposition.

Proposition 4.2 ([11], Proposition 1.10)**.** *Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence of random variables that converges to x in the Ky Fan metric. Then for any $\eta > 0$ and $\epsilon > 0$ there exist $\Omega_\epsilon \subset \Omega$, $\mathbb{P}(\Omega_\epsilon) \geq 1 - \epsilon$, and a subsequence x_{k_j} with*

$$||x_{k_j}(\omega) - x(\omega)|| \leq (1 + \eta)\rho_K(x_{k_j}, x) \qquad \forall\omega \in \Omega_\epsilon.$$

Furthermore there exists a subsequence that converges to x almost surely.

The following estimate for the Ky Fan distance of a Gaussian random variable to its mean is a special case of Proposition 2.5 in [20].

Proposition 4.3. *Let ξ be a random variable with values in $\mathbb{R}^m$. Assume that the distribution of ξ is $\mathcal{N}(0, \sigma^2 I)$ with $\sigma > 0$. Then it holds in $(\mathbb{R}^m, ||\cdot||)$ that*

$$\rho_K(\xi, 0) \leq \min\left\{1, \sqrt{2}\sigma\sqrt{m - \ln^-\left(\sigma^2 2\pi m^2 \left(\frac{e}{2}\right)^m\right)}\right\}, \tag{33}$$

where $f^-(h) := \min\{0, f(h)\}$.

The smallest m for which the $\ln^-$-term vanishes is at the zero of $\ln^-\left(\sigma^2 2\pi m^2 \left(\frac{e}{2}\right)^m\right)$. It is given by

$$m_{\min} = \mathbf{ceil}\left(\frac{2}{1 - \ln 2} W\left(\frac{1 - \ln 2}{2\sqrt{2\pi}}\frac{1}{\sigma}\right)\right),$$

where W is the Lambert W-function defined by $W(z)e^{W(z)} = z$ (cf. [4]), and $\mathbf{ceil}(\cdot)$ the function which maps a real number to the smallest following integer. In practice $m > m_{\min}$ typically is fulfilled. Then $\rho_K(\xi, 0) \leq \min\left\{1, \sqrt{2}\sigma\sqrt{m}\right\} = \mathcal{O}(\mathbb{E}(||\xi||))$, in other words the Ky-Fan distance is of the same order as the expectation of the error. Nevertheless we will give the main results for the complete error estimate.

5 Convergence of maximum a posteriori solutions with Besov priors

Before we analyse convergence properties of the Tikhonov regularization in the stochastic setting, i.e. of the maximum a posteriori solution (13), we want to review facts for the deterministic case proved in [6].

Remark 5.1. *In this paper we assume that for $p = 1$ the operator A is injective in order to guarantee existence of a unique minimizer of the Tikhonov functional (13). This assumption is not needed if $p > 1$ since the functional is strictly convex in that case.*

Theorem 5.2 ([6], Theorem 4.1)**.** *Assume that A is a bounded operator from $\mathcal{X}$ to $\mathcal{Y}$ with $||A|| < 1$, that $1 \le p \le 2$, and that $c < \min_\lambda w_\lambda$, $\{w_\lambda\}_{\lambda\in\Lambda} = \mathbf{w}$ for some constant $c > 0$. Assume that either $p > 1$ or $N(A) = \{0\}$. Let $x^*_{\hat{\alpha}}$ be the minimizer of (11) for given data y^δ with $||y - y^\delta|| \le \delta$ and $\hat{\alpha} > 0$. If $\hat{\alpha} = \hat{\alpha}(\delta)$ satisfies the requirements*

$$\lim_{\delta\to 0} \hat{\alpha}(\delta) \to 0 \quad \text{and} \quad \lim_{\delta\to 0} \frac{\delta^2}{\hat{\alpha}(\delta)} = 0,$$

then we have, for any $x_0 \in \mathcal{X}$,

$$\lim_{\delta\to 0} \left[\sup_{||y-y^\delta||\le\delta} ||x^*_{\hat{\alpha}} - x^\dagger|| \right] = 0,$$

i.e. the regularized solutions converge to $x^\dagger$, where $x^\dagger$ is the solution of the equation $Ax = y$ with minimal value of $\Phi(\cdot)$.

Remark 5.3. *The requirement $||A|| < 1$ was imposed to ensure convergence of the iterative algorithm for the computation of the minimizer. It is not necessary for the minimizer itself since the equation can always be scaled appropriately.*

Remark 5.4. *In both models (MI) and (MII) for the Besov random function the weights are given by $w_\lambda = 2^{\varsigma p|\lambda|} > 0$, $\varsigma = s + d(\frac{1}{2} - \frac{1}{p}) \ge 0$.*

In order to carry Theorem 5.2 over to the stochastic case we need the following Lemma by Egoroff.

Lemma 5.5. *([8], see also [7]) Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. Let x_k and x be measurable functions from Ω into a metric space χ with metric d_χ. Suppose $x_k(\omega) \overset{d_\chi}{\to} x(\omega)$ for μ-almost all $\omega \in \Omega$. Then for any $\epsilon > 0$ there is a set Ω_ϵ with $\mu(\Omega \backslash \Omega_\epsilon) < \epsilon$ such that $x_k \overset{d_\chi}{\to} x(\omega)$ uniformly on Ω_ϵ, that is*

$$\lim_{k\to\infty} \sup\{d_\chi(x_k(\omega), x(\omega)) : \omega \in \Omega_\epsilon\} = 0.$$

In Theorem 4.1 of his PhD thesis [11], Hofinger proved how by means of the Ky Fan metric deterministic results can be lifted to the space of random variables. The same techniques can be used in our situation as well.

Theorem 5.6. *Let $y = y(\omega)$ be the exact right hand side in (7) and $\{y^{\hat\sigma_k}(\omega)\}_{k\in\mathbb{N}}$ be a sequence of noisy realizations of $y(\omega) + \epsilon(\omega)$ such that $\rho_K(y, y^{\hat\sigma_k}) \le \hat\sigma_k$, $\hat\sigma_k \to 0$ as $k \to \infty$. Let $\hat\alpha(\hat\sigma_k)$ be a parameter choice rule such that $\hat\alpha(\hat\sigma_k) \to 0$ and $\hat\sigma^2/\hat\alpha(\hat\sigma_k) \to 0$ as $\hat\sigma_k \to 0$. Furthermore let the minimum norm solution $x^\dagger$ be unique, i.e. $1 < p \le 2$ or $N(A) = \{0\}$. Denote with $x^*_{\hat\alpha(\hat\sigma)}$ the minimizer of (11). Then*

$$\lim_{\hat\sigma\to 0} \rho_K(x^\dagger, x^*_{\hat\alpha(\hat\sigma)}) = 0.$$

Proof. The proof is an adaption of a proof given by Hofinger in [11, Theorem 4.1]. Define $\eta := \limsup_{k\to\infty} \rho_K(x^\dagger, x^*_{\hat\alpha(\hat\sigma_k)})$. (Note that $0 \le \eta \le 1$ due to the properties of the Ky Fan metric.) We show in the following that for arbitrary $\epsilon > 0$ we have $\eta/2 \le \epsilon$ and consequently $\limsup_{k\to\infty} \rho_K(x^\dagger, x^*_{\hat\alpha(\hat\sigma_k)}) = \lim_{k\to\infty} \rho_K(x^\dagger, x^*_{\hat\alpha(\hat\sigma_k)}) = 0$.
As a first step we pick a "worst case" subsequence $\{y^{\hat\sigma_{k^j}}\}$ of $\{y^{\hat\sigma_k}\}$, a subsequence for which the corresponding solutions satisfy $\rho_K(x^\dagger, x^*_{\hat\alpha(\hat\sigma_{k^j})}) \ge \eta/2$. We now show that even from this "worst case" sequence we can pick a subsequence $\{y^{\hat\sigma_{k^j_l}}\}$ for which we have $\limsup \rho_K(x^\dagger, x^*_{\hat\alpha(\hat\sigma_{k^j_l})}) \le \epsilon$ for arbitrary $\epsilon > 0$.
Let $\epsilon > 0$. According to Proposition 4.2 we can pick a subsequence $\{y^{\hat\sigma_{k^j_l}}\}$ and a set Ω_ϵ with $\mathbb{P}(\Omega_\epsilon) \ge 1 - \frac{\epsilon}{2}$ as well as $||y(\omega) - y^{\hat\sigma_{k^j_l}}(\omega)|| \le 2\hat\sigma_{k^j_l}$ on Ω_ϵ. For all $\omega \in \Omega_\epsilon$, the noise tends to zero, we can therefore use the deterministic result and deduce via Theorem 5.2 that $x^*_{\hat\alpha(\hat\sigma_{k^j_l})}(\omega)$ converges to the unique solution $x^\dagger(\omega)$ for $\hat\sigma_{k^j_l} \to 0$, $\omega \in \Omega_\epsilon$ if $\hat\alpha(\hat\sigma) \to 0$ and $\hat\sigma^2/\hat\alpha(\hat\sigma) \to 0$ as $\hat\sigma \to 0$. This convergence is not uniform in ω; nevertheless, pointwise convergence implies uniform convergence except on sets of small measure according to

Lemma 5.5. Therefore there exist $\Omega'_\epsilon \subset \Omega_\epsilon$, $\mathbb{P}(\Omega'_\epsilon) < \frac{\epsilon}{2}$ and $j_0 \in \mathbb{N}$ such that $||x^*_{\hat{\alpha}(\hat{\sigma}_{k_l^j})}(\omega) - x^\dagger(\omega)|| < \epsilon \; \forall \omega \in \Omega_\epsilon \backslash \Omega'_\epsilon$ and $j \geq j_0$. We thus have

$$\mathbb{P}\left(\left\{\omega \in \Omega_\epsilon : ||x^*_{\hat{\alpha}(\hat{\sigma}_{k_l^j})}(\omega) - x^\dagger(\omega)|| > \epsilon\right\}\right) \leq \mathbb{P}(\Omega'_\epsilon) \leq \epsilon/2.$$

Since we split $\Omega = \Omega\backslash\Omega_\epsilon \cup \Omega_\epsilon\backslash\Omega'_\epsilon \cup \Omega'_\epsilon$ with $\mathbb{P}(\Omega\backslash\Omega_\epsilon) < \frac{\epsilon}{2}$, $\mathbb{P}(\Omega\backslash\Omega_\epsilon) + \mathbb{P}(\Omega'_\epsilon) \leq \epsilon$ we have shown existence of a subsequence $\hat{\sigma}_{k_l^j}$ such that

$$\mathbb{P}\left(\left\{\omega \in \Omega : ||x^*_{\hat{\alpha}(\hat{\sigma}_{k_l^j})}(\omega) - x^\dagger(\omega)|| > \epsilon\right\}\right) \leq \epsilon$$

for $\hat{\sigma}_{k_l^j}$ sufficiently small. This ϵ is by definition of the Ky Fan metric an upper bound for the distance between $x^*_{\hat{\alpha}(\hat{\sigma}_{k_l^j})}$ and $x^\dagger$. Therefore we have $\limsup_{l\to\infty} \rho_K(x^*_{\hat{\alpha}(\hat{\sigma}_{k_l^j})}, x^\dagger) \leq \epsilon$. On the other hand, the original sequence satisfied $\liminf_{j\to\infty} \rho_K(x^\dagger, x^*_{\hat{\alpha}(\hat{\sigma}_{k^j})}) \geq \eta/2$. Since $\liminf_{j\to\infty} \rho_K(x^\dagger, x^*_{\hat{\alpha}(\hat{\sigma}_k)}) \leq \limsup_{l\to\infty} \rho_K(x^*_{\hat{\alpha}(\hat{\sigma}_{k_l^j})}, x^\dagger)$ it follows $\eta/2 \leq \epsilon$. Because $\epsilon > 0$ was arbitrary, this implies $\eta = 0$, which concludes the proof. □

Corollary 5.7. *Let $\alpha, \sigma > 0$, $1 \leq p \leq 2$ and $N(A) = 0$ for $p = 1$. Let $x^{\text{MAP}}_{\hat{\alpha}} = x^{\text{MAP}}_{\alpha,\sigma}$ be the solution of (13). If $\alpha = \alpha(\sigma)$ is chosen such that $\alpha\sigma^2 \to 0$ and $\frac{|\log \sigma|}{\alpha} \to 0$ as $\sigma \to 0$, then*

$$\lim_{\sigma \to 0} \rho_K(x^{\text{MAP}}_{\hat{\alpha}}, x^\dagger) = 0.$$

Proof. From (33) we have

$$\rho_K(y, y^\sigma) \leq \sqrt{2}\sigma\sqrt{m - \ln^-\left(\sigma^2 2\pi m^2 \left(\frac{e}{2}\right)^m\right)} =: \hat{\sigma}.$$

Recall the definition of $\hat{\alpha} = \alpha\sigma^2$ and that the maximum a posteriori solution (13) coincides with the minimizer of the Tikhonov functional (11). Theorem 5.6 ensures convergence of $x^{\text{MAP}}_{\hat{\alpha}}$ to $x^\dagger$ with respect to the Ky Fan metric if $\hat{\alpha} = \alpha\sigma^2 \to 0$ and $\hat{\sigma}^2/\hat{\alpha}(\hat{\sigma}) \to 0$. From the definitions of $\hat{\sigma}$ and $\hat{\alpha}$ the condition

$$\frac{\left(\sqrt{2}\sigma\sqrt{m - \ln^-\left(\sigma^2 2\pi m^2 \left(\frac{e}{2}\right)^m\right)}\right)^2}{\alpha\sigma^2} \xrightarrow{\sigma\to 0} 0$$

follows. This is fulfilled if α grows faster than $|\ln \sigma|$ as $\sigma \to 0$. □

Remark 5.8. *As long as the logarithm is inactive, it suffices to require* $\alpha\sigma^2 \to 0$ *and* $\alpha \to \infty$*. The parameter* α *may grow with arbitrarily slow speed.*

From a stochastic point of view, α can be interpreted as a measure for the variance of the prior. If $\alpha \to \infty$, the variance goes to zero. In other words, with high probability the coefficients $|X_\lambda^\alpha|^p$ are very close to zero which emphasizes the sparsity background. However, the actual regularization parameter is $\hat{\alpha} = \sigma^2\alpha$ and goes to zero as $\sigma \to 0$.

Remark 5.9. *Corollary 5.7 also shows the necessity of introducing the extra tuning parameter* α *in the prior distribution (12). If* $\alpha = 1$ *independent of* σ *we can not expect convergence of the algorithm with respect to the Ky Fan metric.*

6 Convergence rates

6.1 Deterministic results

Although we proved that the maximum a posteriori solutions converge to the true solution, it is well-known from deterministic theory (cf. [9]) that in general the convergence can be arbitrarily slow. In order to guarantee a certain decrease of the reconstruction error with respect to the noise parameter, i.e. to derive convergence rates, it is necessary to impose additional conditions on either the true solution, the operator, or both. We will require a smoothing property of the operator A and an a priori bound of the norm of the solution. Since we are only interested in convergence with respect to the noise, we will consider the discretization levels m and n fixed.

Let us first summarize some well known deterministic results. For details, we refer to [6, Section 4.2]. Assume $N(A) = \{0\}$ for $p = 1$ and suppose that we know a priori a bound on the sparsity penalty of the exact solution, i.e. $||\mathbf{x}^*||_{B_p^s(\mathbb{R}^d)} \le \varrho$ for some $\varrho > 0$. If we also know that $\mathbf{y}$ lies within a distance ϵ of $\mathbf{Ax}^*$ in $\mathcal{Y}$, then the exact solution can be localized within the set

$$\mathcal{F}(\epsilon, \varrho) := \{\mathbf{x} \in \mathcal{X} : ||\mathbf{Ax} - \mathbf{y}|| \le \epsilon, ||\mathbf{x}||_{B_p^s(\mathbb{R}^d)} \le \varrho\}.$$

The diameter of this set is a measure of the uncertainty of the solution for a given a priori constant ϱ and noise level ϵ. The maximum diameter of $\mathcal{F}$ is bounded by $2M(\epsilon, \varrho)$ where $M(\epsilon, \varrho)$, defined by

$$M(\epsilon, \varrho) := \sup\{||\mathbf{h}|| : ||\mathbf{A}\mathbf{h}|| \leq \epsilon, ||\mathbf{h}||_{B_p^s(\mathbb{R}^d)} \leq \varrho\}, \tag{34}$$

is called the *modulus of continuity* of A^{-1} under the a priori constraint. It can also be interpreted as the worst case error. An upper bound on the reconstruction error is given by the *modulus of convergence*

$$M_{\hat{\alpha}}(\epsilon, \varrho) := \sup\{||\mathbf{x}_{\hat{\alpha}}^* - \mathbf{x}|| : \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, ||\mathbf{A}\mathbf{x} - \mathbf{y}|| \leq \epsilon, ||\mathbf{x}||_{B_p^s(\mathbb{R}^d)} \leq \varrho\} \tag{35}$$

where $\mathbf{x}_{\hat{\alpha}}^*$ denotes the minimizer of the Tikhonov functional (11). The decay of this modulus of convergence as $\epsilon \to 0$ is governed by the decay of the modulus of continuity, as shown in the following proposition:

Proposition 6.1 ([6], Prop. 4.5)**.** *The modulus of convergence (35) satisfies*

$$M(\epsilon, \varrho) \leq M_{\hat{\alpha}}(\epsilon, \varrho) \leq M(\epsilon + \epsilon', \varrho + \varrho'), \tag{36}$$

where

$$\epsilon' = (\epsilon^2 + \hat{\alpha}\varrho^p)^{\frac{1}{2}}, \qquad \varrho' = (\varrho^p + \epsilon^2\hat{\alpha}^{-1})^{\frac{1}{p}} \tag{37}$$

and $M(\epsilon, \varrho)$ is defined in (34).

Thus it suffices to investigate the convergence behaviour of the modulus of continuity. As in [6], let us additionally assume that the operator A is of smoothing order β, that is, we assume that for some $\beta > 0$ there exist constants A_l and A_u such that for all $\mathbf{h} \in L^2(\mathbb{R}^d)$

$$A_l^2 \sum_{\lambda \in \Lambda} 2^{-2|\lambda|\beta} |\langle \mathbf{h}, \psi_\lambda \rangle|^2 \leq ||\mathbf{A}\mathbf{h}||^2 \leq A_u^2 \sum_{\lambda \in \Lambda} 2^{-2|\lambda|\beta} |\langle \mathbf{h}, \psi_\lambda \rangle|^2. \tag{38}$$

The decay of the modulus of continuity is then characterized as follows.

Proposition 6.2 ([6], Proposition 4.7)**.** *If the operator A satisfies the smoothing property (38), then the modulus of continuity $M(\epsilon, \varrho)$ satisfies*

$$c\left(\frac{\epsilon}{A_u}\right)^{\frac{\varsigma}{\varsigma+\beta}} \varrho^{\frac{\beta}{\beta+\varsigma}} \leq M(\epsilon, \varrho) \leq C\left(\frac{\epsilon}{A_l}\right)^{\frac{\varsigma}{\varsigma+\beta}} \varrho^{\frac{\beta}{\beta+\varsigma}},$$

where $\varsigma = s + d(\frac{1}{2} - \frac{1}{p}) \geq 0$ and c and C are constants depending on ς and β only.

6.2 Lifting the deterministics result into the stochastic setting

In this Section, the Ky Fan metric will be used to lift the deterministic results into the stochastic setting.

Lemma 6.3. *Let $e = (\epsilon_1, \epsilon_2, \dots, \epsilon_m)^T \in \mathbb{R}^m$ where ϵ_i, $i = 1, \dots, m$, are independent identically distributed Gaussian random variables with zero mean and variance σ^2. Then for any $c > 0$*

$$\mathbb{P}(||e|| > c) = \frac{\Gamma(\frac{m}{2}, \frac{c^2}{2\sigma^2})}{\Gamma(\frac{m}{2})}. \tag{39}$$

Proof. We have

$$\mathbb{P}(||e|| > c) = \mathbb{P}\left(\sqrt{\sum_{i=1}^{m} \epsilon_i^2} > c\right) = \mathbb{P}\left(\sum_{i=1}^{m} \epsilon_i^2 > c^2\right). \tag{40}$$

Define $Z := \sum_{i=1}^{m} \epsilon_i^2$. Then Z is the sum of the squares of m Gaussian random variables with zero mean and variance σ^2. Z is χ^2-distributed (see for example [2]) and obeys the probability density function

$$f_Z(\tau) = \frac{1}{2^{\frac{m}{2}} \sigma^m \Gamma(\frac{m}{2})} \tau^{\frac{m}{2}-1} e^{-\frac{\tau}{2\sigma^2}}. \tag{41}$$

Hence

$$\mathbb{P}\left(\sum_{i=1}^{m} \epsilon_i^2 > c^2\right) = \int_{c^2}^{\infty} f_Z(\tau) d\tau = \frac{\Gamma(\frac{m}{2}, \frac{c^2}{2\sigma^2})}{\Gamma(\frac{m}{2})}. \tag{42}$$

□

Now we are ready for the main theorems in which we will prove convergence rates using first the finite model, then the infinite model. To simplify the notation we denote $L_m(\sigma) := \ln^-\left(\sigma^2 2\pi m^2 \left(\frac{e}{2}\right)^m\right)$ from Proposition 4.3 and $\mathcal{E}(\sigma, m, \alpha) := \sqrt{2}\sigma\sqrt{m - L_m(\sigma)} + \sqrt{2}\sigma\sqrt{m - L_m(\sigma) + \frac{\alpha\varrho^p}{2}}$.

Theorem 6.4. *Let $X_n = T_n^* T_n \mathbf{X}$ be defined as $B_p^s(\mathbb{R}^d)$-random variable according to Definition 3.1 for $1 \leq p \leq 2$ and take $s \in \mathbb{R}$ such that $\varsigma = s + d(\frac{1}{2} - \frac{1}{p}) \geq 0$. Let $x_{\hat{\alpha}}^{\mathrm{MAP}}$ be the maximum a-posteriori estimate (13) for* *the solution of (7) in the Bayesian framework with a B^s_p-Besov space prior (12) according to model (MII). Assume we are given noisy data $y^\sigma \in \mathbb{R}^m$ such that the error in each component of y^σ is normally distributed with zero mean and variance σ^2. Assume additionally that the operator A fulfils (38) with $\beta > 0$, $A_l > 0$, and in case $p = 1$, $N(A) = \{0\}$. If the solution is smooth enough, i.e. there is an a-priori estimate $||\mathbf{x}^\dagger||_{B^s_p(\mathbb{R}^d)} \leq \varrho$ for some $\varrho > 0$, then as $\sigma \to 0$ the maximum a-posteriori solution $x^{\mathrm{MAP}}_{\hat{\alpha}}$ converges to the solution with minimal norm $||\cdot||_{B^s_p(\mathbb{R}^d)}$ provided that the parameter $\alpha = \alpha(\sigma, \varrho, \beta, \varsigma, p)$ is chosen such that*

$$\min\left(\left(\tfrac{\sqrt{2}}{A_l}\mathcal{E}(\sigma, m, \alpha)\right)^{\frac{\varsigma}{\beta+\varsigma}}\left(\varrho + \left(\varrho^p + \tfrac{2m - L_m(\sigma)}{\alpha}\right)^{\frac{1}{p}}\right)^{\frac{\beta}{\beta+\varsigma}}, 1\right)$$
$$= \frac{\Gamma(\frac{m}{2}, m - \ln^-(\sigma'(m)))}{\Gamma(\frac{m}{2})} + \frac{\Gamma(\frac{n}{p}, \frac{\alpha\varrho^p}{2})}{\Gamma(\frac{n}{p})} \qquad (43)$$

is fulfilled. Additionally, we have the error estimate

$$\rho_K \quad (x^{\mathrm{MAP}}_{\hat{\alpha}}, x^\dagger) = \mathcal{O}\left(\left(\sigma\sqrt{1 + |\ln(\sigma)| + \frac{\alpha\varrho^p}{2}}\right)^{\frac{\varsigma}{\beta+\varsigma}} \varrho^{\frac{\beta}{\beta+\varsigma}}\right).$$

Proof. To improve readability we define $\eta := \frac{\varsigma}{\beta+\varsigma}$, $\eta' := \frac{\beta}{\beta+\varsigma}$ and $\epsilon := \sqrt{2}\sigma\sqrt{m - L_m(\sigma)}$. Then according to Proposition 4.3, $\rho_K(y, y^\sigma) \leq \epsilon$ holds. From (36) and Proposition 6.2 we know

$$\begin{aligned}\sup\{||x^{\mathrm{MAP}}_{\hat{\alpha}} - x|| : x \in \mathcal{X}, y \in \mathcal{Y}, \qquad & ||Ax - y|| \leq \epsilon, ||T^*x||_{B^s_p(\mathbb{R}^d)} \leq \varrho\} \\ &= M_{\hat{\alpha}}(\epsilon, \varrho) < CA_l^{-\eta}\left(\epsilon + \epsilon'\right)^\eta \left(\varrho + \varrho'\right)^{\eta'}.\end{aligned}$$

$M_{\hat{\alpha}}$ is a deterministic quantity. In particular $||x^{\mathrm{MAP}}_{\hat{\alpha}} - x^\dagger|| \leq CA_l^{-\eta}(\epsilon+\epsilon')^\eta(\varrho+\varrho')^{\eta'}$ whenever $||Ax^\dagger - y^\sigma|| \leq \epsilon$ and $||T^*x^\dagger||_{B^s_p(\mathbb{R}^d)} \leq \varrho$. On the other hand, $||x^{\mathrm{MAP}}_{\hat{\alpha}} - x^\dagger||$ may be larger than $CA_l^{-\eta}(\epsilon+\epsilon')^\eta(\varrho+\varrho')^{\eta'}$ if at least one of the conditions above is violated. Hence

$$\begin{aligned}\mathbb{P}(\qquad & \{\omega \in \Omega : ||x^{\mathrm{MAP}}_{\hat{\alpha}}(\omega) - x^\dagger(\omega)|| > CA_l^{-\eta}(\epsilon+\epsilon')^\eta(\varrho+\varrho')^{\eta'}\}) \\ & \leq \mathbb{P}(\{\omega : ||Ax^\dagger(\omega) - y^\sigma(\omega)|| > \epsilon \vee ||T^*x^\dagger(\omega)||_{B^s_p(\mathbb{R}^d)} \geq \varrho\}) \\ \leq \mathbb{P}(&\{\omega : ||Ax^\dagger(\omega) - y^\sigma(\omega)|| > \epsilon\}) + \mathbb{P}(\{\omega : ||T^*x^\dagger(\omega)||_{B^s_p(\mathbb{R}^d)} \geq \varrho\}) (44)\end{aligned}$$

because for $A, B \subset \Omega : \mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. Note that $\mathbb{P}(\{\omega : ||Ax^\dagger(\omega) - y^\sigma(\omega)|| > \epsilon\})$ corresponds to $\mathbb{P}(\Omega \backslash \Omega_\epsilon)$ with Ω_ϵ from Theorem 5.6, i.e. the subset of Ω for which we do not have a worst-case error bound $||y(\omega) - y^{\hat{\sigma}}(\omega)|| \leq \hat{\sigma}$. For the probability $\mathbb{P}(\{\omega : ||T_n^* x^\dagger(\omega)||_{B_p^s(\mathbb{R}^d)} \geq \varrho\})$, $\varrho > 0$ we derived in Lemma 3.5

$$\mathbb{P}(||T_n^* x^\dagger(\omega)||_{B_p^s(\mathbb{R}^d)} \geq \varrho) = \frac{\Gamma(\frac{n}{p}, \frac{\alpha\varrho^p}{2})}{\Gamma(\frac{n}{p})}. \tag{45}$$

The probability $\mathbb{P}(||Ax^\dagger - y|| \geq \epsilon)$ is given in Lemma 6.3 with $c = \epsilon$ and $e = Ax^\dagger - y$,

$$\mathbb{P}(||Ax^\dagger - y|| \geq \epsilon) \ \ = \frac{\Gamma(\frac{m}{2}, \frac{\epsilon^2}{2\sigma^2})}{\Gamma(\frac{m}{2})} = \frac{\Gamma(\frac{m}{2}, m - L_m(\sigma))}{\Gamma(\frac{m}{2})}. \tag{46}$$

Inserting (45) and (46) into (44) we arrive at

$$\begin{aligned}\mathbb{P}(\{\omega \in \Omega : \ \ &||x_{\hat{\alpha}}^{\text{MAP}}(\omega) - x^\dagger(\omega)|| > CA_l^{-\eta}(\epsilon + \epsilon')^\eta(\varrho + \varrho')^{\eta'}\}) \\ &\leq \frac{\Gamma(\frac{m}{2}, \frac{\epsilon^2}{2\sigma^2})}{\Gamma(\frac{m}{2})} + \frac{\Gamma(\frac{n}{p}, \frac{\alpha\varrho^p}{2})}{\Gamma(\frac{n}{p})}.\end{aligned} \tag{47}$$

Comparing this with the definition of the Ky Fan metric (32), we get an upper bound for $\rho_K(x_{\hat{\alpha}}^{\text{MAP}}, x^\dagger)$ if we choose α such that

$$CA_l^{-\eta}(\epsilon + \epsilon')^\eta(\varrho + \varrho')^{\eta'} = \frac{\Gamma(\frac{m}{2}, \frac{\epsilon^2}{2\sigma^2})}{\Gamma(\frac{m}{2})} + \frac{\Gamma(\frac{n}{p}, \frac{\alpha\varrho^p}{2})}{\Gamma(\frac{n}{p})}. \tag{48}$$

Before we can solve (48) we have to calculate $\epsilon + \epsilon'$ and $\varrho + \varrho'$. Resubstituting the error ϵ and $\hat{\alpha} = \sigma^2\alpha$ into (37) we get

$$\epsilon' = \sqrt{2}\sigma\sqrt{m - L_m(\sigma) + \frac{\alpha\varrho^p}{2}}$$

and

$$\epsilon + \epsilon' \ \ = \sqrt{2}\sigma\sqrt{m - L_m(\sigma)} + \sqrt{2}\sigma\sqrt{m - L_m(\sigma) + \tfrac{\alpha\varrho^p}{2}} =: \mathcal{E}(\sigma, m, \alpha).$$

Analogously we find

$$\varrho + \varrho' = \varrho + \left(\varrho^p + \frac{2m - L_m(\sigma)}{\alpha}\right)^{1/p}$$

and (48) reads

$$C \quad \left(\tfrac{\sqrt{2}}{A_l}\mathcal{E}(\sigma,m,\alpha)\right)^{\eta}\left(\varrho+\left(\varrho^p+\tfrac{2m-L_m(\sigma)}{\alpha}\right)^{1/p}\right)^{\eta'}$$
$$=\tfrac{\Gamma(\frac{m}{2},m-L_m(\sigma))}{\Gamma(\frac{m}{2})}+\tfrac{\Gamma(\frac{n}{p},\frac{\alpha\varrho^p}{2})}{\Gamma(\frac{n}{p})}. \tag{49}$$

The only unknown quantity in (49) is the constant C. Since we have no information about it, we neglect it and set it to one. Solving (49) for α immediately gives an upper bound for $\rho_K(x_{\hat{\alpha}}^{\mathrm{MAP}}, x^{\dagger})$ by definition of the Ky Fan metric. Although (49) does not have an analytical solution, the nonlinear equation can still be solved numerically. By construction the convergence rate is given by both the left hand side and the right hand side of (49). □

We obtain a similar result for the infinite dimensional model. To this end, we only have to replace the probability $\mathbb{P}(||T_n^* x^{\dagger}||_{\cdot,p} \geq \varrho)$ from Lemma 3.5 by the one from Corollary 3.4. We obtain the following Corollary. As before, $\mathcal{E}(\sigma,m,\alpha) := \sqrt{2}\sigma\sqrt{m-L_m(\sigma)}+\sqrt{2}\sigma\sqrt{m-L_m(\sigma)+\frac{\alpha\varrho^p}{2}}$.

Corollary 6.5. *Let $x_{\hat{\alpha}}^{\mathrm{MAP}}$ be the maximum a-posteriori estimate (13) for the solution of (7) in the Bayesian framework with a B_p^s-Besov space prior (12) with $s \in \mathbb{R}$ fulfilling $\varsigma = s + d(\frac{1}{2}-\frac{1}{p}) \geq 0$. Let $\mathbf{X}$ be defined in $B_p^r(\mathbb{R}^d)$ according to model (MI) whit $s < r - \frac{d}{p}$. Assume we are given noisy data $y^\sigma \in \mathbb{R}^m$ such that the error in each component of y^σ is normally distributed with zero mean and variance σ^2. Assume additionally that the operator A fulfils (38) with $\beta > 0$, $A_l > 0$, and in case $p = 1$ $N(A) = \{0\}$. If the solution is smooth enough, i.e. there is an a-priori estimate $||\mathbf{x}^{\dagger}||_{B_p^s(\mathbb{R}^d)} \leq \varrho$ for some $\varrho > 0$, then as $\sigma \to 0$ the maximum a-posteriori solution $x_{\hat{\alpha}}^{\mathrm{MAP}}$ converges to the solution with minimal norm $||\cdot||_{B_p^s(\mathbb{R}^d)}$ provided that the parameter $\alpha = \alpha(\sigma,\varrho,\beta,\varsigma,p)$ is chosen such that*

$$\min\left(\left(\tfrac{\sqrt{2}}{A_u}\mathcal{E}(\sigma,m,\alpha)\right)^{\frac{\varsigma}{\beta+\varsigma}}\left(\varrho+\left(\varrho^p+\tfrac{2m-L_m(\sigma)}{\alpha}\right)^{\frac{1}{p}}\right)^{\frac{\beta}{\beta+\varsigma}},1\right)$$
$$=\tfrac{\Gamma(\frac{m}{2},m-L_m(\sigma))}{\Gamma(\frac{m}{2})}+\tfrac{1}{\varrho}\left(\tfrac{2}{\alpha p}\left(\ell_\phi+\ell_\psi\textstyle\sum_{j=0}^{\infty}2^{-j((s-r)p-d)}\right)\right)^{\frac{1}{p}} \tag{50}$$

is fulfilled. Additionally, we have the error estimate

$$\rho_K \quad (x_{\hat{\alpha}}^{\mathrm{MAP}}, x^{\dagger}) = \mathcal{O}\left(\left(\sigma\sqrt{1+|\ln(\sigma)|+\tfrac{\alpha\varrho^p}{2}}\right)^{\frac{\varsigma}{\beta+\varsigma}}\varrho^{\frac{\beta}{\beta+\varsigma}}\right). \tag{51}$$

7 Numerical examples for $p = 1$

In this section we want to illustrate our theoretical results with a specific example for the case $p = 1$, as it is well known that this choice produces sparse solutions. We will consider a deconvolution problem with a given kernel function. Convolution operators appear in many fields, e.g., in signal processing, where the output of a linear time-invariant system is given by the convolution of the input signal with the impulse response, a fixed function depending on the system. In image processing, blurring effects can often be modelled as convolution of an image with a smoothing kernel. Mathematically, we have an operator equation $\mathbf{A}\mathbf{x} = \mathbf{y}$ where $\mathbf{A} : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ is defined by

$$[\mathbf{A}\mathbf{x}](s) = [\mathbf{k} * \mathbf{x}](s) = \int_{\mathbb{R}^d} \mathbf{k}(s-t)\mathbf{x}(t)dt, \quad s \in \mathbb{R}^d \tag{52}$$

for some kernel function $\mathbf{k} \in L_2(\mathbb{R}^d)$. In order to use our theory, we have to require that $\mathbf{A}$ fulfils (38) for some $\beta > 0$. Since the properties of $\mathbf{A}$ are determined by its kernel $\mathbf{k}$, we just have to choose $\mathbf{k}$ appropriately. Additionally, we have to require $||\mathbf{A}|| < 1$ for the computation of the solutions (see below). Inequality (38) describes the equivalence of $||\mathbf{A}\mathbf{h}||_{L_2}$ with a norm of $\mathbf{h}$ in a Sobolev space of negative order $H^{-\beta}$. Using Fourier analysis we have

$$||\mathbf{h}||_{H^{-\beta}} = \int_{\mathbb{R}^d} (1 + |\xi|^2)^{-\beta} |\widehat{\mathbf{h}}(\xi)|^2 d\xi, \tag{53}$$

where $\widehat{\mathbf{h}}$ denotes the Fourier transform of $\mathbf{h}$. Because of the Fourier-convolution theorem

$$||\mathbf{A}\mathbf{h}||_{L_2} = ||\widehat{\mathbf{k}} \cdot \widehat{\mathbf{h}}||_{L_2} = \int_{\mathbb{R}^d} |\widehat{\mathbf{k}}(\xi) \cdot \widehat{\mathbf{h}}(\xi)|^2 d\xi. \tag{54}$$

Comparing (53) and (54) we may define $\hat{\mathbf{k}}(\xi) := (1 + |\xi|^2)^{-\beta/2}$ and obtain equality in (38) with $A_u = A_l = 1$. To control the width of the convolution filter we introduce an additional constant $\kappa > 0$ and define

$$\widehat{\mathbf{k}}(\xi) = \frac{c_{\kappa,\beta}}{(1 + \kappa|\xi|^2)^{\beta/2}}, \quad \xi \in \mathbb{R}^d \text{ with } c_{\kappa,\beta} \text{ suchthat } ||\widehat{\mathbf{k}}||_{L_2(\mathbb{R}^d)} < 1. \tag{55}$$

Now (38) holds with $A_l = c_{\kappa,\beta}^2$ and $A_u = \frac{c_{\kappa,\beta}^2}{\kappa}$ for $\kappa \leq 1$ and vice versa for $\kappa > 1$. The maximum a-posteriori solution, i.e. the minimizer of the

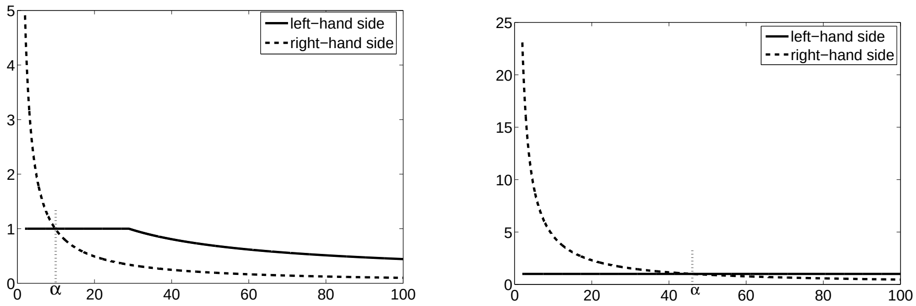

Figure 1: Plot of the left-hand side and right-hand side of (43) (MI, left) and (50) (MII, right) for $\sigma = 0.01$, $m = 2500$, $\varsigma = 0.5$, $\beta = 1$, $\varrho = 2.16$. The optimal α is the one for which the intersection occurs.

Tikhonov functional (13) was calculated with the *iterative soft-shrinking algorithm* proposed in [6]. Starting from an initial guess x_0, the iterates are given by

$$x_{k+1} = \mathcal{S}_{\mathbf{t},p}\left(x_k + A^*(y - Ax_k)\right), \qquad k = 1, 2, \dots \tag{56}$$

where the tresholding operator $\mathcal{S}_{\mathbf{t},p}(h) := \sum_{\lambda\in\Lambda} S_{\tau_\lambda,p}(\langle h, \psi_\lambda\rangle)\psi_\lambda$ is defined componentwise. For $p = 1$ we have

$$S_{\tau_\lambda,1}(x) := \begin{cases} x - \frac{\tau}{2} & x \geq \frac{\tau}{2} \\ 0 & |x| < \frac{\tau}{2} \\ x + \frac{\tau}{2} & x \leq -\frac{\tau}{2} \end{cases}.$$

The tresholding parameter τ depends on the regularization parameter $\hat{\alpha}$ and the weights $2^{\varsigma p|\lambda}$ from the definition of the Besov-space norm $||\cdot||_{B^s_p(\mathbb{R}^d)}$ in (18). Written in full dependence of all parameters, $\tau = \alpha\sigma^2 2^{s+d(\frac{1}{2}-\frac{1}{p})p|\lambda|}$ where $|\lambda|$ is the scale of the wavelet. Since $||\mathbf{A}|| < 1$, the algorithm converges to $x^{MAP}_{\hat{\alpha}}$. Because the kernel is symmetric we have $A^* = A$ and (56) can easily be implemented using Fourier transformation and the convolution theorem. In order to calculate the regularization parameter $\hat{\alpha} = \alpha\sigma^2$, we have to solve (43) or (50), respectively, for α. This can be done with Newton's method after obtaining a good initial guess, for example with the bisection method. A typical situation is shown in figure 1.

Each of the two definitions of the Besov space random variable allows for a slightly different implementation of the parameter choice rule. To illustrate the behaviour of both variants, we consider an academic example of

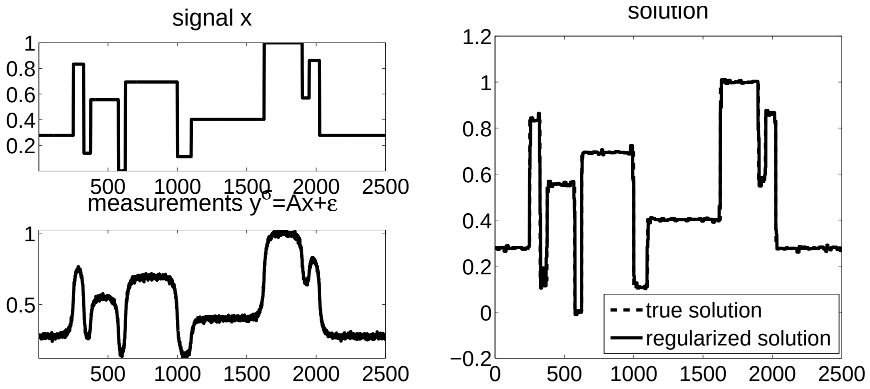

Figure 2: Signal, noisy measurements and regularized solution for $\sigma = 0.01$, obtained with our parameter choice rule for the finite dimensional model (MII). We used the exact ϱ and chose $s = 1$ such that $\varsigma = 0.5$. Resulting from $\alpha = 45.85$ as solution of (49) we obtained the effective regularization parameter $\hat{\alpha} = \alpha \cdot \sigma^2 = 0.0045$.

a one dimensional signal $\mathbf{x}$ that is sparse with respect to the Haar basis in $L_2(\mathbb{R})$ and its convolution with a kernel of type (55). A sample of signal, measurements and corresponding regularized solution is shown in figure 2, where $m = n = 2500$ and $\beta = 1$. Next we want to compare the predicted convergence properties to the numerical results. The behaviour of our parameter choice rules (43) and (50) with respect to σ is demonstrated in figure 3. Both models (MI) and (MII) lead to parameters α and $\hat{\alpha}$ fulfilling the theoretical conditions. The numerically obtained errors follow the theoretically predicted convergence rates.
So far we did not address the question of convergence of the solutions if m and n are not fixed anymore but increasing. Although theoretical results are missing at this point, figure 4 shows a comparison of the parameter choice rules for the models (MI) and (MII). In contrary to the convergence with respect to decreasing variance σ, the two models show a distinct convergence behaviour with respect to m. It is future work to give detailed analysis on this.

Until now we only considered the one dimensional case. However, figure 5 shows that also for two dimensional problems our parameter choice rule works and leads to reasonable reconstructions. figure 6 shows that in the 2-dimensional case the regularization parameter $\hat{\alpha} = \alpha\sigma^2$ with α chosen ac-

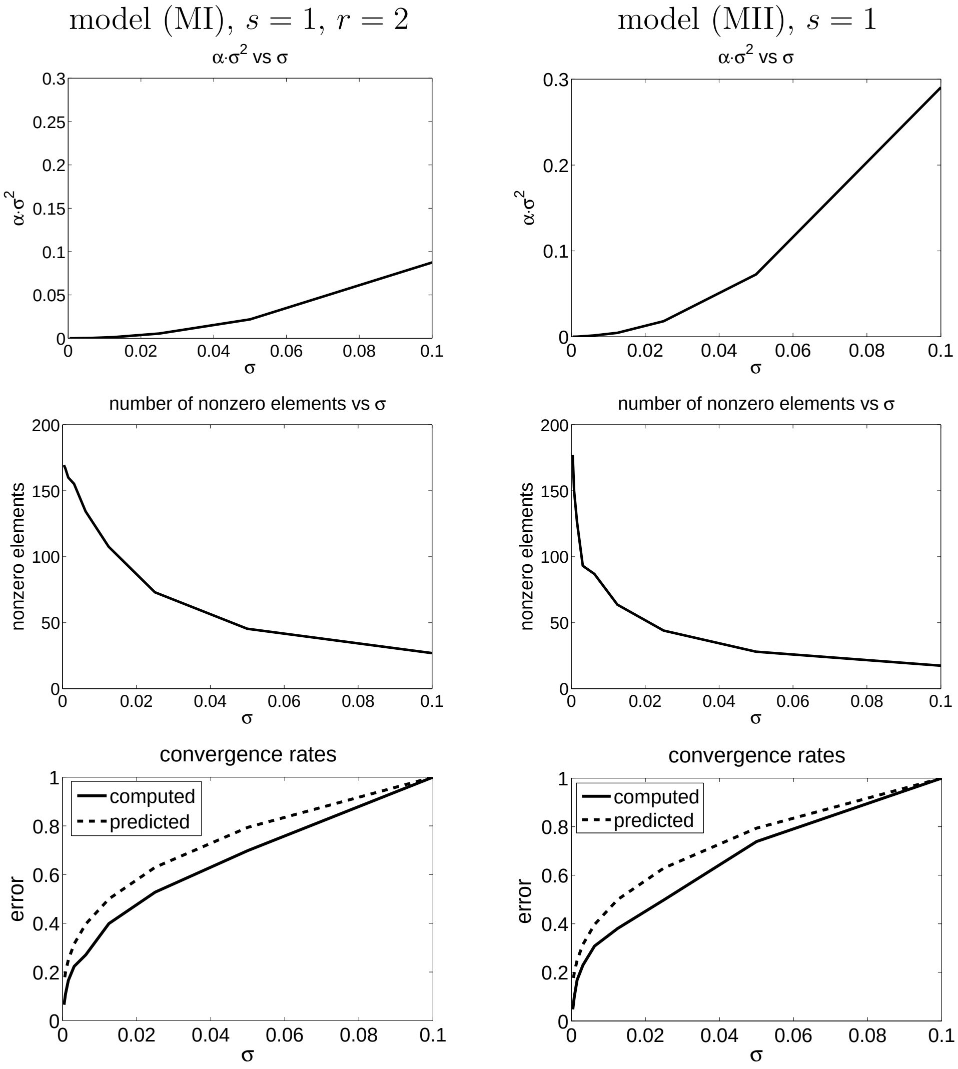

Figure 3: Numerical comparison of the two parameter choice rules (50) and (43), respectively, for the one dimensional deconvolution problem. The values for $\alpha{\cdot}\sigma^2$, plotted against σ, are shown in the first row. For sufficiently small σ, α starts to grow, following the theory developed in Corollary 5.7. However, $\alpha \cdot \sigma^2$ still goes to zero. The number of recovered non-zero coefficients in the solution is shown in the second row. Since $n = 2500$ we end up with a sparse solution for all σ used in the simulations. In the last row we plotted the obtained reconstruction error $||x^{MAP} - x^*||$ and compare it with the convergence rates $O\left((\sigma\sqrt{m})^{\frac{\varsigma}{\varsigma+\beta}}\right)$ predicted in Theorem 6.4 and Corollary 6.5, respectively.

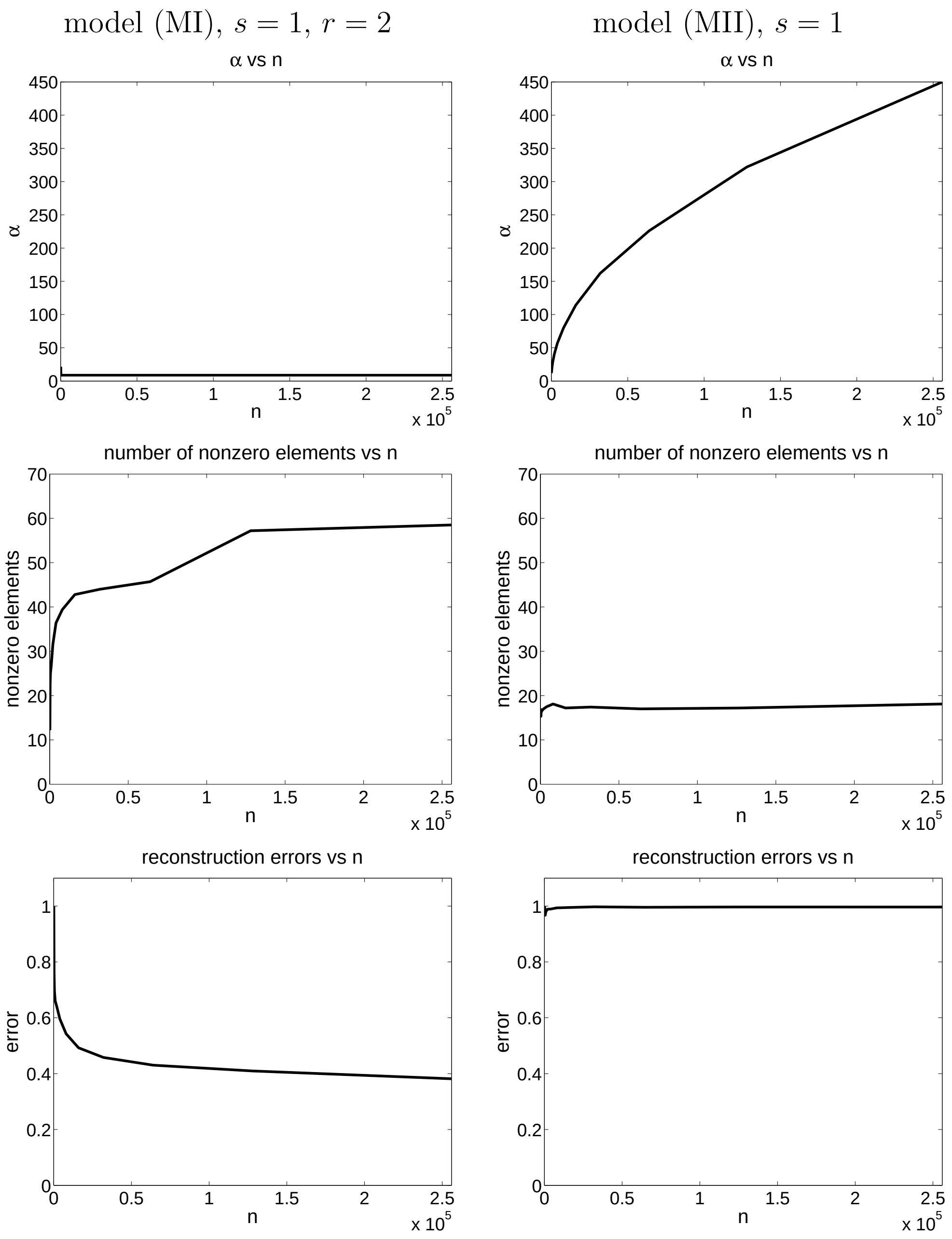

Figure 4: Comparison of the two implementations of the parameter choice rule with respect to varying level of discretization n. The variance $\sigma = 0.01$ was held constant. As estimate for ϱ we used the exact value calculated from the true solution scaled according to the respective models. While the finite model leads to growing α for increasing n, the infinite model keeps the regularization parameter constant. The number of non-zero coefficients in the recovered solution behaves conversely. To plot the reconstruction errors we used the same scaling for both models. While in the finite model the error stays the same, it decreases for growing n in the infinite model.

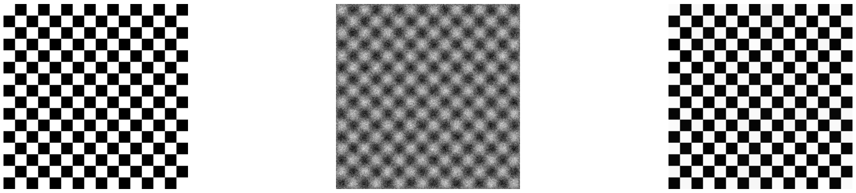

Figure 5: 2D example, left: signal x with 68 non-zero coefficients, middle: measurements $Ax+\epsilon$, $\sigma = 0.1$, $\beta = 1$, right: solution with $\alpha = 130.5$, exactly the 68 true coefficients were recovered.

cording to (50) for the infinite dimensional model (MI) keeps the number of recovered non-zero coefficients nearly constant as $\sigma \to 0$.

8 Conclusions

We have investigated convergence properties of Tikhonov regularization in a stochastic setting. Aiming for the maximum a posteriori solution, the Tikhonov-type functional (13) was derived from a Bayesian approach. We proved convergence of the regularized solution to the true solution in the Ky Fan metric. The special properties of this metric allowed us to establish non-standard parameter choice rules (43) and (50), respectively, leading to convergence rates.

Acknowledgements

D. Gerth and R. Ramlau were funded by the Austrian Science Fund (FWF): W1214-N15, project DK08. The authors would like to thank Samuli Siltanen (University of Helsinki, Finland) and his group for the fruitful discussions.

References

References

[1] Abramowitz M and Stegun I A 1964 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (National Bu-*

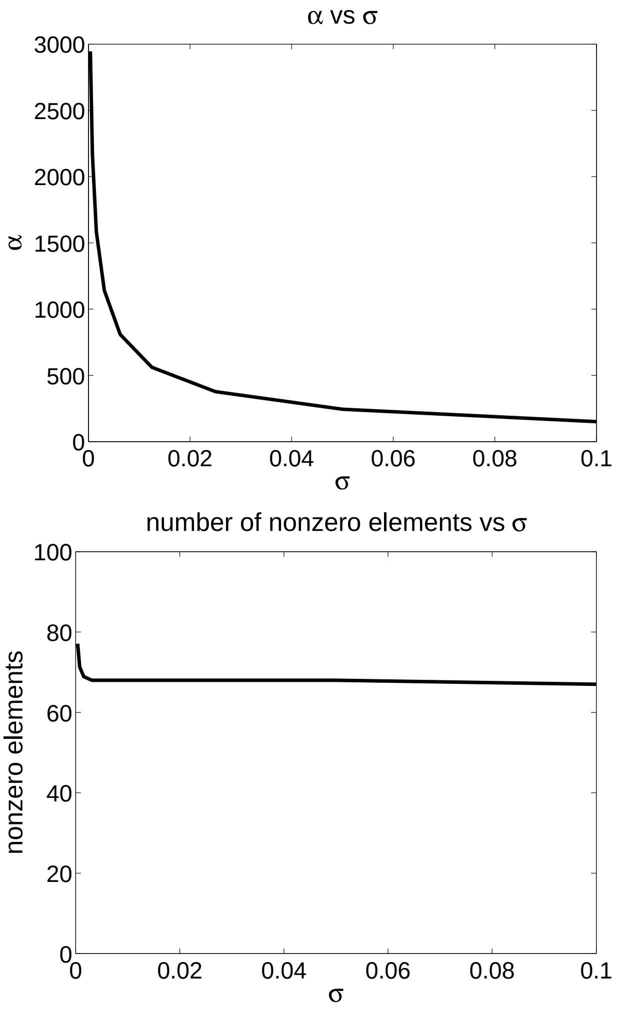

Figure 6: Upper plot: regularization parameter α against σ. α grows for decreasing σ according to Corollary 5.7; lower plot: number of recovered non-zero elements against σ. Out of $n = 65536$ coefficients of the original image 68 had non-zero values.

reau of Standards Applied Mathematics Series vol 55) (Washington D.C.: U.S. Government Printing Office)

[2] Blake I F 1979 *An Introduction to Applied Probability* (New York-Chichester-Brisbane: John Wiley & Sons)

[3] Bredies K 2009 An iterative thresholding-like algorithm for inverse problems with sparsity constraints in Banach space *J. Inverse Ill-Posed Probl.* **17** 1 19–26

[4] Corless R M, Gonnet G H, Hare D E G, Jeffrey D J and Knuth D E 1996 On the Lambert W function *Adv. Comput. Math.* **5** 4 329–59

[5] Daubechies I 1992 *Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics vol 61)* (Philadelphia: SIAM)

[6] Daubechies I, Defrise M and De Mol C 2004 An iterative thresholding algorithm for linear inverse problems with a sparsity constraint *Comm. Pure Appl. Math.* **57** 11 1413–57

[7] Dudley R M 1989 *Real Analysis and Probability (The Wadsworth & Brooks/Cole Mathematics Series)* (Pacific Grove: Wadsworth & Brooks/Cole Advanced Books & Software)

[8] Egoroff D Th 1911 Sur les suites de fonctions mesurables *C. R.* **152** 244–246

[9] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems (Mathematics and its Applications vol 375)* (Dordrecht: Kluwer Academic Publishers Group)

[10] Fan K 1944 Entfernung zweier zufälligen Grössen und die Konvergenz nach Wahrscheinlichkeit *Math. Z.* **49** 681–83

[11] Hofinger A 2006 Ill-posed problems: Extending the Deterministic Theory to a Stochastic Setup *PhD-thesis (Johannes Kepler University Linz)*

[12] Hofinger A and Pikkarainen H K 2007 Convergence rate for the Bayesian approach to linear inverse problems *Inverse Problems* **23** 6 2469–84

[13] Hofmann B 1986 *Regularization for Applied Inverse and Ill-posed Problems* (Leipzig: B. G. Teubner)

[14] Kaipio J and Somersalo E 2005 *Statistical and Computational Inverse Problems (Applied Mathematical Sciences vol 160)* (New York: Springer)

[15] Kolehmainen V, Lassas M, Niinimäki K and Siltanen S 2012 Sparsity-promoting Bayesian inversion *Inverse Problems* **28** 2

[16] Lassas M, Saksman E and Siltanen S 2009 Discretization-invariant Bayesian inversion and Besov space priors *Inverse Probl. Imaging* **3** 1 87–122

[17] Lassas M and Siltanen S 2004 Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems* **20** 5 1537–63

[18] Louis A K 1989 *Inverse und schlecht gestellte Probleme* (Stuttgart: B. G. Teubner)

[19] Meyer Y 1992 *Wavelets and Operators (Cambridge Studies in Advanced Mathematics vol 37)* (Cambridge: Cambridge University Press)

[20] Neubauer A and Pikkarainen H K 2008 Convergence results for the Bayesian inversion theory *J. Inverse Ill-Posed Probl.* **16** 6 601–13

[21] Ramlau R 2008 Regularization properties of Tikhonov regularization with sparsity constraints *Electron. Trans. Numer. Anal.* **30** 54–74

[22] Ramlau R and Teschke G 2006 A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints *Numer. Math.* **104** 2 177–203

[23] Vänskä S, Lassas M and Siltanen S 2009 Statistical X-ray tomography using empirical Besov priors *Int. J. Tomogr. Stat.* **11** S09 3–32

Technical Reports of the Doctoral Program "Computational Mathematics"

2013

2013-01 U. Langer, M. Wolfmayr: *Multiharmonic Finite Element Analysis of a Time-Periodic Parabolic Optimal Control Problem* January 2013. Eds.: W. Zulehner, R. Ramlau

2013-02 M.T. Khan: *Translation of* MiniMaple *to Why3ML* February 2013. Eds.: W. Schreiner, F. Winkler

2013-03 J. Kraus, M. Wolfmayr: *On the robustness and optimality of algebraic multilevel methods for reaction-diffusion type problems* March 2013. Eds.: U. Langer, V. Pillwein

2013-04 H. Rahkooy, Z. Zafeirakopoulos: *On Computing Elimination Ideals Using Resultants with Applications to Gröbner Bases* May 2013. Eds.: B. Buchberger, M. Kauers

2013-05 G. Grasegger: *A procedure for solving autonomous AODEs* June 2013. Eds.: F. Winkler, M. Kauers

2013-06 M.T. Khan *On the Formal Verification of Maple Programs* June 2013. Eds.: W. Schreiner, F. Winkler

2013-07 P. Gangl, U. Langer: *Topology Optimization of Electric Machines based on Topological Sensitivity Analysis* August 2013. Eds.: R. Ramlau, V. Pillwein

2013-08 D. Gerth, R. Ramlau: *A stochastic convergence analysis for Tikhonov regularization with sparsity constraints* October 2013. Eds.: U. Langer, W. Zulehner

The complete list since 2009 can be found at
`https://www.dk-compmath.jku.at/publications/`

Doctoral Program
“Computational Mathematics”

Director:

Prof. Dr. Peter Paule
Research Institute for Symbolic Computation

Deputy Director:

Prof. Dr. Bert Jüttler
Institute of Applied Geometry

Address:

Johannes Kepler University Linz
Doctoral Program “Computational Mathematics”
Altenbergerstr. 69
A-4040 Linz
Austria
Tel.: ++43 732-2468-6840

E-Mail:

office@dk-compmath.jku.at

Homepage:

http://www.dk-compmath.jku.at