CONVERGENCE RATES OF GENERAL REGULARIZATION METHODS FOR STATISTICAL INVERSE PROBLEMS AND APPLICATIONS

BY N. BISSANTZ[1], T. HOHAGE[2], A. MUNK[2] AND F. RUYMGAART[3]

[1]*UNIVERSITY OF BOCHUM*, [2]*UNIVERSITY OF GÖTTINGEN*, [3]*TEXAS TECH UNIVERSITY*

Abstract. During the past the convergence analysis for linear statistical inverse problems has mainly focused on spectral cut-off and Tikhonov type estimators. Spectral cut-off estimators achieve minimax rates for a broad range of smoothness classes and operators, but their practical usefulness is limited by the fact that they require a complete spectral decomposition of the operator. Tikhonov estimators are simpler to compute, but still involve the inversion of an operator and achieve minimax rates only in restricted smoothness classes. In this paper we introduce a unifying technique to study the mean square error of a large class of regularization methods (spectral methods) including the aforementioned estimators as well as many iterative methods, such as ν-methods and the Landweber iteration. The latter estimators converge at the same rate as spectral cut-off, but only require matrix-vector products. Our results are applied to various problems, in particular we obtain precise convergence rates for satellite gradiometry, L^2-boosting, and errors in variable problems.

AMS subject classifications. 62G05, 62J05, 62P35, 65J10, 35R30

Key words. Statistical inverse problems, iterative regularization methods, Tikhonov regularization, nonparametric regression, minimax convergence rates, satellite gradiometry, Hilbert scales, boosting, errors in variable

1. Introduction. This paper is concerned with estimating an element f of a Hilbert space $\mathbb{H}_1$ from indirect noisy measurements

$$Y = Kf + \text{"noise"} \tag{1.1}$$

related to f by a (known) operator $K : \mathbb{H}_1 \to \mathbb{H}_2$ mapping $\mathbb{H}_1$ to another Hilbert space $\mathbb{H}_2$. The operator K is assumed to be linear, bounded, and injective, but *not* necessarily compact. We are interested in the case that the operator equation (1.1) is ill-posed in the sense that the Moore-Penrose inverse of K is unbounded. The analysis of regularization methods for the stable solution of (1.1) depends on the mathematical model for the noise term on the right hand side of (1.1): If the noise is considered as a deterministic quantity, it is natural to study the worst-case error. In the literature a number of efficient methods for the solution of (1.1) has been developed, and it has been shown under certain conditions that the worst-case error converges at optimal order as the noise level tends to 0 (see Engl *et al.* [14]). If the noise is modeled as a random quantity, the convergence of estimators $\hat{f}$ of f should be studied in statistical terms. Here we consider the expected square error $\mathbf{E}\,\|\hat{f} - f\|^2$, also called *mean integrated square error* (MISE). This problem has also been studied extensively in the statistical literature, but the numerical efficiency has not been a major issue so far. It is the purpose of this paper to provide an analysis of a class of computationally efficient regularization methods including Landweber iteration, ν-methods, and iterated Tikhonov regularization, which is applicable to linear inverse problems with random noise as they occur for example in parameter identification problems in partial differential equations, deconvolution or errors in variable models.

There exists a considerable amount of literature on regularization methods for linear inverse problems with random noise. For surveys we refer to O'Sullivan [37], Nychka & Cox [36], Evans & Stark [16] and Kaipio & Somersalo [26]. A large part of the literature focusses on methods which require the explicit knowledge of a spectral decomposition of the operator K^*K. The simplest of these methods is spectral cut-off (or truncated singular value decomposition for compact operators) where an estimator is constructed by a truncated expansion of f w.r.t. the eigenfunctions of K^*K (e.g. Diggle & Hall [10], Healy, Hendriks & Kim [21]). It has been shown in a number of papers that spectral cut-off estimators are order optimal in a minimax-sense under certain conditions (e.g. Mair & Ruymgaart [30], Efromovich [13], Kim & Koo [27]). Based on a singular value decomposition (SVD) of K it is also possible to construct exact minimax estimators for given smoothness classes (see Johnstone & Silverman [25]).

Another major approach are wavelet-vaguelette (and vaguelette-wavelet) based methods which lead to estimators of a similar functional form as SVD methods. In general these estimators are based on expansions of f and Kf with respect to different bases of the respective function spaces than those provided by the SVD of K (e.g. Donoho [11], Abramovich & Silverman [1], Johnstone et *al.* [24]).

A well-known method both in the statistical and the deterministic inverse problems literature is Tikhonov regularization. This has been studied for certain classes of linear statistical inverse problems by Cox [9], Nychka & Cox [36] and Mathé & Pereverzev [31, 33].

The main restriction of the usefulness of spectral cut-off and related estimators is the need of the spectral data of the operator (i.e. an SVD if K is compact) to implement these estimators. This is known explicitly only in a limited number of special cases, and numerical computation of the spectral data is prohibitively expensive for many situations. Although Tikhonov regularization does not require the spectral data of the operator, there is still a need of setting up and inverting a matrix representing the operator. For iterative regularization methods such as Landweber iteration or ν-methods (see Nemirovskii & Polyak [35], Brakhage [6], and Engl *et al.* [14]) only matrix-vector multiplications are required. Furthermore, it is known that Tikhonov regularization achieves minimax rates of convergence only in a restricted number of smoothness classes, which is highlighted by the fact that its qualification number is 1, whereas Landweber iteration has infinitely large qualification, and $\nu-$methods with qualification ν are available for every $\nu > 0$ (see [14]).

Iterative regularization methods are particularly attractive for inverse problems in partial differential equations (pde's). Here the operator K maps an unknown parameter f in a pde to (part of) the solution to this pde. Hence, applying K to a vector f simply means solving the pde with the parameter f, whereas inverting or even setting up the matrix for K is often not feasible. We will discuss two linear inverse problem for pde's (the backwards heat equation and satellite gradiometry) in §5. However, most inverse problems for pde's are nonlinear even if the pde is linear. Such problems are often solved by regularized Newton methods. In this case the methods and the analysis of this paper can be applied to the linearized operator equations in each Newton step as discussed in the forthcoming paper [2].

In this paper we will show that general spectral regularization methods as defined in section 2 achieve the same rates of convergence of the MISE as spectral cut-off, which is known to be optimal in most cases (see above). Whereas the bias or approximation error is exactly the same in a deterministic and a statistical framework, the analysis significantly differs in the estimation of the noise term. In spectral cut-off for compact operators, the noise (or variance) part of the estimators $\hat{f}_\alpha$ belongs to a finite-dimensional space of "low-frequencies". The main difficulty in the analysis of general spectral regularization methods is the estimation of the "high-frequency" components of the noise. Unlike in a deterministic framework, the bound on the noise term depends not only on the regularization parameter, but also on the distribution of the singular values of K (if K is compact). Therefore, a statistical analysis has to impose additional conditions on the operator. We will verify these conditions for several important problems including inverse problems in partial differential equations and errors in variable models. As an example of particular interest in the machine learning context we obtain optimal rates of convergence of L^2-boosting by interpreting L^2-boosting as a Landweber iteration (see also Bühlmann & Yu [7], Yao *et al.* [42]).

The plan of this paper is as follows: The following section gives a brief overview of regularization methods and source conditions and introduce an abstract noise model. §3 contains the main results of this paper on the rates of convergence of general spectral regularization methods. In §4 we demonstrate how a number of commonly used statistical noise models fit into our general framework. Finally, in §5 we discuss the application of our results to the backwards heat equation, satellite gradiometry, errors in variable models with dependent random variables, L^2-boosting, and operators in Hilbert scales. Proofs of §3 are collected in §6.

2. Framework. We first review some basic notions of regularization theory.

2.1. Spectral theorem. Halmos' version of the spectral theorem (see, for instance, Halmos [20], Taylor [40]) turns out to be particularly convenient for the construction and statistical analysis of regularized inverses of a self-adjoint operator. This has been demonstrated by Mair & Ruymgaart [30] for the spectral cut-off estimator. The theorem claims that for a (not necessarily bounded) self-adjoint operator $A : D(A) \to \mathbb{H}$ defined on a dense subset $D(A)$ of a separable Hilbert space $\mathbb{H}$ there exists a σ-compact space $\mathbb{S}$, a Borel measure Σ on $\mathbb{S}$, a unitary operator $U : \mathbb{H} \to L^2(\Sigma)$, and a measurable function $\rho : \mathbb{S} \to \mathbb{R}$ such that

$$UAf = \rho \cdot Uf, \qquad \Sigma\text{-almost everywhere}, \tag{2.1}$$

for all $f \in D(A)$. Introducing the multiplication operator $M_\rho : D(M_\rho) \to L^2(\Sigma)$, $M_\rho \varphi := \rho \cdot \varphi$ defined on $D(M_\rho) := \{\varphi \in L^2(\Sigma) : \rho\varphi \in L^2(\Sigma)\}$, we can rewrite (2.1) as $A = U^* M_\rho U$, i.e. A is unitarily equivalent to a multiplication operator. The essential range of ρ is the spectrum $\sigma(A)$ of A. If A is bounded and positive definite as below, then $0 < \rho \leq \|A\|$, Σ-a.e.

Remark 1. *In the special case that A is compact, a well-known version of the spectral theorem states that A has a complete orthonormal system of eigenvectors u_i with corresponding eigenvalues ρ_i, and $Af = \sum_{j=0}^{\infty} \rho_j \langle u_j, f\rangle u_j$. This can be rewritten in the multiplicative form (2.1) by choosing Σ as the counting measure on $\mathbb{S} = \mathbb{N}$, i.e. $L^2(\Sigma) = l^2(\mathbb{N})$, the multiplicator function as $\rho(i) = \rho_i$, $i \in \mathbb{N}$, and defining the unitary operator $U : \mathbb{H} \to l^2(\mathbb{N})$ by $(Uf)(i) := \langle f, u_i\rangle$, $i \in \mathbb{N}$.*

2.2. Regularized estimators. Recall Halmos' spectral theorem from §2.1. For a self-adjoint operator $A : D(A) \to \mathbb{H}$ and a bounded, measurable function $\Phi : \sigma(A) \to \mathbb{R}$ one defines an operator $\Phi(A) \in L(\mathbb{H})$ by

$$\Phi(A) = U^* M_{\Phi(\rho)} U, \tag{2.2}$$

(see e.g. Taylor [40]). The mapping $\Phi \mapsto \Phi(A)$, called the *functional calculus* at A, is an algebra homomorphism from the algebra of bounded measurable functions on $\sigma(A)$ to the algebra $L(\mathbb{H})$ of bounded linear operators on $\mathbb{H}$, and

$$\|\Phi(A)\| \leq \sup_{\lambda \in \sigma(A)} |\Phi(\lambda)|, \tag{2.3}$$

with equality if Φ is continuous. We will construct estimators of the input function by regularization methods of the form

$$\hat{f}_{\alpha,\sigma} = \Phi_\alpha(K^*K)K^*Y. \tag{2.4}$$

Here $\Phi_\alpha : \sigma(K^*K) \to \mathbb{R}$ is a collection of bounded filter functions approximating the unbounded function $t \mapsto \frac{1}{t}$ on $\sigma(K^*K)$, which are parametrized by a *regularization parameter* $\alpha > 0$.

A particular example of a regularization method of the form (2.4) is the *spectral cut-off* estimator (also known as *truncated singular value decomposition*) described by the functions

$$\Phi_\alpha^{\mathrm{SC}}(t) := \begin{cases} t^{-1}, & t \geq \alpha, \\ 0, & t < \alpha. \end{cases}$$

As explained in the introduction, we will focus on regularization methods which can be implemented without explicit knowledge of the spectral decomposition of the operator K^*K. This includes both *implicit methods* such as Tikhonov regularization ($\Phi_\alpha(t) = (\alpha + t)^{-1}$), iterated Tikhonov regularization and Lardy's method, which involve the inversion of an operator and *explicit methods* such as Landweber iteration ($\Phi_{1/(k+1)}(t) = \sum_{j=0}^{k-1}(1-\beta t)^j$ where $\beta \in (0, \|K^*K\|^{-2}$ is a step-length parameter) and ν-methods, which require only matrix-vector products in a discrete setting. For a derivation and discussion of these methods we refer to the monograph [14].

2.3. Smoothness classes. We will measure the smoothness of the input function f relative to the smoothing properties of K in terms of *source conditions*: Let $\Lambda : [0,\infty) \to [0,\infty)$ be a continuous, strictly increasing function with $\Lambda(0) = 0$, and assume that there exists a "source" $w \in \mathbb{H}_1$ such that

$$f = \Lambda(K^*K)w \tag{2.5}$$

(see [14, 15, 32]). The set of all f satisfying this condition with $\|w\|_{\mathbb{H}_1} \leq \overline{w}$, $\overline{w} > 0$ will be denoted by $F_{\Lambda,\overline{w},K^*K} := \{\Lambda(K^*K)w : w \in \mathbb{H}_1, \|w\| \leq \overline{w}\}$. We will shortly write $F_{\Lambda,\overline{w}} := F_{\Lambda,\overline{w},K^*K}$ if there is no ambiguity. The most common choice, which is usually appropriate for finitely smoothing operators K is

$$\Lambda(t) = t^\nu, \qquad \nu > 0. \tag{2.6}$$

In particular, (2.5) with $\Lambda(t) = \sqrt{t}$ is equivalent to $f = K^*v$, $\|v\|_{\mathbb{H}_2} \le 1$ (see Engl *et al.* [14, Prop. 2.18]). For exponentially ill-posed problems such as the backwards heat equation, (2.6) is usually too restrictive and *logarithmic source conditions* corresponding to the choice

$$\Lambda(t) = (-\ln t)^{-p}, \qquad p > 0, \tag{2.7}$$

are more appropriate (see Hohage [23], Mair [29]). Since Λ is singular at $t = 1$, we assume that the norms in $\mathbb{H}_1$ and $\mathbb{H}_2$ are scaled such that $\|K^*K\| < 1$ in this case. For a further discussion of source conditions and interpretations as smoothness conditions in Sobolev spaces we refer to the applications in §5.

If f belongs to the smoothness class $F_{\Lambda,\overline{w}}$ and we are given exact data $Y = g$, then the error is bounded by

$$\|\Phi_\alpha(K^*K)K^*g - f\| = \|(\Phi_\alpha(K^*K)K^*K - I)\Lambda(K^*K)w\| \le \sup_{t\in\sigma(K^*K)} |(\Phi_\alpha(t)t - 1)\Lambda(t)|\overline{w}, \tag{2.8}$$

where we have used (2.3).

2.4. Assumptions on smoothness and the regularization method. In the following we discuss a number of standard assumptions on the filter functions Φ_α satisfied for all commonly used regularization methods, in particular those discussed in §2.2 (see [14]). First, we assume that there exists a constant $C_2 > 0$ such that

$$\sup_{t\in\sigma(K^*K)} |t\Phi_\alpha(t)| \le C_2, \qquad \text{uniformly in } \alpha > 0. \tag{2.9a}$$

To bound the so-called propagated deterministic noise error $\tau\|\Phi_\alpha(K^*K)K^*\xi\|$, we impose the condition

$$\text{there exists } C_3 > 0: \ \sup_{\alpha>0} \sup_{t\in\sigma(K^*K)} |\alpha\Phi_\alpha(t)| \le C_3. \tag{2.9b}$$

In view of the bound (2.8) on the approximation error, we also assume that there exists a number $\nu_0 > 0$ called *qualification* of the method and constants $\gamma_\nu > 0$ such that

$$\sup_{t\in\sigma(K^*K)} |t^\nu(1 - t\Phi_\alpha(t))| \le \gamma_\nu\alpha^\nu, \qquad \text{for all } \alpha \text{ and all } 0 \le \nu \le \nu_0. \tag{2.9c}$$

The qualification of a method is a measure of the maximal degree of smoothness in terms of the Hölder-type conditions (2.5), (2.6) under which the approximation error (2.8) converges at optimal order. The qualification some commonly used methods is: Tikhonov regularization: 1, K-times iterated Tikhonov regularization: K, Landweber iteration: ∞ (in the sense that it is greater than any real number), ν-methods: ν (where $\nu > 0$ is a parameter in the method), see references in the Introduction.

Note that the condition (2.9c) with $\nu_0 > 0$ implies that $\lim_{\alpha\searrow 0} \Phi_\alpha(t) = \frac{1}{t}$ for all $t \in \sigma(K^*K)$. For $\nu = 0$ the condition (2.9c) implies (2.9a) with $C_2 = 1+\gamma_0$. However, this value of C_2 is usually not optimal as for most regularization methods (2.9a) holds true with $C_2 = 1$.

For general source conditions we assume that there exists a constant γ_Λ such that

$$\sup_{t\in\sigma(K^*K)} |\Lambda(t)(1 - t\Phi_\alpha(t))| \le \gamma_\Lambda\Lambda(\alpha), \qquad \alpha \searrow 0. \tag{2.10}$$

Under Hölder-type source conditions (2.6) this holds true for $\nu \le \nu_0$ by assumption (2.9c). For the choice $\Lambda(t) = (-\ln t)^{-p}$, it has been shown in Hohage [23] that (2.9c) with $\nu_0 > 0$ implies (2.10). For more general functions Λ we refer to Mathé & Pereverzev [32] for similar implications.

2.5. Noise model. In this subsection we introduce an abstract noise model which will be used in the proof of our main result. In §4 we will demonstrate that several noise models commonly encountered in statistical modelling fit into this general framework.

Following Mathé & Pereverzev [31] we assume that our given data can be written as

$$Y = g + \sigma\varepsilon + \tau\xi, \qquad g := Kf, \tag{2.11}$$

where $\xi \in \mathbb{H}_2$, $\|\xi\| \leq 1$ is a deterministic error, ε is a stochastic error, and $\tau, \sigma > 0$ are the corresponding noise levels. Note that model (2.11) allows for stochastic and deterministic noise, simultaneously.

Often, the stochastic error is modelled as a Hilbert-space valued random variable Ξ, i.e. a measurable function $\Xi : \Omega \to \mathbb{H}_2$ where $(\Omega, \mathcal{P}, P)$ is the underlying probability space. However, we will assume more generally that it is a Hilbert-space process, i.e. a continuous linear operator

$$\varepsilon : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P).$$

Every Hilbert-space valued random variable Ξ with finite second moments, $\mathbf{E}\,\|\Xi\|^2 < \infty$ can be identified with a Hilbert-space process $\varphi \mapsto \langle \Xi, \varphi \rangle$, $\varphi \in \mathbb{H}_2$, but not vice versa. We will use the notation $\langle \varepsilon, \varphi \rangle := \varepsilon\varphi$, $\varphi \in \mathbb{H}_2$. The covariance $\mathbf{Cov}_\varepsilon : \mathbb{H}_2 \to \mathbb{H}_2$ of a Hilbert-space process $\varepsilon : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P)$ is the bounded linear operator defined implicitly by $\langle \mathbf{Cov}_\varepsilon \varphi_1, \varphi_2 \rangle = \mathbf{Cov}\left(\langle \varepsilon, \varphi_1 \rangle, \langle \varepsilon, \varphi_2 \rangle\right)$, $\varphi_1, \varphi_2 \in \mathbb{H}_2$. We call ε a *white noise process* if $\mathbf{Cov}_\varepsilon = I$ and $\mathbf{E}\,\langle \varepsilon, \varphi \rangle = 0$ for all $\varphi \in \mathbb{H}_2$. Note that a Gaussian white noise process in an infinite-dimensional Hilbert space cannot be identified with a Hilbert-space valued random variable.

If $\varepsilon : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P)$ is a Hilbert-space process and $A : \mathbb{H}_2 \to \mathbb{H}_1$ is a bounded linear operator, we define the Hilbert-space process $A\varepsilon : \mathbb{H}_1 \to L^2(\Omega, \mathcal{P}, P)$ by $\langle A\varepsilon, \varphi \rangle := \langle \varepsilon, A^*\varphi \rangle$, $\varphi \in \mathbb{H}_1$. Its covariance operator is given by $\mathbf{Cov}_{A\varepsilon} = A\mathbf{Cov}_\varepsilon A^*$.

Assumption 1. *In the noise model (2.11) $\xi \in \mathbb{H}_2$ is a deterministic vector with $\|\xi\| = 1$, and ε is a Hilbert space process such that*

$$\mathbf{E}\,\langle \varepsilon, \varphi \rangle = 0, \qquad \|\mathbf{Cov}_\varepsilon\| \leq 1 \tag{2.12}$$

for all $\varphi \in \mathbb{H}_2$. Moreover, $K^*\varepsilon$ is a Hilbert-space valued random variable satisfying

$$\mathbf{E}\,\|K^*\varepsilon\|^2 < \infty, \tag{2.13}$$

and there exists a spectral decomposition (2.1) of K^*K such that for almost all $s \in \mathbb{S}$

$$\mathbf{Var}\left(UK^*\varepsilon(s)\right) \leq \rho(s). \tag{2.14}$$

The first condition in (2.12) is not a restriction since an expected value different from zero can be included in $\tau\xi$, and the second condition is a scaling condition analogous to $\|\xi\| \leq 1$. Assumption (2.13) ensures that the estimators defined in (2.4) are Hilbert-space valued random variables with finite second moments. (2.13) is usually a mild assumption, but it excludes e.g. very mildly ill-posed problems in combination with white noise. The following lemma implies that (2.14) is a condition on the choice of U in the Halmos representation (2.1) rather than a condition on the noise model. Moreover, we can arrange that $\rho \in L^1(\Sigma)$ as required in §3 below. Noise models with a finite number of observations satisfying Assumption 1 are discussed in §4 below.

Lemma 2. *If ε is a Hilbert space process satisfying (2.12), $K^*\varepsilon$ is a Hilbert space valued random variable satisfying (2.13), and K is injective, then there exists a spectral decomposition (2.1) of K^*K such that (2.14) holds true, and $\rho \in L^1(\Sigma)$.*

Proof. According to Halmos' spectral theorem there exists a Borel measure $\tilde{\Sigma}$ on a σ-compact space $\mathbb{S}$, and a unitary operator $\tilde{U} : L^2(\mathbb{R}^d) \to L^2(\tilde{\Sigma})$ such that $K^*K = \tilde{U}^* M_\rho \tilde{U}$. For any $\tilde{\Sigma}$-measurable function $\chi > 0$ on $\mathbb{S}$ we can construct another Halmos representation of K^*K by introducing the Borel measure $\Sigma := \chi\tilde{\Sigma}$ on $\mathbb{S}$ and the mapping $U : L^2(\mathbb{R}^d) \to L^2(\Sigma)$, $Uf :=$

$\chi^{-1/2} \cdot \tilde{U} f$ since U is unitary and $UK^*Kf = \rho \cdot Uf$ Σ-a.e. for all $f \in \mathbb{H}_1$. In particular, we may define

$$\chi(s) := \frac{\mathbf{Var}\,(\tilde{U} K^* \varepsilon)(s)}{\rho(s)} \qquad \text{for } s \in M, \quad M := \{s \in \mathbb{S} : \mathbf{Var}\,(\tilde{U} K^* \varepsilon)(s) > 0\}. \tag{2.15}$$

Here we use that $\rho > 0$ $\tilde{\Sigma}$-a.e since K and hence K^*K is injective by assumption. We first consider the case $\tilde{\Sigma}(M^c) = 0$ where $M^c := \mathbb{S} \setminus M$. Then (2.14) holds true for $s \in M$ as $\mathbf{Var}\,(UK^*\varepsilon)(s) = \chi(s)^{-1} \mathbf{Var}\,(\tilde{U} K^* \varepsilon)(s) = \rho(s)$. Moreover,

$$\int \rho \, \mathrm{d}\Sigma = \int \mathbf{Var}\,(UK^*\varepsilon) \, \mathrm{d}\Sigma = \mathbf{E} \int |UK^*\varepsilon|^2 \, \mathrm{d}\Sigma = \mathbf{E}\, \|K^*\varepsilon\|^2 < \infty, \tag{2.16}$$

which is the assertion. Now assume that $\tilde{\Sigma}(M^c) > 0$. Let ψ be an arbitrary strictly positive function in $L^1(\tilde{\Sigma})$, e.g. $\psi(s) := \left(j(s)^2 \tilde{\Sigma}(A_{j(s)})\right)^{-1}$, where $j(s) := \min\{j : \; s \in A_j\}$ for a sequence $A_1 \subset A_2 \subset \ldots \subset \Omega$ with $\tilde{\Sigma}(A_j) < \infty$ and $\tilde{\Sigma}(\mathbb{S} \setminus \bigcup_j A_j) = 0$. Such a sequence exists because $\tilde{\Sigma}$ is σ-finite. We define $\chi(s)$ by (2.15) for $s \in M$ and $\chi(s) := \frac{\psi(s)}{\rho(s)}$ for $s \in M^c$. Then (2.14) is trivially satisfied for $s \in M^c$, and $\rho \in L^1(\Sigma)$ since $\int_M \rho \, \mathrm{d}\Sigma < \infty$ as in (2.16) and $\int_{M^c} \rho \, \mathrm{d}\Sigma \leq \int \psi \, \mathrm{d}\tilde{\Sigma} < \infty$. This finishes the proof. ∎

3. MISE estimates. In this section the main results of this paper are presented. Recall the definition of the estimator $\hat{f}_{\alpha,\sigma}$ of the input function f in (2.4). Since $\mathbf{E}\, \Phi_\alpha(K^*K)K^*\varepsilon = 0$, the MISE satisfies the bias-variance decomposition

$$\mathbf{E}\, \|\hat{f}_{\alpha,\sigma} - f\|^2 = \mathbf{B}\,(\hat{f}_{\alpha,\sigma})^2 + \mathbf{E}\, \|\hat{f}_{\alpha,\sigma} - \mathbf{E}\, \hat{f}_{\alpha,\sigma}\|^2, \tag{3.1}$$

with the bias term $\mathbf{B}\,(\hat{f}_{\alpha,\sigma}) := \|\mathbf{E}\, \hat{f}_{\alpha,\sigma} - f\|$. As discussed in the introduction, the bias term can be bounded by standard estimates whereas the variance term requires a special treatment involving a splitting in the frequency domain.

3.1. Estimation of the bias. The bias in our model coincides with the error in a deterministic setting and can be estimated by standard techniques (see [14]). Using the triangle inequality, the noise model (2.11), (2.12), and the definition (2.4) of $\hat{f}_{\alpha,\sigma}$, we get

$$\mathbf{B}\,(\hat{f}_{\alpha,\sigma}) \leq \|\Phi_\alpha(K^*K)K^*Kf - f\| + \tau \|\Phi_\alpha(K^*K)K^*\xi\|.$$

The first term (called approximation error) is bounded by $\gamma_\Lambda \Lambda(\alpha) \overline{w}$ due to (2.8) and (2.10). For the second term (called propagated deterministic noise error) we obtain the bound

$$\|\Phi_\alpha(K^*K)K^*\xi\|^2 = \langle \Phi_\alpha(KK^*)\xi, KK^*\Phi_\alpha(KK^*)\xi \rangle \leq \frac{C_2 C_3}{\alpha} \tag{3.2}$$

using the identity $\Phi_\alpha(K^*K)K^* = K^*\Phi_\alpha(KK^*)$, (see [14, eq. (2.43)]) and (2.9). Hence,

$$\mathbf{B}\,(\hat{f}_{\alpha,\sigma}) \leq \gamma_\Lambda \Lambda(\alpha) \overline{w} + \sqrt{\frac{C_2 C_3}{\alpha}} \tau. \tag{3.3}$$

Since we aim to show optimality of general regularization methods by comparison to spectral cut-off (see Introduction and §3.3), we now compare the approximation errors of general regularization methods and spectral cut-off. To this end, we introduce the following notations.

Notation: For two real-valued functions f, g defined on an interval $(0, \bar{\alpha}]$ we write

$$f(\alpha) \sim g(\alpha) \qquad (\text{or } f(\alpha) \lesssim g(\alpha)), \qquad \alpha \searrow 0,$$

if $g(\alpha) \neq 0$ for α in some neighborhood of 0 and $\lim_{\alpha \searrow 0} \frac{f(\alpha)}{g(\alpha)} = 1$ or $\limsup_{\alpha \searrow 0} \frac{f(\alpha)}{g(\alpha)} \leq 1$. Furthermore, we write

$$f(\alpha) \asymp g(\alpha), \qquad \alpha \searrow 0,$$

if there exist constants $\bar{\alpha} > 0$ and $C_{\bar{\alpha}} \geq 1$ such that $(1/C_{\bar{\alpha}})f(\alpha) \leq g(\alpha) \leq C_{\bar{\alpha}} f(\alpha)$ for $0 < \alpha \leq \bar{\alpha}$.

Recall that $\Lambda : [0, \infty) \to [0, \infty)$ is assumed to be a strictly increasing, continuous function with $\Lambda(0) = 0$ and that $1 - t\Phi_\alpha^{\mathrm{SC}}(t) = \chi_{[0,\alpha]}(t)$, i.e. $(I - K^*K\Phi_\alpha^{\mathrm{SC}}(K^*K))$ is an orthogonal projection operator. Therefore,

$$\sup_{f \in F_{\Lambda,\overline{w}}} \|(I - K^*K\Phi_\alpha^{\mathrm{SC}}(K^*K))f\| = \sup_{t \in \sigma(K^*K)} (1 - t\Phi_\alpha^{\mathrm{SC}}(t))\Lambda(t)\overline{w} \sim \Lambda(\alpha)\overline{w}, \quad \alpha \searrow 0.$$

The last relation holds since 0 is not an isolated point of the spectrum $\sigma(K^*K)$ for ill-posed operator equations. Using (2.8) and (2.10) we obtain the estimate

$$\sup_{f \in F_{\Lambda,\overline{w}}} \|(I - K^*K\Phi_\alpha(K^*K))f\| \leq \gamma_\Lambda \Lambda(\alpha)\overline{w} \sim \gamma_\Lambda \sup_{f \in F_{\Lambda,\overline{w}}} \|(I - K^*K\Phi_\alpha^{\mathrm{SC}}(K^*K))f\| \tag{3.4}$$

as $\alpha \searrow 0$. For many regularization methods and smoothness classes we have $\gamma_\Lambda \leq 1$.

3.2. Estimation of the integrated variance and rate of convergence of the MISE. The more difficult part is the estimation of the integrated variance of the error $\hat{f}_{\alpha,\sigma} - f$. Under Assumption 1 we have

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2 = \sigma^2 \mathbf{E}\,\|\Phi_\alpha(\rho)UK^*\varepsilon\|^2 \leq \sigma^2 \int_{\mathbb{S}} \Phi_\alpha^2(\rho)\rho \,\mathrm{d}\Sigma. \tag{3.5}$$

A crucial point in the following analysis is the estimation of the tails of the spectral function ρ. To this end, we bound the variance in terms of the function

$$R(\alpha) := \Sigma(\{\rho \geq \alpha\}), \qquad \alpha > 0. \tag{3.6}$$

In order to control the MISE of $\hat{f}_{\alpha,\sigma}$ as $\alpha \searrow 0$ it is tempting to assume that R is smooth in a neighborhood around 0. However, this is not true in general. Therefore, we will pose instead that R can be approximated suitably by a smooth function S with similar properties as R, as $\alpha \searrow 0$. Obviously, R is monotonically decreasing (see (3.8a) below). If $\rho \geq 0$ belongs to $L^1(\Sigma)$, then $-\int_0^\infty \alpha \,\mathrm{d}R(\alpha) = \int_{\mathbb{S}} \rho \,\mathrm{d}\Sigma < \infty$ (see (3.8b)), and it follows from Lebesgue's dominated convergence theorem that $\lim_{\alpha\searrow 0} \alpha R(\alpha) = \lim_{\alpha\searrow 0} \int_{\mathbb{S}} \alpha \, 1_{\{\rho\geq\alpha\}} \,\mathrm{d}\Sigma = 0$ (see (3.8c)).

Assumption 2. *There exists a constant $\bar{\alpha} \in (0, \|\rho\|_\infty]$ and a function $S \in C^2((0, \bar{\alpha}])$ such that*

$$R(\alpha) \sim S(\alpha), \qquad \alpha \searrow 0, \tag{3.7}$$

with R defined by (3.6) in terms of the spectral decomposition (2.1), and S satisfies

$$S' < 0, \tag{3.8a}$$

$$-\alpha S'(\alpha) \text{ is integrable on } (0, \bar{\alpha}], \tag{3.8b}$$

$$\lim_{\alpha\searrow 0} \alpha S(\alpha) = 0, \tag{3.8c}$$

$$\exists \gamma_{\mathrm{S}} \in (0, 2)\ \forall \alpha \in (0, \bar{\alpha}] : \frac{S''(\alpha)}{-S'(\alpha)} \leq \frac{\gamma_{\mathrm{S}}}{\alpha}. \tag{3.8d}$$

We will show in §5 for a number of examples that this assumption is satisfied. Now we are in the position to give an estimate of the MISE. The estimate of the MISE in the image space $\mathbb{H}_2$ in (3.10) is needed in the analysis of L^2-boosting (§5.4) and for nonlinear inverse problems.

Theorem 3. *Consider the model (2.11), and let Assumptions 1 and 2 hold true. We define a general spectral estimator $\hat{f}_{\alpha,\sigma}$ by (2.4) and assume that Φ_α satisfies (2.9).*

1. *If condition (2.10) is satisfied for the function Λ defining the smoothness class $F_{\Lambda,\overline{w},K^*K}$, then for all $f \in F_{\Lambda,\overline{w},K^*K}$ the MISE can be asymptotically bounded by*

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|_{\mathbb{H}_1}^2 \lesssim \left(\gamma_\Lambda \Lambda(\alpha)\overline{w} + \sqrt{\frac{C_2C_3}{\alpha}}\tau\right)^2 + \frac{(C_2^2 + C_3^2)\sigma^2}{\alpha^2} \int_0^\alpha S(\beta)\,\mathrm{d}\beta, \quad \alpha \searrow 0. \tag{3.9}$$

2. *Assume that $g \in F_{\tilde{\Lambda},\overline{w},KK^*} \subset \mathbb{H}_2$ and that $\tilde{\Lambda}$ satisfies (2.10). (If $g = Kf$ with $f \in F_{\Lambda,\overline{w},K^*K}$, then $\tilde{\Lambda}(t) := \sqrt{t}\Lambda(t)$, but we do* not *assume $g \in R(K)$ here!) Then*

$$\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - g\|^2_{\mathbb{H}_2} \lesssim \left(\gamma_{\tilde{\Lambda}}\tilde{\Lambda}(\alpha)\overline{w} + C_2\tau\right)^2 + \frac{(C_2^2 + C_3^2)\sigma^2}{\alpha^2}\int_0^\alpha \beta S(\beta)\,\mathrm{d}\beta, \qquad \alpha \searrow 0. \tag{3.10}$$

Note that for statistical inverse problems as opposed to deterministic inverse problems, the estimates of the noise term and hence the rates of convergence of the MISE do not only depend on the relative smoothness of the solution (i.e. on Λ), but also on the operator (i.e. on S).

Remark 4. *We comment on the choice of the regularization parameter $\alpha > 0$. If the noise levels σ and τ, the spectral properties of K^*K (i.e. S) and the smoothness of f (i.e. Λ) are known, one can choose α by minimizing the right hand side of (3.9). Since typically the smoothness of the solution is not known a-priori, so-called* adaptive *methods must be employed for the selection of α. We do not intend to review the considerable amount of literature on this topic here, but want to mention that the explicit bounds on the variance given in Theorem 3 allow the application of the Lepskij balancing principle as proposed for inverse problems by Mathé & Pereverzev [32, 33] and Bauer & Pereverzev [3]. We will discuss this in more detail elsewhere. With this method one typically loses a log factor in the asymptotic rates of convergence. In most cases this can be avoided by using Akaike's method as studied for spectral cut-off and related methods by Cavalier et al. [8]. Unfortunately, Assumption 2 in this paper is not satisfied for the methods discussed here.*

3.3. Comparison with spectral cut-off. To show that with an optimal choice of α our estimators can achieve the best possible order of convergence among all estimators as $\sigma \searrow 0$, we compare them to the spectral cut-off estimator for which minimax results are known in many situations (see references in the introduction). Since we are mainly interested in the case that the statistical noise is asymptotically dominant, we will assume that $\tau = 0$ for simplicity. Moreover, we assume in addition to (2.14) that the lower bound

$$\mathbf{Var}\,(UK^*\varepsilon(s)) \geq \gamma_{\mathbf{var}}\rho(s). \tag{3.11}$$

holds true for some constant $\gamma_{\mathbf{var}} > 0$. For the white noise model this is satisfied with $\gamma_{\mathbf{var}} = 1$ and for the inverse regression model with $\gamma_{\mathbf{var}} = C_{v,l}/C_1$ (see (4.13)). Moreover, we need the following assumption to prove optimal rates in many mildly ill-posed problems.

Assumption 3. *There exists a constant $C_4 > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$*

$$\frac{C_4}{\alpha} \leq \frac{-S'(\alpha)}{S(\alpha)}. \tag{3.12}$$

Theorem 5. *Let Assumptions 1 and 2 and the lower bound (3.11) hold true and assume that the family of functions $\{\Phi_\alpha\}$ satisfies (2.9). Moreover, assume that either $S = R$, or Assumption 3 holds true. Then the integrated variance of the estimator $\hat{f}_{\alpha,\sigma}$ is bounded by the integrated variance of the spectral cut-off estimator $\hat{f}^{\mathrm{SC}}_{\alpha,\sigma}$*

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2 \lesssim \frac{C_2^2 + \kappa C_3^2}{\gamma_{\mathbf{var}}}\,\mathbf{E}\,\|\hat{f}^{\mathrm{SC}}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}^{\mathrm{SC}}_{\alpha,\sigma}\|^2, \qquad \alpha \searrow 0, \tag{3.13}$$

with C_2 and C_3 as in Theorem 3 and $\kappa := \gamma_{\mathrm{S}}/(2-\gamma_{\mathrm{S}})$, γ_{S} defined in (3.8d). Moreover, if condition (2.10) is satisfied for the function Λ defining the smoothness class $F_{\Lambda,\overline{w}}$ and if $\tau = 0$, then there exists a constant $C > 0$ such that

$$\sup_{f\in F_{\Lambda,\overline{w}}} \mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|^2 \leq C \sup_{f\in F_{\Lambda,\overline{w}}} \mathbf{E}\,\|\hat{f}^{\mathrm{SC}}_{\alpha,\sigma} - f\|^2, \tag{3.14}$$

for all $\sigma > 0$ and all $\alpha > 0$ sufficiently small.

Whereas condition (3.12) is usually satisfied for mildly ill-posed problems, it is not satisfied for exponentially ill-posed problems where $S(\alpha) \sim c(-\ln\alpha)^q$ for constants $c, q > 0$. Nevertheless, the error bounds in Theorem 3 yield optimal rates of convergence in the limit $\sigma \searrow 0$ for logarithmic source conditions after taking the infimum over all α. This is made precise in the following result which relies on a comparison of the rates for general regularization methods and bounds on the spectral cut-off rates, which are known to be optimal in many situations (e.g. Mair & Ruymgaart [30]).

THEOREM 6. *Under the assumptions of Theorem 3, Part 1 with $\tau = 0$ define the increasing functions $\gamma_1(\alpha) := -\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}R(\beta)$ and $\gamma_2(\alpha) := \frac{1}{\alpha^2}\int_0^\alpha S(\beta)\,\mathrm{d}\beta$ and assume that*

$$\Lambda\left(\overline{\gamma_2}(\gamma_1(\alpha))\right) \lesssim C\Lambda(\alpha), \qquad \alpha \searrow 0, \tag{3.15}$$

with the inverse function $\overline{\gamma_2}$ of γ_2 and a constant $C > 0$. Then

$$\inf_{\alpha>0} \mathbf{E}\,\|\hat f_{\alpha,\sigma} - f\|^2 \lesssim \inf_{\alpha>0}\left(C\gamma_\Lambda \Lambda(\alpha)\overline{w} + (C_3^2 + C_2^2)\sigma^2\gamma_1(\alpha)\right), \qquad \sigma \searrow 0,$$

i.e. if we choose the optimal value of α for every noise level σ, all spectral regularization methods achieve the same rate of convergence of the MISE as spectral cut-off.

Assumption (3.15) is satisfied if $\Lambda(t) = (-\ln t)^{-p}$ and $\gamma_1(\alpha) \le \gamma_2(\alpha^2)$ since

$$\Lambda(\overline{\gamma_2}(\gamma_1(\alpha))) \le \Lambda(\alpha^2) = (-2\ln\alpha)^{-p} = 2^{-p}\Lambda(\alpha).$$

4. Noise models satisfying Assumption 1. In this section we show that several commonly used noise models fit into the general framework described in Assumption 1. We start with an (infinite dimensional) white noise model, and then continue with several models based on finitely many observations.

4.1. White noise. A frequently used model is to assume that ε in (2.11) is a white noise process in $\mathbb{H}_2$ (see e.g. Donoho [12, 11], Mathé & Pereverzev [31]). Moreover, we assume that K^*K is a trace-class operator, i.e. it is compact and the eigenvalues ρ_j of K^*K satisfy $\mathrm{tr}(K^*K) := \sum_{j=0}^\infty \rho_j < \infty$. Then $\mathbf{Cov}_{K^*\varepsilon} = K^*K$, so

$$\mathbf{E}\,\|K^*\varepsilon\|^2 = \mathrm{tr}(\mathbf{Cov}_{K^*\varepsilon}) = \mathrm{tr}(K^*K) < \infty.$$

Therefore, $K^*\varepsilon$ can be identified with a Hilbert-space valued random variable. Using the notation introduced in Remark 1 and defining $e_j : \mathbb{N} \to \mathbb{R}$ by $e_j(k) := \delta_{jk}$, $u_j = U^*e_j \in \mathbb{H}_1$ is a unit-length eigenvector of K^*K to the eigenvalue ρ_j, and

$$\mathbf{Var}\,(UK^*\varepsilon(j)) = \mathbf{Var}\,\langle UK^*\varepsilon, e_j\rangle = \mathbf{Var}\,\langle \varepsilon, Ku_j\rangle = \|Ku_j\|^2 = \rho_j,$$

for $j = 0, 1, 2, \ldots$. Therefore, (2.14) is satisfied with equality.

4.2. Quasi-deconvolution, errors in variable, non-compact operators. Suppose we want to estimate the density f of a random variable Z with values in $\mathbb{R}^d$, but we can only observe a random variable $X = Z + W$ perturbed by a random variable W. Hence, our data are

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X = Z + W. \tag{4.1}$$

The density g of X is given by

$$g = \int_{\mathbb{R}^d} h(\cdot - z|z)f(z)\,\mathrm{d}z =: Kf, \tag{4.2}$$

where $h(\cdot|z)$ is the conditional density of W given $Z = z$. If Z and W are stochastically independent, K is a convolution operator. Recovering of f is known as *deconvolution problem* and has been studied extensively (e.g. Stefanski & Carroll [38], Fan [17] and Diggle & Hall [10]). Dependent Z and W in (4.1) occur in many scientific applications, e.g. brightness determination of extragalactic star clusters in astrophysics, where the variance σ^2 of the noise W increases monotonically with decreasing brightness of the object Z. Here, a reasonable model is described by $h(y|z) = (2\pi\sigma^2(z))^{-1/2}\exp(-y^2/\sigma^2(z))$ (see Bissantz [4]).

We assume that $f \in L^2(\mathbb{R}^d)$ and that K is a bounded, injective operator in $L^2(\mathbb{R}^d)$. As opposed to the previous section, in general, K is not compact here. Obviously, an unbiased estimator of $q := K^*g$ is given by

$$\hat{q}_n(y) := \frac{1}{n}\sum_{j=1}^{n} h(X_j - y|y). \tag{4.3}$$

To fit this into our general framework, we show that $\hat{q}_n = q + K^*\tilde{\varepsilon}$ for a Hilbert-space process $\tilde{\varepsilon} : L^2(\mathbb{R}^d) \to L^2(\Omega, \mathcal{P}, P)$ defined by

$$\langle \tilde{\varepsilon}, \varphi \rangle := \frac{1}{n}\sum_{j=1}^{n} \varphi(X_j) - \langle g, \varphi \rangle . \tag{4.4}$$

In fact, for $\psi \in L^2(\mathbb{R}^d)$,

$$\langle K^*\tilde{\varepsilon}, \psi \rangle = \langle \tilde{\varepsilon}, K\psi \rangle = \frac{1}{n}\sum_{j=1}^{n} \int_{\mathbb{R}^d} h(X_j - z|z)\psi(z)\,\mathrm{d}z - \langle K^*g, \psi \rangle = \langle \hat{q}_n - q, \psi \rangle .$$

The next result states that Assumption 1 is satisfied:

Proposition 7. *Assume that the operator K defined by (4.2) is injective and satisfies $\|K\|_{2,2} < \infty$ and $\|K\|_{2,\infty} < \infty$, where $\|K\|_{r,s}$ is defined as the operator norm of $K : L^r(\mathbb{R}^d) \to L^s(\mathbb{R}^d)$. Moreover, let $\hat{q}_n$ and $\tilde{\varepsilon}$ be defined by (4.3) and (4.4), and let*

$$\sigma := \frac{1}{\sqrt{n}}\left(\|g\|_{L^\infty} + \|g\|_{L^2}^2\right)^{1/2} \qquad \text{and} \qquad \varepsilon := \tilde{\varepsilon}/\sigma. \tag{4.5}$$

Then ε satisfies Assumption 1, and $\hat{q}_n = q + \sigma K^*\varepsilon$.

Proof. We have to show that (2.12)–(2.14) hold true. Since the X_j are assumed to be independent, it suffices to consider the case $n = 1$. The first part of (2.12), i.e. $\langle \varepsilon, \varphi \rangle = 0$ for $\varphi \in L^2(\mathbb{R}^d)$ follows from $E\varphi(X) = \int \varphi g\,\mathrm{d}x$. Since

$$\mathbf{Cov}(\langle \tilde{\varepsilon}, \varphi_1 \rangle, \langle \tilde{\varepsilon}, \varphi_2 \rangle) = \int_{\mathbb{R}^d} \varphi_1\varphi_2 g\,\mathrm{d}x - \langle g, \varphi_1 \rangle \langle g, \varphi_2 \rangle \quad \text{for all } \varphi_1, \varphi_2 \in \mathbb{H}_2,$$

the covariance operator of $\tilde{\varepsilon}$ is given by $\mathbf{Cov}_{\tilde{\varepsilon}} = M_g - g \otimes g$, where M_g means multiplication by g, and $g \otimes g : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is the rank-1 operator defined by $(g \otimes g)\varphi := g\langle \varphi, g \rangle$. Now $\|\mathbf{Cov}_{\varepsilon}\| \leq 1$ follows from the estimate $\|\mathbf{Cov}_{\tilde{\varepsilon}}\| \leq \|g\|_{L^\infty} + \|g\|_{L^2}^2$, which completes the proof of (2.12).

To show (2.13), i.e. $\mathbf{E}\|\hat{q}_n - q\|^2 < \infty$, note that

$$\mathbf{Cov}_{\hat{q}_1} = K^*\mathbf{Cov}_{\tilde{\varepsilon}}K = K^*M_gK - (K^*g) \otimes (K^*g).$$

We have to show that this is a trace class operator. Obviously $(K^*g) \otimes (K^*g)$ is trace class as a rank-1 operator. It is not obvious, however, that K^*M_gK is trace class since neither K nor M_g are even compact in general. To show this, we rewrite the kernel of K as $k(x, z) := h(x - z|z)$ and note that $\operatorname{ess\,sup} \|k(x, \cdot)\|_{L^2} = \|K\|_{2,\infty} < \infty$. Since $g \geq 0$, the operator K^*M_gK is self-adjoint and positive semi-definite. Let $\{\varphi_j : j \in \mathbb{N}\}$ be a complete orthonormal system in the separable Hilbert space $L^2(\mathbb{R}^d)$. The B. Levi Theorem yields

$$\begin{aligned}
&\sum_{j\in\mathbb{N}} \langle \varphi_j, K^*M_gK\varphi_j \rangle = \sum_{j\in\mathbb{N}} \int g(x)|(K\varphi_j)(x)|^2 dx \\
&= \sum_{j\in\mathbb{N}} \int g(x)|\langle k(x, \cdot), \varphi_j \rangle|^2 dx \leq \|g\|_{L^1} \operatorname*{ess\,sup}_{x\in\mathbb{X}_2} \|k(x, \cdot)\|_{L^2}^2 < \infty,
\end{aligned}$$

which implies that K^*M_gK is trace class with $\operatorname{tr}(K^*M_gK) \leq \|K\|_{2,\infty}^2$. Finally, (2.14) follows from Lemma 2. ■

If K is a convolution operator with convolution kernel $w(x-z)$, then the canonical choice of the unitary operator U in the Halmos decomposition is the Fourier transform

$$(U\varphi)(\xi) = (\mathcal{F}\varphi)(\xi) = \int_{\mathbb{R}^d} \varphi(x)e^{-2\pi i\xi\cdot x}dx, \tag{4.6}$$

and the multiplier function is then $\rho = |\mathcal{F}w|^2$. In this case the condition (2.14) in Assumption 1 can be verified explicitly, see Mair & Ruymgaart [30].

4.3. Inverse regression. We now review another commonly used noise model (see Wabha [41], O'Sullivan [37], Nychka & Cox [36], Bissantz et *al.* [5]) and show how it is related to the model (2.11). Suppose that $\mathbb{H}_i = L^2(\mu_i)$ are L^2-spaces with respect to measure spaces $(\mathbb{X}_i, \mathcal{X}_i, \mu_i)$, $i = 1, 2$, $\mathbb{H}_1$ is separable, and that $K : L^2(\mu_1) \to L^2(\mu_2)$ is an integral operator

$$(Kf)(x) := \int_{\mathbb{X}_1} k(x,y)f(y)\,\mathrm{d}\mu_1(y), \qquad x \in \mathbb{X}_2, \tag{4.7}$$

with kernel k. Recall that K^*K is trace class if and only if K is Hilbert-Schmidt and that K is a Hilbert-Schmidt operator if and only if $k \in L^2(\mu_2 \times \mu_1)$ (see Taylor [39]). The latter condition is easy to verify in most applications.

We will assume in the following that the measure space $\mathbb{H}_2$ is finite. Then we can arrange that $\mu_2(\mathbb{X}_2) = 1$. We consider the regression model

$$Y_i = (Kf)(X_i) + \varepsilon_i, \qquad f \in \mathbb{H}_1, \quad i = 1, \ldots, n, \tag{4.8}$$

where we assume for simplicity that the random variables $X_i \in \mathbb{X}_2$ have uniform distribution on $\mathbb{X}_2$ (see also Remark 9). Moreover, we assume that $(Y_i, X_i) \sim (Y, X)$, $i = 1, \ldots, n$ are i.i.d. random variables with values in $\mathbb{R} \times \mathbb{X}_2$ such that

$$\mathbf{E}[Y|X] = (Kf)(X), \tag{4.9}$$

and hence $\mathbf{E}[\varepsilon|X] = 0$ for $\varepsilon := Y - (Kf)(X)$. Finally we assume that that $v(X) := \sqrt{\mathbf{E}[\varepsilon^2|X]}$ satisfies

$$0 < C_{v,l} \leq v(X) \leq C_{v,u} < \infty \qquad \text{a.s.}, \tag{4.10}$$

for some constants $C_{v,l}, C_{v,u} > 0$. A straightforward computation shows that

$$\hat{q}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i k(X_i, \cdot). \tag{4.11}$$

is an unbiased estimator of the vector $q := K^*Kf$. To fit the inverse regression model with random design in our general framework, we introduce the Hilbert-space (noise) process $\tilde{\varepsilon} : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P)$ by

$$\langle \tilde{\varepsilon}, \varphi \rangle := \frac{1}{n}\sum_{j=1}^{n} Y_j\varphi(X_j) - \langle g, \varphi \rangle, \qquad \varphi \in \mathbb{H}_2, \tag{4.12}$$

and show that

$$\langle K^*\tilde{\varepsilon}, \psi \rangle = \langle \tilde{\varepsilon}, K\psi \rangle = \frac{1}{n}\sum_{j=1}^{n} Y_j \int_{\mathbb{X}_j} k(X_j, y)\psi(y)\,\mathrm{d}\mu_1(y) - \langle K^*g, \psi \rangle = \langle \hat{q}_n - q, \psi \rangle$$

for all $\psi \in \mathbb{H}_1$, i.e. $\hat{q}_n = q + K^*\tilde{\varepsilon}$.

Proposition 8. *Assume the inverse regression model (4.7)–(4.10), and let $\hat{q}_n$ and $\tilde{\varepsilon}$ be defined by (4.11) and (4.12). Moreover, let $K : L^2(\mu_1) \to L^2(\mu_2)$ be Hilbert-Schmidt, and $\mu_2 - \operatorname{ess\,sup} \|k(x,\cdot)\|_{L^2(\mu_1)} < \infty$. Define*

$$\sigma := \sqrt{\frac{C_1}{n}} \qquad \text{and} \qquad \varepsilon := \tilde{\varepsilon}/\sigma,$$

with $C_1 := C_{v,u} + \|g\|^2_{L^\infty(\mu_2)} + \|g\|^2_{L^2(\mu_2)}$. Then ε satisfies Assumption 1 for the unitary transform U defined in Remark 1, and $\hat{q}_n = q + \sigma K^* \varepsilon$. Moreover,

$$\frac{C_{v,l}}{n}\rho(j) \leq \mathbf{Var}\left((U\hat{q}_n)(j)\right), \qquad j = 0,1,2,\ldots. \tag{4.13}$$

Proof. It suffices to prove this for $n = 1$. Since X is uniformly distributed and (4.9) holds true, we have

$$\mathbf{E}\left(Y\varphi(X)\right) = \mathbf{E}\left(\mathbf{E}\left[\varepsilon|X\right]\varphi(X)\right) + \mathbf{E}\left(g(X)\varphi(X)\right) = \int g\varphi \, \mathrm{d}\mu_2 = \langle g, \varphi\rangle$$

for all $\varphi \in \mathbb{H}_2$ and hence the first part of eq. (2.12) holds true. Using once more the same properties of X and Y we find that

$$\begin{aligned}
&\mathbf{Cov}(\langle\tilde{\varepsilon},\varphi_1\rangle, \langle\tilde{\varepsilon},\varphi_2\rangle) = \mathbf{E}\left\{Y^2\varphi_1(X)\varphi_2(X)\right\} - \langle g,\varphi_1\rangle\langle g,\varphi_2\rangle \\
&= \mathbf{E}\left\{(\varepsilon^2 + 2\varepsilon g(X) + g(X)^2)\varphi_1(X)\varphi_2(X)\right\} - \langle g,\varphi_1\rangle\langle g,\varphi_2\rangle \\
&= \int \varphi_1\left(v^2 + g^2\right)\varphi_2 \, \mathrm{d}\mu_2 - \langle g,\varphi_1\rangle\langle g,\varphi_2\rangle
\end{aligned}$$

for all $\varphi_1, \varphi_2 \in \mathbb{H}_2$. Hence, $\mathbf{Cov}_{\tilde{\varepsilon}} = M_{v^2+g^2} - g\otimes g$ where $M_{v^2+g^2}\varphi := (v^2+g^2)\cdot\varphi$ and $(g\otimes g)\varphi := \langle g,\varphi\rangle g$. This implies $\|\mathbf{Cov}_{\tilde{\varepsilon}}\| \leq C_1$ and finishes the proof of (2.12). Using the notation of Remark 1, condition (2.13) can be seen as follows:

$$\mathbf{E}\,\|\hat{q}_1 - q\|^2 = \operatorname{tr}\left(\mathbf{Cov}_{\hat{q}_1 - q}\right) = \sum_{j=0}^{\infty}\langle Ku_j, \mathbf{Cov}_{\tilde{\varepsilon}}Ku_j\rangle \leq C_1\sum_{j=0}^{\infty}\|Ku_j\|^2 = C_1\operatorname{tr}(K^*K) < \infty.$$

Since

$$\mathbf{Var}\left(U\hat{q}_1\right)(j) = \langle u_j, \mathbf{Cov}_{\hat{q}_1}u_j\rangle = \langle Ku_j, \mathbf{Cov}_{\tilde{\varepsilon}}Ku_j\rangle \leq C_1\|Ku_j\|^2 = C_1\rho_j,$$

we obtain the bound (2.14). The lower bound in (4.13) holds true since the operator $M_g^2 - g\otimes g$ is positive definite as covariance operator of $\tilde{\varepsilon}$ for the case $\varepsilon \equiv 0$. ■

As opposed to [30] we do not need the assumption that the singular vectors $u_j \in \mathbb{H}_1$ and $v_j \in \mathbb{H}_2$ in the singular value decomposition $Kf = \sum_{j=0}^{\infty}\sqrt{\rho_j}\langle f, u_j\rangle_{\mathbb{H}_1} v_j$ be uniformly bounded sequences in $L^\infty(\mu_1)$ and $L^\infty(\mu_2)$, respectively. We only require that $\mu_2 - \operatorname{ess\,sup}\|k(x,\cdot)\|_{L^2(\mu_1)} < \infty$. This condition is often less restrictive and easier to verify.

Remark 9. *Generalizations:*

1. *We can replace $L^2(\mu_1)$ by an arbitrary Hilbert space $\mathbb{H}_1$ (e.g. a Sobolev space) by replacing $k(x,\cdot)$ by $\tilde{k}(x) := \sum_{j=0}^{\infty}\sqrt{\rho_j}v(x)u_j$, $x \in \mathbb{X}_2$. Then (4.7) and (4.11) read $(Kf)(x) = \langle\tilde{k}(x), f\rangle_{\mathbb{H}_1}$ and $\hat{q}_n = \frac{1}{n}\sum_{i=1}^{n}Y_i\tilde{k}(X_i)$, respectively. Proposition 8 remains valid with literally the same proof if $L^2(\mathbb{X}_1)$ is replaced by $\mathbb{H}_1$.*
2. *(deterministic and nonuniform design). The noise model (2.11) also allows to treat models of the form (4.8) where the measurement points are either nonuniformly distributed on $\mathbb{X}_2$ or $x_i = x_i^{(n)}$ are deterministic quantities (see, for instance, Nychka & Cox [36], O'Sullivan [37]). For conditions on the design density see Munk [34].*

5. Applications. In this section we discuss how Assumption 2 of our main result (Theorem 5) can be verified for some specific operators A of practical interest.

A remarkable number of interesting inverse problem can be expressed in the form

$$K^*K = \Theta(-\Delta) \tag{5.1}$$

in terms of the Laplace operator Δ on some compact, smooth d-dimensional Riemannian manifold M with a (possibly empty) boundary ∂M. Our first three examples are of this form. Here $\Theta : [0, \infty) \to (0, \infty)$ is a function satisfying $\lim_{\lambda\to\infty} \Theta(\lambda) = 0$. Under the given assumptions the Laplace operator $-\Delta$ defined on $D(-\Delta) := H_0^1(M) \cap H^2(M) \subset L^2(M)$ (i.e. with Dirichlet condition on ∂M) is a positive, self-adjoint operator, which has a complete orthonormal system of eigenvectors u_i in $L^2(M)$ with corresponding eigenvalues λ_i (see e.g. Taylor [40, Chap. 8.2]). Hence the operator on the right hand side of (5.1) defined in (2.2) can be written as $\Theta(-\Delta)f = \sum_i \Theta(\lambda_i) \langle f, u_i\rangle u_i$ for $f \in L^2(M)$. Due to a famous result of Weyl (see Taylor [40, Ch.8, Thm 3.1. and Cor.3.5]), the distribution of the eigenvalues

$$N(\lambda) := \#\{\lambda_i : \lambda_i \le \lambda\}, \qquad \lambda \ge 0$$

has the asymptotic behavior

$$N(\lambda) \sim c_M \lambda^{d/2}, \qquad c_M := \frac{\operatorname{vol} M}{\Gamma\left(\frac{d}{2}+1\right)(4\pi)^{d/2}} \tag{5.2}$$

as $\lambda \to \infty$, where $\operatorname{vol} M = \int_M 1\,dx$ denotes the volume of M. Under the given assumptions the operator A is compact as operator norm limit of the finite rank operators $\sum_{i=1}^k \Theta(\lambda_i) \langle u_i, \cdot\rangle u_i$ as $k \to \infty$. Assume that $\Theta(\lambda)$ is monotonically decreasing for $\lambda \ge \lambda_0$ and that $\Theta(\lambda) > \alpha_0 := \Theta(\lambda_0)$ for $\lambda < \lambda_0$. As $\lim_{\lambda\to\infty} \Theta(\lambda) = 0$, the inverse function $\overline{\Theta} : (0, \alpha_0] \to [\lambda_0, \infty)$ satisfies $\lim_{\alpha\searrow 0} \overline{\Theta}(\alpha) = \infty$. If the spectral decomposition of A is chosen as in Remark 1, then the function R defined in (3.6) satisfies

$$R(\alpha) = \#\{\lambda_i : \Theta(\lambda_i) \ge \alpha\} = N\left(\overline{\Theta}(\alpha)\right) \sim c_M \left(\overline{\Theta}(\alpha)\right)^{d/2}, \qquad \alpha \searrow 0. \tag{5.3}$$

5.1. Backwards heat equation. We consider the inverse problem to reconstruct the temperature at time $t = 0$ on M from measurements of the temperature at time $t = T$. The forward problem is described by the parabolic equation

$$\begin{array}{ll} \partial_t u(x,t) = \Delta u(x,t), & x \in M, t \in (0,T) \\ u(x,t) = 0, & x \in \partial M, t \in (0,T] \\ u(x,0) = f(x), & x \in M, \end{array} \tag{5.4}$$

with an initial temperature $f \in L^2(M)$ and the final temperature in $g(x) := u(x,T)$, $x \in M$. We have $g = \exp(-T\Delta)f$, i.e. $K = \exp(-T\Delta) \in L(L^2(M))$ and $K^*K = \exp(-2T\Delta)$. Hence,

$$\Theta(\lambda) = \exp(-2T\lambda)$$

in (5.1). By virtue of (5.3) the condition $R \sim S$ is satisfied for

$$S(\alpha) := c_M \left(-\frac{1}{2T} \ln \alpha\right)^{d/2}.$$

It is easy to check that this function satisfies the conditions (3.8). In particular

$$\frac{S''(\alpha)}{-S'(\alpha)} = \frac{1}{\alpha}\left(1 - \frac{d-2}{2}\frac{1}{\ln \alpha}\right) \tag{5.5}$$

so (3.8d) holds with any $\gamma_{\mathrm{S}} \in (1,2)$ for sufficiently small $\bar{\alpha}$ if $d \ge 3$ and with $\gamma_{\mathrm{S}} = 1$ for all $\bar{\alpha} < 1$ for $d \le 2$.

If M is a compact Riemannian manifold without boundary, then the smoothness class $F_{\Lambda,1}$ for a logarithmic source condition (2.7) is the unit ball in the Sobolev space $H^{2p}(M)$ with respect to some equivalent norm (see Hohage [23]). Similar results hold true if M has a boundary with a Dirichlet or Neumann condition. In this case we additionally need to impose boundary conditions. Hence, if the initial temperature is bounded in some Sobolev norm, $\|f\|_{H^s} \le C$, $s = 2p > 0$, if the regularization parameter is chosen such that $\alpha \asymp \sigma$, and if $\tau = O(\sigma^\mu)$ with $\mu > \frac{1}{2}$ as $\sigma \searrow 0$, then it follows from Theorem 3 after some elementary computations that the MISE decays like

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|_{L^2}^2 = \mathrm{O}\left((-\ln\sigma)^{-s}\right), \qquad \sigma \searrow 0$$

for all regularization methods satisfying (2.9).

5.2. Satellite gradiometry. In satellite gradiometry measurements of the gravitational force of the earth at a distance a from the center are used to reconstruct the gravitational potential u at the surface of the earth (see Hohage [23], Bauer & Pereverzev [3] and references therein). Let the earth be described by $E := \{x : |x| < 1\}$, and let $M := \partial E$ denote the surface of the earth. Then u satisfies the Laplace equation

$$\Delta u(x) = 0, \qquad x \in \mathbb{R}^3 \setminus \overline{E}$$

and decays like $|u(x)| = O\left(|x|^{-1}\right)$ as $|x| \to \infty$. The available data consist of noisy measurements of the rate of change of the gravitational force $-\nabla u$ in radial direction $r = |x|$, i.e.

$$g(x) := \frac{\partial^2 u}{\partial r^2}(x), \qquad \text{for } |x| = a.$$

A discussion of the measurement errors shows that they are mainly of random nature (see [3]). The problem is to determine the potential $f = u\big|_M$ at the surface M of the earth. Introducing the operator $K : L^2(M) \to L^2(aM)$ mapping the solution f to the data g, we can write K^*K in the form (5.1) with $\Theta(\lambda) = \Phi(\Lambda(\lambda))$ and

$$\Phi(t) := c\left(\frac{1}{2} + t\right)^2\left(\frac{3}{2} + t\right)^2 a^{-2t}, \qquad \Lambda(\lambda) := \sqrt{\lambda + \frac{1}{2}}$$

(see Hohage [23]). It is easy to show that $\Phi(t)$ is decreasing for sufficiently large t and that $\Lambda(\lambda)$ is monotonic increasing for all $\lambda > 0$. Obviously, $\overline{\Theta}(\alpha) = \overline{\Lambda}\left(\overline{\Phi}(\alpha)\right) = \overline{\Phi}(\alpha)^2 - \frac{1}{2}$. The function $\overline{\Phi}(\alpha)$ cannot be computed explicitly, but we can estimate its asymptotic behavior as $\alpha \searrow 0$. Writing $t = \overline{\Phi}(\alpha)$ for α sufficiently small and $p(t) := c\left(\frac{1}{2} + t\right)^2\left(\frac{3}{2} + t\right)^2$ we obtain

$$\overline{\Phi}(\alpha) = -\log_a \alpha \frac{t}{-\log_a \alpha} = -\log_a \alpha \frac{t}{-\log_a\left(p(t)a^{-2t}\right)} = -\log_a \alpha \left(\frac{t}{-\log_a(p(t)) + 2t}\right) \sim -\frac{\ln\alpha}{2\ln a},$$

as $\alpha \searrow 0$. Therefore, using (5.3), we get

$$R(\alpha) = c_M \overline{\Theta}(\alpha) \sim \left(-\frac{\ln\alpha}{2\ln a}\right)^2, \qquad \alpha \searrow 0.$$

The function $S(\alpha) := \left(-\frac{\ln\alpha}{2\ln a}\right)^2$ satisfies the conditions (3.8) (see (5.5)). Moreover, the smoothness classes $F_{\Lambda,1}$ for logarithmic source conditions (2.7) are unit balls in the Sobolev spaces $H^p(M)$ w.r.t. equivalent norms (see Hohage [23]). Since the gravitational potential satisfies the Poisson equation $\Delta u = -\phi$ in $\mathbb{R}^3$ and since the mass density ϕ of the earth E belongs to $L^2(E)$, it follows from elliptic regularity results that $u \in H^2(E)$, so $f = u|_M \in H^{3/2}(M)$ in the sense of the trace operator (see e.g. Taylor [39]). Therefore,

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|_{L^2}^2 = \mathrm{O}\left((-\ln\sigma)^{-3}\right), \qquad \sigma \searrow 0,$$

if $\tau = O(\sigma)$ and if we choose $\alpha \asymp \sigma$.

5.3. Operators in Hilbert scales. In the following we show that our assumptions are satisfied for operators acting in Hilbert scales (see Mair & Ruymgaart [30], Mathé & Pereverzev [31]). Hence, spectral regularization methods yield optimal rates of convergence for this class of operators.

Let $L : D(L) \subset H \to H$ be an unbounded, positive, self-adjoint operator defined on a dense domain $D(L) \subset H$, and assume the inverse $L^{-1} : H \to H$ is bounded. Then L generates a scale of Hilbert spaces H_μ, $\mu \in \mathbb{R}$ defined as completion of $\bigcap_{n\in\mathbb{N}} D(L^n)$ under the norm generated by the inner product $\langle f, g\rangle_\mu := \langle L^\mu f, L^\mu g\rangle$. We have $H_\mu \subset H_\lambda$ for $\mu, \lambda \in \mathbb{R}$ with $\mu > \lambda$. A prototype is $L = \sqrt{I - \Delta}$ with the Laplace operator Δ on a closed manifold M, which leads to the usual Sobolev spaces on M.

We assume that K is a-times smoothing ($a > 0$) in (part of) the Hilbert scale (H_μ), i.e. $K : H_{\mu-a} \to H_\mu$ is a bounded operator for all $\mu \in [\underline{\mu}, \overline{\mu}]$ which has a bounded inverse $K^{-1} : H_\mu \to H_{\mu-a}$. This is equivalent to $\frac{1}{C_\mu}\|f\|_{\mu-a} \le \|Kf\|_\mu \le C_\mu \|f\|_{\mu-a}$ for all $f \in H_{\mu-a}$ and some constants $C_\mu \ge 1$. Such conditions are satisfied for many boundary integral operators, multiplication operators, convolution operators and compositions of such operators (see also the discussion after (5.10)). We do not assume here that K is self-adjoint or that K^*K and L commute, i.e. that they can be diagonalized by the same unitary operator U.

Usually the nature of the noise dictates the choice $\mathbb{H}_2 = H_0$, and one is interested in error bounds for the estimator in positive norm, i.e. $\mathbb{H}_1 = H_{\mu-a}$ for $\mu \ge a$. Then the operator equation $Kf = g$ is ill-posed with $K = K_{0\leftarrow\mu-a}$ considered as an operator from $H_{\mu-a}$ to H_0.

To verify Assumptions 2 and 3 with $R \sim S$ in (3.7) replaced by $R \asymp S$ (see Remark 14), we assume that L has a complete orthonormal system of eigenvectors with eigenvalues $0 < \lambda_0 \le \lambda_1 \le \lambda_2 \le \dots$ tending to infinity. Then the embedding operator $J : H_\mu \hookrightarrow H_0$ is compact, and its singular values are given by $\sigma_j(J) = \lambda_j^{-\mu}$. It follows from the decomposition $K_{0\leftarrow\mu-a} = JK_{\mu\leftarrow\mu-a}$ and the Courant mini-max characterization of the singular values $\sigma_j = \sigma_j(K_{0\leftarrow\mu-a})$ (see e.g. Kreß [28]) that

$$\frac{1}{\|K^{-1}\|_{\mu-a\leftarrow\mu}}\lambda_j^{-\mu} \le \sigma_j(K_{0\leftarrow\mu-a}) \le \|K\|_{\mu\leftarrow\mu-a}\lambda_j^{-\mu}, \qquad j = 0, 1, \dots$$

Hence if $N(\lambda) := \#\{\lambda_j : \lambda_j \le \lambda\}$ and $C := \max(\|K\|_{\mu\leftarrow\mu-a}, \|K^{-1}\|_{\mu-a\leftarrow\mu})$, then $R(\alpha) := \{\sigma_j : \sigma_j^2 \ge \alpha\}$ satisfies

$$N\left((\alpha/C^2)^{-1/2\mu}\right) \le R(\alpha) \le N\left((C^2\alpha)^{-1/2\mu}\right).$$

If the counting function has the asymptotic behavior $N(\lambda) \asymp \lambda^d$ for some $d > 0$, then $R(\alpha) \asymp \alpha^{-d/2\mu}$. For the case $L = \sqrt{I - \Delta}$, d is the space dimension (see (5.2)). A straightforward computation shows that $S(\alpha) := \alpha^{-d/2\mu}$ satisfies (3.8) and (3.12) in Assumption 2 and 3 if and only if $d/(2\mu) \in (0, 1)$. Under this condition, it follows from Remark 14 that Theorems 3 and 5 hold true with different constants.

It remains to discuss the Hölder-type source conditions (2.6) in this setting. To do this we assume for simplicity that $\mathbb{H}_1 = \mathbb{H}_2 = H_0$. Let K^* denote the adjoint of K with respect to the inner product in H_0. It is easy to show that $K^* : H_{-\mu} \to H_{-\mu+a}$ is bounded and boundedly invertible for all $\mu \in [\underline{\mu}, \overline{\mu}]$. Let $l \in \mathbb{N}$ such that $[-2al + 1, 2al - 1] \subset [\underline{\mu}, \overline{\mu}]$. Then there exists a constant $\gamma \ge 1$ such that

$$\gamma^{-1}\|L^{2al}f\|_{H_0} \le \|(K^*K)^{-l}f\|_{H_0} \le \gamma\|L^{2al}f\|_{H_0}$$

for all $f \in H_{2al}$. It follows from the Heinz inequality (see *et al.* [14], Heinz [22]) that

$$\gamma^{-\sigma}\|L^{2a\sigma l}f\|_{H_0} \le \|(K^*K)^{-\sigma l}f\|_{H_0} \le \gamma^{\sigma}\|L^{2a\sigma l}f\|_{H_0}$$

for all $\sigma \in [0, 1]$ and $f \in H_{2a\sigma l}$. Therefore, the source condition $f = (K^*K)^\nu w$, $w \in H_0$ is equivalent to $f \in H_{2a\nu}$. Let $u := 2a\nu$ and $f \in H_u$. Then

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|_{H_0}^2 = \mathrm{O}\left(\sigma^{\frac{2u}{u+a+d/2}}\right), \qquad \sigma \searrow 0,$$

for the choice $\alpha \asymp \sigma^{\frac{2a}{u+a+d/2}}$ if $\tau = \mathrm{O}\left(\sigma^{\frac{u+a}{u+a+d/2}}\right)$ and $\mu_0 \geq u/2a$.

5.4. L^2-Boosting. Boosting algorithms include a large class of iterative procedures which improve stagewise the performance of estimators. They have achieved significant interest in the machine learning context and more recently in statistics (see Freund & Shapire [18] or Friedman [19] among many others). One of the main challenges is to provide a proper convergence analysis and proper stopping rules for the iteration depth (see Zhang & Yu [43]). L^2-Boosting has been introduced in the context of regression analysis by Bühlmann & Yu [7] for classification and more general learning problems. We consider the inverse regression problem described in §4.3 if K is an embedding operator and $\mathbb{X}_2$ is a d-dimensional smooth, compact Riemannian manifold (e.g. a smooth compact subset of $\mathbb{R}^d$). Consider a *weak learner* of the form

$$\hat{f}_{0,n} = \frac{1}{n}\sum_{j=1}^{n} Y_i k(y, X_j), \tag{5.6}$$

with a continuous, symmetric kernel $k : \mathbb{X}_2 \times \mathbb{X}_2 \to \mathbb{R}$ such that the integral operator $\tilde{K} : L^2(\mathbb{X}_2) \to L^2(\mathbb{X}_2)$ with kernel k is compact and strictly positive definite with eigenvalues $\kappa_0 \geq \kappa_1 \geq \ldots$ and satisfies

$$\operatorname{ess\,sup}_{x\in\mathbb{X}_2} k(x,x) < \infty \qquad \text{and} \qquad \#\{\kappa_j \geq \alpha\} \asymp \alpha^{-d/(2\mu_0)} \text{ as } \alpha \to 0, \tag{5.7}$$

for some $\mu_0 > 0$. Further, let H_μ, $\mu \in \mathbb{R}$ be the Hilbert scale generated by the operator $L := \tilde{K}^{-1/(2\mu_0)}$ as described in §5.3. If we set $\mathbb{H}_1 := H_{\mu_0}$ and $\mathbb{H}_2 = H_0 = L^2(\mathbb{X}_2)$, then $\mathbb{H}_1 \subset \mathbb{H}_2$, and the adjoint of the embedding operator $K : \mathbb{H}_1 \hookrightarrow \mathbb{H}_2$ is given by $K^*\varphi = \tilde{K}\varphi$ since $\langle \varphi, \tilde{K}\psi\rangle_{\mathbb{H}_1} = \langle L^{\mu_0}\varphi, L^{\mu_0}\tilde{K}\psi\rangle_{L^2} = \langle \varphi, \psi\rangle_{L^2}$ for all $\psi \in L^2(\mathbb{X}_2)$ and all $\varphi \in \mathbb{H}_1$. By a similar reasoning one can show that $\mathbb{H}_1$ is a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k(\cdot, x)$. A typical example of a weak learner is a spline smoother which leads to Sobolev spaces H_μ (see [7]).

Note that the weak learner (5.6) can shortly be written as $\hat{f}_{0,n} = K^*Y$. Boosting this learner results in a recursive iteration

$$\hat{f}_{j+1,n} = \hat{f}_{j,n} - \beta K^*(Y - K\hat{f}_{j,n}), \qquad j = 0, 1, 2, \ldots, \tag{5.8}$$

which is in fact Landweber iteration (see §2.2). Hence, Theorem 3 gives the following bound.

Corollary 10. *Assume that k satisfies (5.7) with $\mu_0 > \frac{d}{2}$, let $g \in H_\mu$ with $\mu > 0$ and $\beta \in (0, \|KK^*\|^2]$. Then the MISE is bounded by*

$$\mathbf{E}\,\|\hat{f}_{j,n} - g\|^2_{L^2(\mathbb{X}_2)} \leq C\left((j+1)^{-\mu/\mu_0} + n^{-1}(j+1)^{d/(2\mu_0)}\right). \tag{5.9}$$

For the optimal stopping index $j_*(n) \asymp n^{2\mu_0/(2\mu+d)}$ we obtain the rate $\mathbf{E}\,\|\hat{f}_{j_*(n),n} - f\|^2_{L^2(\mathbb{X}_2)} \leq Cn^{-2\mu/(2\mu+d)}$, which is the well-known minimax rate in the case of Sobolev spaces.

Proof. It follows easily from the definitions that $g \in H_\mu$ is equivalent to $g \in F_{\tilde{\Lambda},\overline{w}}$ with $\tilde{\Lambda}(t) = t^{\mu/2\mu_0}$ for some $\overline{w} > 0$. Since Landweber iteration has infinite qualification (see [14]), $\tilde{\Lambda}$ satisfies (2.10). Moreover, as the singular values of K are $\sigma_j(K) = \sqrt{\kappa_j}$, (5.7) implies that $R(\alpha) = \#\{\sigma_j(K)^2 \geq \alpha\} \asymp S(\alpha)$ with $S(\alpha) := \alpha^{-d/2\mu_0}$, and S satisfies (3.8) in Assumption 2 for $\mu_0 > \frac{d}{2}$. To verify the assumptions of Prop. 8, we note that $\operatorname{tr}(K^*K) = -\int_0^\infty \alpha \,\mathrm{d}R(\alpha) < \infty$ for $\mu_0 > \frac{d}{2}$ (i.e. K is Hilbert-Schmidt) and that $\operatorname{ess\,sup}_{x\in\mathbb{X}_2}\|k(x,\cdot)\|_{\mathbb{H}_1} = \operatorname{ess\,sup}_{x\in\mathbb{X}_2}\sqrt{\langle k(x,\cdot), k(x,\cdot)\rangle_{\mathbb{H}_1}} = \operatorname{ess\,sup}_{x\in\mathbb{X}_2}\sqrt{k(x,x)} < \infty$. Therefore, Prop. 8 and Remark 9.1 imply that Assumption 1 is satisfied with $\sigma \asymp n^{-1/2}$. Hence (5.9) follows from (3.10) in Theorem 3 with $\alpha = (j+1)^{-1}$ and Remark 14. ∎

Corollary 10 immediately applies to all other regularization methods covered by Theorem 3. In particular, ν-methods require only the square root of the number of Landweber iterations to achieve the optimal rate, but they seem to be unknown in statistics and machine learning.

Often a discretized sample variant of the iteration (5.8) is considered. Convergence of this algorithm has been analyzed by Yao *et al.*[42], but without optimal rates. It is still an open problem whether this discretized version achieves the minimax rates of Corollary 10 in the general context of RKHS as it has been shown in [7] for the particular case of spline learning.

5.5. Errors in variable problems. We now further discuss the errors in variable problems introduced in §4.2. Our aim is to establish rates of convergence of estimators of the density f of $Z \in \mathbb{R}^d$ as the sample size n tends to infinity. Therefore, with a slight abuse of notation, we will write $\hat{f}_{\alpha,n} = \hat{f}_{\alpha,\sigma(n,g)}$ in this context. It follows from the definition (4.5) of σ and the boundedness of $\|\Lambda(K^*K)\|_{2,2}$ that

$$\sup_{f \in F_{\Lambda,\overline{w}}} \sigma(n, Kf) = \sup_{\|w\|=\overline{w}} \sigma(n, K\Lambda(K^*K)w) \leq \frac{\overline{w}}{\sqrt{n}} \left(\|K\Lambda(K^*K)\|_{2,\infty}^2 + \|K\Lambda(K^*K)\|_{2,2}^2 \right)$$

where the expression in parenthesis is finite under the assumptions of Proposition 7.

We first consider two important special cases

$$h_1(y|z) = w_1(y) := \exp(-\pi\|y\|_2^2), \qquad h_2(y|z) = w_2(y) := c_d \exp(-\|y\|_2), \quad y, z \in \mathbb{R}^d$$

with normalization constant $c_d := \pi^{-d/2}\Gamma(d/2+1)/\Gamma(d+1)$ corresponding to an error variable W independent of Z. Here K is a convolution operator, the canonical unitary transformation U in the spectral decomposition is the Fourier transform $\mathcal{F}$ defined in (4.6), and the multiplier function is $\rho_j = |\mathcal{F}w_j|^2$, i.e. $\rho_1(\xi) = \exp(-2\pi\|\xi\|_2^2)$, and $\rho_2(\xi) = (1 + 4\pi^2\|\xi\|^2)^{-d-1}$. Hence, the corresponding functions R are given by

$$R_1(\alpha) = V_d \left(-\frac{1}{2\pi} \ln \alpha \right)^{d/2}, \qquad R_2(\alpha) = V_d (2\pi)^{-d} \left(\alpha^{-1/(d+1)} - 1 \right)^{d/2}, \qquad 0 < \alpha < 1,$$

where V_d denotes the volume of the unit ball in $\mathbb{R}^d$. Hence, Assumption 2 is satisfied for R_1 with $S = R_1$ (see (5.5)) and for R_2 with $S(\alpha) = V_d(2\pi)^{-d}\alpha^{-d/(2d+2)}$. Since the norm of the Sobolev space $H^s(\mathbb{R}^d)$ is defined by $\|f\|_{H^s(\mathbb{R}^d)} = (\int (1+|\xi|^2)^s |\mathcal{F}f(\xi)|^2 \,\mathrm{d}\xi)^{1/2}$, a simple computation shows that in the first case a logarithmic source condition (2.7) is equivalent to $f \in H^{2p}(\mathbb{R}^d)$, and in the second case a Hölder-type source condition (2.6) is equivalent to $f \in H^{2(d+1)\nu}(\mathbb{R}^d)$. Suppose that $f \in H^s(\mathbb{R}^d)$. Then we find in the first case for the choice $\alpha \asymp n^{-1/2}$ the asymptotic rates

$$\mathbf{E}\,\|\hat{f}_{\alpha,n} - f\|_{L^2}^2 = \mathrm{O}\left((\ln n)^{-s} \right), \qquad n \to \infty,$$

and in the second case the rate

$$\mathbf{E}\,\|\hat{f}_{\alpha,n} - f\|_{L^2}^2 = \mathrm{O}\left(n^{-\frac{s}{s+3d/2+1}} \right), \qquad n \to \infty$$

for the choice $\alpha \asymp n^{-\frac{d+1}{s+3d/2+1}}$. This generalizes results in Mair & Ruymgaart [30] for spectral cut-off to arbitrary regularization methods and to the multivariate setting.

We now consider the case that the random variables Z and W are not stochastically independent. We assume that the conditional density h is of the form

$$h(x-z|z) = w(x-z) + p(x,z) \tag{5.10}$$

where $\underline{c}(1+\|\xi\|_2^2)^{-a} \leq |\mathcal{F}w(\xi)|^2 \leq \overline{c}(1+\|\xi\|_2^2)^{-a}$ for some constants $a, \underline{c}, \overline{c} > 0$, and p is C^∞-smooth and decays exponentially as $\|x\|, \|z\| \to \infty$. Then the convolution operator $\tilde{K}$ with kernel w is bounded and boundedly invertible from $H^{\mu-a}(\mathbb{R}^d)$ to $H^\mu(\mathbb{R}^d)$ for all $\mu \in \mathbb{R}$, and the integral operator P with kernel p is compact from $H^{\mu-a}(\mathbb{R}^d)$ to $H^\mu(\mathbb{R}^d)$ for all $\mu \in \mathbb{R}$. Under our general assumption that $K = \tilde{K} + P$ is injective, it follows from Riesz theory that $K : H^{\mu-a}(\mathbb{R}^d) \to H^\mu(\mathbb{R}^d)$ has a bounded inverse. Hence, it follows from the arguments of the previous paragraph that Hölder source condition (2.6) for K are equivalent to $f \in H^{2a\nu}(\mathbb{R}^d)$. If we additionally assume periodicity of w and p with arbitrary size of the periodicity interval, then it follows from our results on operators in Hilbert scales that also Assumption 1 and 2 are satisfied.

6. Appendix: Proofs and auxiliary results. This section contains the proofs of our main results on the MISE. First, we require some technical lemmas.

LEMMA 11. *If Assumption 1 holds true and the family of functions $\{\Phi_\alpha\}$ satisfies (2.9), then*

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2 \le -\frac{(\sigma C_3)^2}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta) - (\sigma C_2)^2\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta), \tag{6.1a}$$

$$\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - \mathbf{E}\,K\hat{f}_{\alpha,\sigma}\|^2 \le (\sigma C_2)^2 R(\alpha) - \frac{(\sigma C_3)^2}{\alpha^2}\int_0^\alpha \beta^2\,\mathrm{d}R(\beta). \tag{6.1b}$$

(Recall that R is decreasing, i.e. the right hand sides of the inequalities above are non-negative.)

Proof. Recall the bound (3.5) on $\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2$. We split the integral on the right hand side of (3.5) of the variance over the "frequency domain" $\mathbb{S}$ into low frequency components $\{\rho \ge \alpha\}$ and high frequency components $\{\rho < \alpha\}$. The low frequency components are bounded by

$$\int_{\{\rho\ge\alpha\}} \Phi_\alpha(\rho)^2\rho\,\mathrm{d}\Sigma \le C_2^2\int_{\{\rho\ge\alpha\}} \frac{1}{\rho}\,\mathrm{d}\Sigma = -C_2^2\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta),$$

where the latter equality holds by a transformation of the integral on the l.h.s. to an integral with respect to the image measure Σ^ρ, and subsequent reformulation as the Lebesgue-Stieltjes integral given on the r.h.s. of the equation. Similarly, the high frequency components of the variance can be estimated by

$$\int_{\{\rho<\alpha\}} \Phi_\alpha(\rho)^2\rho\,\mathrm{d}\Sigma \le \frac{C_3^2}{\alpha^2}\int_{\{\rho<\alpha\}} \rho\,\mathrm{d}\Sigma = -\frac{C_3^2}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta)$$

using (2.9b). This completes the proof of (6.1a).
In analogy to (3.5) we have $\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - \mathbf{E}\,K\hat{f}_{\alpha,\sigma}\|^2 \le \sigma^2\int_{\mathbb{S}}\Phi_\alpha(\rho)^2\rho^2\,\mathrm{d}\Sigma$, and the right hand side of this inequality can be estimated as above to establish the bound (6.1b). ∎

The next lemma shows that for $R = S$ the high frequency components of the variance are asymptotically bounded by low frequency components and that the relative magnitude of these components is determined by the constant γ_{S} in (3.8d).

LEMMA 12. *Assume that $S \in C^2((0,\bar{\alpha}])$ satisfies (3.8), and define $\kappa := \frac{\gamma_{\mathrm{S}}}{2-\gamma_{\mathrm{S}}}$, i.e. $\frac{2\kappa}{\kappa+1} = \gamma_{\mathrm{S}}$. Then*

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}S(\beta) \le -\kappa\int_\alpha^{\bar{\alpha}} \frac{1}{\beta}\,\mathrm{d}S(\beta) - \frac{\kappa+1}{2}S'(\bar{\alpha}), \qquad \alpha \in (0,\bar{\alpha}]. \tag{6.2}$$

Proof. We rewrite (3.8d) as $(\kappa+1)S''(\alpha) \le 2\kappa\frac{-S'(\alpha)}{\alpha}$. Integrating this inequality from α to $\bar{\alpha}$ yields $(\kappa+1)(S'(\bar{\alpha}) - S'(\alpha)) \le 2\kappa\int_\alpha^{\bar{\alpha}}\frac{-S'(\beta)}{\beta}\,\mathrm{d}\beta$, or equivalently

$$0 \le \alpha S'(\alpha) + \kappa\alpha S'(\alpha) + 2\kappa\alpha\int_\alpha^{\bar{\alpha}}\frac{-S'(\beta)}{\beta}\,\mathrm{d}\beta - \alpha(\kappa+1)S'(\bar{\alpha}). \tag{6.3}$$

It follows that

$$0 \le \int_0^\alpha \beta\,\mathrm{d}S(\beta) - \kappa\alpha^2\int_\alpha^{\bar{\alpha}}\frac{1}{\beta}\,\mathrm{d}S(\beta) - \frac{\alpha^2}{2}(\kappa+1)S'(\bar{\alpha}), \qquad \alpha \in (0,\bar{\alpha}]. \tag{6.4}$$

To verify this we check that the derivative of the right hand side of eq. (6.4) is the right hand side of (6.3) and that the limit of the right hand side of (6.4) as $\alpha \searrow 0$ is nonnegative by assumptions (3.8a) and (3.8b). (6.4) is equivalent to (6.2). ∎

Next we show under an additional assumption that the asymptotic balance between high and low frequency components of the variance also holds true if R is not smooth.

LEMMA 13. *If Assumption 2 holds true, then for $j \in \{1,2\}$*

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta^j \,\mathrm{d}S(\beta) \le \frac{1}{\alpha^2}\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta, \tag{6.5a}$$

$$\left|\frac{1}{\alpha^2}\int_0^\alpha \beta^j \,\mathrm{d}(R-S)(\beta)\right| = \mathrm{o}\left(\frac{1}{\alpha^2}\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta\right), \tag{6.5b}$$

$$-\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}S(\beta) \le \frac{1}{\alpha}S(\alpha), \tag{6.5c}$$

$$\left|\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}(R-S)(\beta)\right| = \mathrm{o}\left(\frac{1}{\alpha}S(\alpha)\right) \tag{6.5d}$$

as $\alpha \searrow 0$. If additionally Assumption 3 is satisfied, then

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta^j \,\mathrm{d}R(\beta) \sim -\frac{1}{\alpha^2}\int_0^\alpha \beta^j \,\mathrm{d}S(\beta), \tag{6.6a}$$

$$-\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}R(\beta) \sim -\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}S(\beta). \tag{6.6b}$$

Proof. Using (3.8c), a partial integration yields

$$-\int_0^\alpha \beta^j \,\mathrm{d}T(\beta) = -\alpha^j T(\alpha) + \int_0^\alpha j\beta^{j-1}T(\beta)\,\mathrm{d}\beta \qquad \text{for } T = S \text{ and } T = R-S. \tag{6.7}$$

Due to assumption (3.7) and (3.8b), the left hand side of (6.7), and hence $\int_0^\alpha j\beta^{j-1}T(\beta)\,\mathrm{d}\beta$ is finite. (6.5a) follows from (6.7) with $T = S$ since $R(\alpha)$, and hence $S(\alpha)$ are positive for small α. By assumption (3.7), there exists for all $\epsilon > 0$ a $\delta = \delta(\epsilon) > 0$ such that

$$|R(\alpha) - S(\alpha)| \le \epsilon S(\alpha) \qquad \text{for } \alpha < \delta. \tag{6.8}$$

Therefore, using (6.7) with $T = S - R$,

$$\left|\int_0^\alpha \beta^j \,\mathrm{d}(S(\beta) - R(\beta))\right| \le \epsilon\alpha^j S(\alpha) + \epsilon\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta$$

for $\alpha < \delta$. As $\alpha^j S(\alpha) = \int_0^\alpha j\beta^{j-1}S(\alpha)\,\mathrm{d}\beta \le \int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta$ due to (3.8a), we obtain (6.5b).

To prove (6.5c) and (6.5d), again partial integration yields for $T = S$ or $T = R - S$

$$-\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}T(\beta) = \frac{1}{\alpha}T(\alpha) - \frac{1}{\bar\alpha}T(\bar\alpha) - \int_\alpha^{\bar\alpha} \frac{1}{\beta^2}T(\beta)\,\mathrm{d}\beta. \tag{6.9}$$

For $T = S$ this yields (6.5c). Let $\epsilon > 0$ and choose $\delta_1 := \delta(\epsilon)$ according to (6.8) and $\delta_2 := \delta_1\epsilon$. Then

$$\left|\int_\alpha^{\bar\alpha} \frac{R(\beta) - S(\beta)}{\beta^2}\,\mathrm{d}\beta\right| \le \epsilon\int_\alpha^{\delta_1} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta + \int_{\delta_1}^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta + \int_{\delta_1}^{\bar\alpha} \frac{R(\beta)}{\beta^2}\,\mathrm{d}\beta$$

for $\alpha \le \delta_2$. Due to the monotonicity of S we have

$$\int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \ge \int_{\delta_2}^{\delta_1} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \ge S(\delta_1)\int_{\delta_2}^{\delta_1} \frac{\mathrm{d}\beta}{\beta^2} = S(\delta_1)\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right) = \frac{1-\epsilon}{\epsilon}\frac{S(\delta_1)}{\delta_1},$$

so

$$\int_{\delta_1}^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \le S(\delta_1)\int_{\delta_1}^{\infty} \frac{\mathrm{d}\beta}{\beta^2} = \frac{S(\delta_1)}{\delta_1} \le \frac{\epsilon}{1-\epsilon}\int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta,$$

$$\int_{\delta_1}^{\bar\alpha} \frac{R(\beta)}{\beta^2}\,\mathrm{d}\beta \le \frac{1}{\delta_1}R(\delta_1) \le (1+\epsilon)\frac{S(\delta_1)}{\delta_1} \le \epsilon\frac{1+\epsilon}{1-\epsilon}\int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta.$$

Since $S(\alpha) > 0$ for all $\alpha \in (0, \bar{\alpha}]$ due to (3.12) we can extend the integrals over $[\delta_2, \delta_1]$ and $[\alpha, \delta_1]$ to $[\alpha, \bar{\alpha}]$ and obtain

$$\left| \int_\alpha^{\bar{\alpha}} \frac{R(\beta) - S(\beta)}{\beta^2} \, \mathrm{d}\beta \right| \leq \epsilon \left(1 + \frac{1}{1-\epsilon} + \frac{1+\epsilon}{1-\epsilon} \right) \int_\alpha^{\bar{\alpha}} \frac{S(\beta)}{\beta^2} \, \mathrm{d}\beta$$

for $\epsilon < 1$ and $\alpha \leq \delta_2$. Since $\int_\alpha^{\bar{\alpha}} \frac{S(\beta)}{\beta^2} \, \mathrm{d}\beta \leq S(\alpha)/\alpha - S(\bar{\alpha})/\bar{\alpha} \lesssim S(\alpha)/\alpha$ due to (6.9) and (3.8a), we obtain (6.5d).

Assume now that Assumption 3 holds true. Written as $-\alpha^j S'(\alpha) \geq C_4 \alpha^{j-1} S(\alpha)$ and integrated from 0 to α, eq. (3.12) yields

$$-\int_0^\alpha \beta^j \, \mathrm{d}S(\beta) \geq C_4 \int_0^\alpha \beta^{j-1} S(\beta) \, \mathrm{d}\beta \qquad \text{for } \alpha \in (0, \bar{\alpha}].$$

Together with (6.5b) this implies (6.6a). Writing (3.12) as $-S'(\alpha)/\alpha \geq C_4(S(\alpha)/\alpha^2)$ and adding $C_4(-S'(\alpha)/\alpha)$ on both sides, we obtain

$$(C_4 + 1) \frac{-S'(\alpha)}{\alpha} \geq C_4 \left(\frac{-S'(\alpha)}{\alpha} + \frac{S(\alpha)}{\alpha^2} \right) = C_4 \frac{d}{d\alpha} \left\{ -\frac{1}{\alpha} S(\alpha) \right\}.$$

Integrating this inequality from α to $\bar{\alpha}$ and multiplying by $(C_4 + 1)^{-1}$ yields

$$-\int_\alpha^{\bar{\alpha}} \frac{1}{\beta} \, \mathrm{d}S(\beta) \geq \frac{C_4}{C_4 + 1} \left(\frac{1}{\alpha} S(\alpha) - \frac{1}{\bar{\alpha}} S(\bar{\alpha}) \right) \gtrsim \frac{C_4}{C_4 + 1} \frac{S(\alpha)}{\alpha}, \qquad \alpha \searrow 0.$$

This together with (6.5d) implies (6.6b). ∎

Remark 14. *Assume*

$$R(\alpha) \asymp S(\alpha), \qquad \alpha \searrow 0, \tag{6.10}$$

i.e. there exist constants $C \geq 1$ and $\bar{\alpha} > 0$ such that $(1/C)R(\alpha) \leq S(\alpha) \leq CR(\alpha)$ for $0 < \alpha \leq \bar{\alpha}$. In this case (6.8) holds true with $\delta = \bar{\alpha}$ and $\epsilon = \max(C - 1, 1 - 1/C)$. Proceeding as in the proof of Lemma 13 and choosing $\delta_1 = \delta_2 = \bar{\alpha}$, we find that (6.5) holds true with $\mathrm{o}(\ldots)$ replaced by $\mathrm{O}(\ldots)$ if S satisfies (3.8). If additionally (3.12) holds true, then

$$-\frac{1}{\alpha^2} \int_0^\alpha \beta \, \mathrm{d}R(\beta) \asymp \frac{1}{\alpha^2} \int_0^\alpha S(\beta) \, \mathrm{d}\beta \asymp -\frac{1}{\alpha^2} \int_0^\alpha \beta \, \mathrm{d}S(\beta),$$
$$-\int_\alpha^{\bar{\alpha}} \frac{1}{\beta} \, \mathrm{d}R(\beta) \asymp \frac{1}{\alpha} S(\alpha) \asymp -\int_\alpha^{\bar{\alpha}} \frac{1}{\beta} \, \mathrm{d}S(\beta).$$

Therefore similar convergence rate results with different constants can be shown if condition (3.7) in Assumption 2 is replaced by (6.10).

Proof of Theorem 3. To prove (3.9), we use the bias-variance decomposition (3.1) and the bound (3.3) of the bias. To bound the variance we start from (6.1a) in Lemma 11. From (6.5a) and (6.5b) we obtain $-\alpha^{-2} \int_0^\alpha \beta \, \mathrm{d}R(\beta) \lesssim \alpha^{-2} \int_0^\alpha S(\beta) \, \mathrm{d}\beta$. For the second term in (6.1a) the partial integration (6.9) with $T = R$ and $\bar{\alpha} > \|K^*K\|$ and (3.7) yield

$$-\int_\alpha^\infty \frac{1}{\beta} \, \mathrm{d}R(\beta) \leq \frac{1}{\alpha} R(\alpha) \sim \frac{1}{\alpha} S(\alpha) \quad \alpha \searrow 0.$$

Using the partial integration (6.7) with $T = S$ and (3.8a) we obtain

$$\frac{1}{\alpha} S(\alpha) = \frac{1}{\alpha^2} \int_0^\alpha \beta \, \mathrm{d}S(\beta) + \frac{1}{\alpha^2} \int_0^\alpha S(\beta) \, \mathrm{d}\beta \leq \frac{1}{\alpha^2} \int_0^\alpha S(\beta) \, \mathrm{d}\beta.$$

This completes the proof of (3.9). The proof of (3.10) also relies on the bias-variance decomposition $\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - g\|^2 = B^2 + V$ where the bias term satisfies

$$\begin{aligned} B &= \|\mathbf{E}\,K\hat{f}_{\alpha,\sigma} - g\| \le \|K\Phi_\alpha(K^*K)K^*g - g\| + \tau\|K\Phi_\alpha(K^*K)K\xi\| \\ &\le \|(\Phi_\alpha(KK^*)KK^* - I)\tilde{\Lambda}(KK^*)w\| + \tau\|(\Phi_\alpha(KK^*)KK^*\| \le \gamma_{\tilde{\Lambda}}\tilde{\Lambda}(\alpha)\overline{w} + \tau C_2. \end{aligned}$$

The bound on the variance term $V = \mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - \mathbf{E}\,K\hat{f}_{\alpha,\sigma}\|^2$ we start from (6.1b) in Lemma 11. By (6.5a) and (6.5b), the first term on the right hand side satisfies $\int_0^\alpha \beta^2\,\mathrm{d}R(\beta) \lesssim \frac{1}{\alpha^2}\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta$, and for the second term we obtain $R(\alpha) \sim S(\alpha) = \alpha^{-2}\int_0^\alpha 2\beta S(\alpha)\,\mathrm{d}\beta \le \alpha^{-2}\int_0^\alpha 2\beta S(\beta)\,\mathrm{d}\beta$ due to (3.8a). This shows that $V \le (\sigma/\alpha)^2(C_2^2 + C_3^2)\int_0^\alpha 2\beta S(\beta)\,\mathrm{d}\beta$ and finishes the proof of (3.10). ■

Proof of Theorem 5. Using (3.11) we can bound the variance of the spectral cut-off estimator as follows:

$$\sigma^{-2}\mathbf{E}\,\|\hat{f}^{\mathrm{SC}}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}^{\mathrm{SC}}_{\alpha,\sigma}\|^2 = \int_{\mathbb{S}} \Phi^{\mathrm{SC}}_\alpha(\rho)^2\mathbf{Var}\,(UK^*\varepsilon)\,\mathrm{d}\Sigma \ge \gamma_{\mathbf{var}}\int_{\{\rho\ge\alpha\}}\frac{1}{\rho}\,\mathrm{d}\Sigma = -\gamma_{\mathbf{var}}\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta).$$

On the other hand, using Lemma 12 and 13 we can bound the first term on the right hand side of (6.1a) as follows

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta) \sim -\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}S(\beta) \lesssim -\kappa\int_\alpha^{\bar{\alpha}}\frac{1}{\beta}\,\mathrm{d}S(\beta) \lesssim -\kappa\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta).$$

This yields (3.13). (3.14) follows from (3.1), (3.4), and (3.13). ■

Proof of Theorem 6. Using the substitution $\alpha = \overline{\gamma_2}(\gamma_1(\beta))$ and Theorem 3, this follows from

$$\begin{aligned} \inf_{\alpha>0}\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|^2 &\lesssim \inf_{\alpha>0}\left(\gamma_\Lambda^2\Lambda(\alpha)^2\overline{w}^2 + (C_3^2 + C_2^2)\sigma^2\gamma_2(\alpha)\right) \\ &= \inf_{\beta>0}\left(\gamma_\Lambda^2\Lambda(\overline{\gamma_2}(\gamma_1(\beta)))^2\overline{w}^2 + (C_3^2 + C_2^2)\sigma^2\gamma_1(\beta)\right) \\ &\le \inf_{\beta>0}\left(C\gamma_\Lambda^2\Lambda(\beta)^2\overline{w}^2 + (C_3^2 + C_2^2)\sigma^2\gamma_1(\beta)\right). \end{aligned}$$

Acknowledgment. We would like to thank P. Bühlmann and L. Rosasco for helpful discussions on L^2-boosting. Moreover, we are grateful to unknown referees for their valuable comments which helped to improve the presentation of the material. N.Bissantz, T. Hohage and A. Munk gratefully acknowledge financial support of the Graduiertenkolleg 1023 "Identification in Mathematical Models", DFG grant MU 1230/8-1, and by the DFG "Sonderforschungsbereich" 475.

REFERENCES

[1] F. Abramovich and B. W. Silverman. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85:115–129, 1998.
[2] F. Bauer, T. Hohage, and A. Munk. Regularized newton methods for nonlinear inverse problems with random noise. in preparation.
[3] F. Bauer and S. Pereverzev. Regularization without preliminary knowledge of smoothness and error behavior. *Eur. J. of Applied Mathematics*, 16:303–317, 2005.
[4] N. Bissantz. Iterative inversion methods for statistical inverse problems. In L. Lyons et al., editors, *Phystat05: Proceedings of the conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology*, 2006.
[5] N. Bissantz, T. Hohage, and A. Munk. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20:1773–1791, 2004.
[6] H. Brakhage. On ill-posed problems and the method of conjugate gradients. In H. W. Engl and C. W. Groetsch, editors, *Inverse and ill-posed Problems*, pages 191–205. Academic Press, Orlando, 1987.
[7] P. Bühlmann and B. Yu. Boosting with l_2 loss: Regression and classification. *J. Am. Stat. Ass.*, 98:324–339, 2003.
[8] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Stat.*, 30:843–874, 2002.
[9] D. D. Cox. Approximation of method of regularization estimators. *Ann. Stat.*, 16:694–712, 1988.

[10] P. J. Diggle and P. Hall. A Fourier approach to nonparametric deconvolution of a density estimate. *J. R. Statist. Soc. B*, 55:523–531, 1993.

[11] D. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2:101–126, 1995.

[12] D. L. Donoho. Statistical estimation and optimal recovery. *Ann. Stat.*, 22:238–270, 1994.

[13] S. Efromovich. Robust and efficient recovery of a signal passed through a filter and then contaminated by non-gaussian noise. *IEEE Trans. Inform. Theory*, 43:1184–1191, 1997.

[14] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publisher, Dordrecht, Boston, London, 1996.

[15] H. W. Engl and J. Zou. A new approach to convergence rate analysis of tikhonov regularization for parameter identification problems in heat conduction. *Inverse Problems*, 16:1907–1923, 2000.

[16] S. N. Evans and P. B. Stark. Inverse problems as statistics. *Inverse Problems*, 18:R55–R97, 2002.

[17] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Stat.*, 19:1257–1272, 1991.

[18] Y. Freund and R. E. Schapire. Experiments wtih a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International conference*, pages 148–156, San Francisco, 1996. Morgan Kauffman.

[19] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, 29:1189–1232, 2001.

[20] P. R. Halmos. What does the spectral theorem say? *Amer. Math. Monthly*, 70:241–247, 1963.

[21] D. M. Healy, H. Hendriks, and P. T. Kim. Spherical deconvolution. *J. Multivariate Anal.*, 67:1–22, 1998.

[22] E. Heinz. Beiträge zur Störungstheorie der Spektralzerlegung. *Math. Ann.*, 123:425–438, 1951.

[23] T. Hohage. Regularization of exponentially ill-posed problems. *Numer. Funct. Anal. Optim.*, 21:439–464, 2000.

[24] I. M. Johnstone, G. Kerkyacharian, D. Picard, and M. Raimondo. Wavelet deconvolution in a periodic setting. *J. R. Statist. Soc. B*, 66:547–573, 2004.

[25] I. M. Johnstone and B. W. Silverman. Speed of estimation in positron emission tomography and related inverse problems. *Ann. Stat.*, 18:251–280, 1990.

[26] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, New York, 2004.

[27] P. T. Kim and J.-Y. Koo. Optimal spherical deconvolution. *J. Multivariate Anal.*, 80:21–42, 2002.

[28] R. Kreß. *Linear Integral Equations*. Springer Verlag, Berlin, Heidelberg, New York, 2nd edition, 1999.

[29] B. A. Mair. Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM J. Math. Anal.*, 25:135–147, 1994.

[30] B. A. Mair and F. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56:1424–1444, 1996.

[31] P. Mathé and S. Pereverzev. Optimal discretization of inverse problems in Hilbert scales. regularization and self-regularization of projection methods. *SIAM J. Numer. Anal.*, 38:1999–2021, 2001.

[32] P. Mathé and S. Pereverzev. Geometry of ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19:789–803, 2003.

[33] P. Mathé and S. Pereverzev. Regularization of some linear ill-posed problems with discretized random noisy data. *Math. Comp.*, 2006. to appear.

[34] A. Munk. Testing the goodness of fit of parametric regression models with random Toeplitz forms. *Scand. J. Statist.*, 29:501–535, 2002.

[35] A. S. Nemirovskii and B. T. Polyak. Iterative methods for solving linear ill-posed problems under precise information i. *Engrg. Cybernetics*, 22:1–11, 1984.

[36] D. W. Nychka and D. Cox. Convergence rates for regularized solutions of integral equations from discrete noisy data. *Ann. Stat.*, 17(2):556–572, 1989.

[37] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 4:502–527, 1986.

[38] L. Stefanski and R. J. Carroll. Deconvoluting kernel density estimators. *Statistics*, 21:169–184, 1990.

[39] M. Taylor. *Partial Differential Equations: Basic Theory*, volume 1. Springer Verlag, New York, 1996.

[40] M. Taylor. *Partial Differential Equations: Qualitative Studies of Linear Equations*, volume 2. Springer Verlag, New York, 1996.

[41] G. Wahba. Practical approximate solutions to linear operator equations when data are noisy. *SIAM J. Numer. Anal.*, 14:651–667, 1977.

[42] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. http://mathberkeley.edu/yao/publications/earlystop.pdf.

[43] T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. *Ann. Stat.*, 33:1539–1579, 2005.