SIAM J. NUMER. ANAL.
Vol. 58, No. 6, pp. 3504–3529

CONVERGENCE RATES OF SPECTRAL REGULARIZATION METHODS: A COMPARISON BETWEEN ILL-POSED INVERSE PROBLEMS AND STATISTICAL KERNEL LEARNING*

SABRINA GUASTAVINO† AND FEDERICO BENVENUTO†

Abstract. In this paper we study the relation between convergence rates of spectral regularization methods under Hölder-type source conditions resulting from the theory of ill-posed inverse problems, when the noise level δ goes to 0, and convergence rates resulting from statistical kernel learning, when the number of samples n goes to infinity. Toward this aim, we introduce a family of hybrid estimators in the statistical learning context whose convergence rates have the following properties: first, they are equal to those of spectral methods, and second, they are connected to the rates of spectral regularization in ill-posed inverse problems, provided that a suitable inverse proportionality relation between n and δ holds true. This family of estimators allows us to convert upper rates depending on n to upper rates depending on δ and to convert lower rates vice versa, quantifying their deviation. The analysis is carried out under general source conditions in the case the rank of the forward operator is both finite and infinite, and, in the latter case, both by not making any assumptions on the eigenvalues and by assuming a polynomial eigenvalue decay.

Key words. linear ill-posed inverse problems, statistical kernel learning, spectral regularization, convergence rates

AMS subject classifications. 62G20, 65J20, 68Q32

DOI. 10.1137/19M1256038

1. Introduction.

Finding approximate solutions to linear operator equations when the data are noisy is a common issue of ill-posed inverse problems and supervised kernel learning [16, 19, 37, 38]. In both contexts the method of regularization is used [14, 15, 34, 9, 41, 30, 39, 29, 1, 22]. However, there is a substantial difference between the two fields: in ill-posed inverse problems noisy data are usually considered infinite dimensional, whereas in kernel learning data are a finite set of samples. In the literature, regularization methods have been studied in these two contexts providing error convergence rates under smoothness assumptions on the solution. For spectral regularization in ill-posed inverse problems, convergence rates depending on the noise level δ have been known for years under Hölder-type source conditions [12, 2]. Moreover, a lot of work has been done in the case of general source conditions [18, 27, 26, 25, 28] and variational source conditions which allow for nonquadratic fidelity and regularization terms (see, e.g., [13, 35] and references therein). In addition to variational source conditions, smoothness based on stability estimates [40] and also penalty-based smoothness [17] have been recently considered. On the other hand, results on optimal convergence rates for kernel learning and inverse learning depending on the number of samples n in the case of general source conditions are quite recent [4, 6, 8, 21, 32, 33, 39, 41, 31]. An advanced monograph with error bounds for a regularization algorithm in learning theory is [23]. The question naturally arises

*Received by the editors April 12, 2019; accepted for publication (in revised form) September 2, 2020; published electronically December 18, 2020.
https://doi.org/10.1137/19M1256038
Funding: The work of the authors was partially supported by the INdAM Research group GNCS. The work of the authors was also supported by the European Union's Horizon 2020 research and innovation programme under grant 640216, (FLARECAST) project "Flare Likelihood And Region Eruption foreCASTing."
†Department of Mathematics, Università degli Studi di Genova, 16146 Genova, Italy (guastavino@dima.unige.it, benvenut@dima.unige.it).

whether the above rates are comparable and, if this is the case, which relation occurs between δ and n for quantifying the difference between optimal rates in the two contexts. An interesting approach that sheds some light on this topic has been proposed in [11]. The main objective of that paper was to draw a connection between consistency in kernel learning and regularization in inverse problems: in order to do so, authors formalized kernel learning as an inverse problem where the forward operator is an inclusion, but the inverse problem thus obtained is essentially well-posed and no connection of convergence rates is provided.

In this paper we answer this question by introducing a statistical estimator with the following two properties. First, it has the same upper rates of the spectral regularization considered in statistical learning: our analysis of the convergence rates of the proposed estimator is based on the results in [5], where a comprehensive study on the convergence rates with infinite dimensional deterministic and stochastic noise is given. Second, the rates of the proposed estimator are related to those of the classical spectral regularization for deterministic ill-posed inverse problems. Indeed, we prove that the expected error of this estimator given n samples is an upper bound of the error of the spectral regularization given the noise level δ, provided that a suitable relation between n and δ holds true. This allows us to convert upper rates with respect to the number of samples n to upper rates with respect to the noise level δ and, conversely, to convert lower rates depending on δ to lower rates depending on n. The conversion procedure turns out to be independent of the considered source condition and therefore it is applicable in any case. In this paper, we give the general conversion result and then we apply it for comparing optimal convergence rates obtained in the two contexts for spectral regularization under general source conditions and for quantifying their difference, showing that they exactly match when the rank of the linear operator is finite. However, we prove that, in general, they do not correspond to each other being the conversion procedure based on a tight inequality.

The paper is organized as follows. In section 2 we introduce the spectral regularization, and we present the two main contexts in which optimal rates have been considered: inverse problems and statistical learning theory. Then, we briefly recall the hypotheses on the sought solution, namely the general source condition, and on the model, namely the polynomial decay of the eigenvalues of the linear operator. In section 3 we establish the main result of the paper, i.e., a theorem for converting results about convergence rates obtained with respect to the number of samples (in the statistical framework) to results on convergence rates with respect to the noise level (in the infinite dimensional inverse problems framework) and vice versa. Moreover, we compare and quantify the difference between the achieved rates in the two frameworks, showing that the optimal rates obtained in statistical learning are usually weaker than the optimal ones obtained in the deterministic error analysis. In section 4 we present the conclusions of our study. In section 5 we provide the proofs of the results.

2. Preliminaries.

Both ill-posed inverse problems and statistical learning deal with a bounded linear operator A: in inverse problems A is the operator to be (approximately) inverted; in statistical learning A relates to the feature map [6]. We start by making some assumptions on the operator A used to prove the convergence rates in the context of statistical inverse learning. Let $\mathcal{X}$ be a standard Borel space endowed with a measure ν. Let $\mathcal{H}_1$ be a separable Hilbert space, $\mathcal{H}_2 := L^2(\mathcal{X}, \nu)$, and let A be a bounded linear operator, $A : \mathcal{H}_1 \to \mathcal{H}_2$. We assume that A is uniformly bounded, i.e., there exists a constant $c > 0$ such that

$$|Af(x)| \leq c\|f\|_{\mathcal{H}_1}, \tag{2.1}$$

for all $x \in \mathcal{X}$ and for all $f \in \mathcal{H}_1$. This assumption together with the Riesz representation theorem implies that for all x there exists an element $\phi_x \in \mathcal{H}_1$ such that

$$(Af)(x) = \langle f, \phi_x \rangle_{\mathcal{H}_1}. \tag{2.2}$$

Moreover, the range of A is a subset of $L^2(\mathcal{X}, \nu)$, and it is well known that it is a reproducing kernel Hilbert space with kernel $k(x, x') = \langle \phi_x, \phi_{x'} \rangle_{\mathcal{H}_1}$ (e.g., see [36, 20, 3]). Therefore, the adjoint operator $A^* : \mathcal{H}_2 \to \mathcal{H}_1$ of A takes the form

$$A^* g = \int_{\mathcal{X}} g(x) \phi_x d\nu(x), \tag{2.3}$$

and the operator A^*A is given by

$$A^*A = \int_{\mathcal{X}} \langle \cdot, \phi_x \rangle_{\mathcal{H}_1} \phi_x d\nu(x). \tag{2.4}$$

As the operator A^*A is self-adjoint and compact, there exists an orthonormal basis consisting of eigenfunctions of A with real eigenvalues. Furthermore, the operator A^*A is of trace class. The proofs of such properties can be found in [10].

Finally, for any set $\{X_1, \ldots, X_n\} \subset \mathcal{X}$ we consider the operator $A_n : \mathcal{H}_1 \to \mathbb{R}^n$ as follows:

$$(A_n f)_i := \langle f, \phi_{X_i} \rangle_{\mathcal{H}_1} \tag{2.5}$$

for $i = 1, \ldots, n$ and $f \in \mathcal{H}_1$. Its adjoint operator $A_n^* : \mathbb{R}^n \to \mathcal{H}_1$ is given by

$$A_n^* z = \frac{1}{n} \sum_{i=1}^{n} z_i \phi_{X_i} \text{ for } z \in \mathbb{R}^n, \tag{2.6}$$

and the operator $A_n^* A_n : \mathcal{H}_1 \to \mathcal{H}_1$ is given by

$$A_n^* A_n = \frac{1}{n} \sum_{i=1}^{n} \langle \cdot, \phi_{X_i} \rangle_{\mathcal{H}_1} \phi_{X_i}. \tag{2.7}$$

The assumption $\mathcal{H}_2 := L^2(\mathcal{X}, \nu)$ allows us to consider convergence rate results in both inverse problems and statistical learning theory.

2.1. Spectral regularization. A spectral regularization is a map $R : \mathcal{B} \times \mathcal{H}_2 \times \mathbb{R}_+ \to \mathcal{H}_1$ defined by

$$R(A, y, \lambda) := s_\lambda(A^*A) A^* y, \tag{2.8}$$

where $\mathcal{B}$ denotes the space of bounded linear operators $\mathcal{H}_1 \to \mathcal{H}_2$, and s_λ denotes the *regularization function* defined as follows [12].

DEFINITION 2.1. *The regularization (or filtering) function s_λ for $\lambda > 0$ is defined on the spectrum of A^*A, denoted by $\sigma(A^*A)$, and satisfies the following properties:*

1. *There exists a constant $D > 0$ such that*

$$\sup_{t \in \sigma(A^*A)} |t s_\lambda(t)| \le D \quad \textit{uniformly in} \quad \lambda > 0. \tag{2.9}$$

2. *There exists a constant $E > 0$ such that*

$$\sup_{\lambda > 0} \sup_{t \in \sigma(A^*A)} |\lambda s_\lambda(t)| \le E. \tag{2.10}$$

3. *There exist $q > 0$ called qualification of the method and constants $C_\nu > 0$ such that*

$$\sup_{t\in\sigma(A^*A)} |t^\nu(1-ts_\lambda(t))| \le C_\nu\lambda^\nu \quad \forall\ \lambda > 0 \quad \textit{and} \quad 0 \le \nu \le q. \tag{2.11}$$

The idea of spectral regularization is to provide approximated solutions of linear operator equations with noisy data. A typical example is Tikhonov regularization: in this case the regularization function is given by $s_\lambda(t) = (\lambda+t)^{-1}$ and the qualification is $q = 1$.

Now we introduce the two main assumptions on the noise: the first is usually considered in the study of ill-posed inverse problems, whereas the second has been considered both in the case of statistical (inverse) learning and of ill-posed inverse problems (see [38]).

2.2. Infinite dimensional deterministic noise. Spectral regularization has been introduced in ill-posed inverse problems theory to approximately solve

$$Af = y \tag{2.12}$$

when $y \in \mathcal{H}_2$ is not known and only a noisy version y^δ of the data is available. The spectral regularized solution takes the form

$$f_\delta^\lambda := R(A, y^\delta, \lambda) = s_\lambda(A^*A)A^*y^\delta\ . \tag{2.13}$$

In this context the noisy data are infinite dimensional and the relation with the exact data y is $\|y^\delta - y\|_{\mathcal{H}_2} \le \delta$ for some $\delta > 0$ representing the noise level. As f_δ^λ continuously depends on the data, it converges to the generalized solution $f^\dagger$. The convergence rates are studied with respect to $\delta \to 0$, and source conditions are considered in order to obtain a rate in the following form:

$$\|f_\delta^\lambda - f^\dagger\|_{\mathcal{H}_1} \in O(\alpha(\delta)) \tag{2.14}$$

where $\lambda = \lambda(\delta)$ is suitably chosen so that $\alpha : (0,\infty) \to (0,\infty)$ is an increasing function with respect to δ.

2.3. Finite dimensional stochastic noise. In the context of supervised (inverse) learning the noise is formalized differently. In this case, instead of knowing an infinite dimensional noisy version of the data y, we assume to know a set of noisy samples $\{(X_i, Y_i)\}_{i=1}^n$. In particular, one supposes that each (X_i, Y_i) is independently drawn from a given (but unknown) probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$ where the input space $\mathcal{X} \subseteq \mathbb{R}^p$ and the output space $\mathcal{Y} \subseteq \mathbb{R}$. We assume that ρ satisfies the factorization property $\rho(X, Y) = \rho(Y|X)\nu(X)$, where ν is the marginal distribution on $\mathcal{X}$ and $\rho(\cdot|X = x)$ is the conditional distribution on $\mathcal{Y}$ for almost all $x \in \mathcal{X}$. We assume that the conditional expectation with respect to $\rho(\cdot|\cdot)$ of Y given X is equal to

$$\mathbb{E}(Y|X = x) = Af^\dagger(x) = y(x) \tag{2.15}$$

and that the variance of the conditional probability is

$$Var(Y|X = x) = \sigma^2 \tag{2.16}$$

for ν-almost $x \in \mathcal{X}$, where σ is a constant. In this setting, spectral regularization takes the form

$$\hat{f}^{\lambda}_{n,\mathrm{learn}} = R(A_n, \mathbf{y}, \lambda) = s_\lambda(A_n^* A_n) A_n^* \mathbf{y}, \tag{2.17}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)$. The convergence rates of the estimator $\hat{f}^{\lambda}_{n,\mathrm{learn}}$ are studied with respect to $n \to \infty$ in expectation (or in probability), and source conditions are considered in order to obtain rates in the following form:

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}^{\lambda}_{n,\mathrm{learn}} - f^\dagger\|_{\mathcal{H}_1}) \in O\left(\alpha(n^{-1})\right) , \tag{2.18}$$

where $\rho^{\otimes n}$ indicates the distribution tensor product and $\lambda = \lambda(n)$ is chosen so that $\alpha : (0, \infty) \to (0, \infty)$ is an increasing function with respect to n^{-1}.

2.4. Model and source conditions. In the following we give a brief review about the main results on convergence rates for spectral regularized estimators in statistical learning theory and in deterministic infinite dimensional inverse problems under general source conditions characterized by the so-called index function [7, 18, 27].

The general source condition used in inverse problems is a requirement on the form of the solution, i.e.,

$$f^\dagger \in \omega(\Phi, R) := \{f \in \mathcal{H}_1 \ : \ f = \Phi(B)w, \ \|w\|_{\mathcal{H}_1} \leq R\}, \tag{2.19}$$

where $B := A^*A$, $R > 0$, and Φ is the index function, i.e., it is continuous and strictly increasing on the interval $[0, \|B\|]$, with $\Phi(0) = 0$. The general source condition reduces to the Hölder-type source condition when $\Phi(t) = t^r$, with $r > 0$, and to the logarithmic source condition when $\Phi(t) = t^p \log^{-m}(\frac{1}{t})$, with $p \in \mathbb{N}$ and $m > 0$. General source conditions are common in both statistical learning and infinite dimensional deterministic inverse problems theory, although in statistical learning source conditions are imposed in terms of restrictions of the conditional probability $\rho(\cdot|\cdot)$ (for details see [6, 4, 32]).

Furthermore, in the statistical learning setting another assumption on the eigenvalue decay of the operator B is considered in order to improve the convergence rates. We assume that

$$\frac{c}{j^b} \leq \mu_j \leq \frac{d}{j^b}, \tag{2.20}$$

where μ_j are the eigenvalues of B for each $j \in \mathbb{N}$, $j \geq 1$, $d, c > 0$, and $b > 1$. Again, in the statistical framework this assumption is given as a requirement on the probability ν on which B depends. Together with the first assumption, they can be expressed in statistical learning as a single restriction on the probability space by requiring that ρ belongs to a suitable subspace (for details see [6]). As we will see in the next subsection, it is well known that the latter assumption does not improve the convergence rates given in the deterministic infinite dimensional inverse problems setting, which, instead, are independent of any restriction on the eigenvalue decay. For further details on this topic see the introduction of [5].

2.5. Existing convergence rates. In Table 2.1 we report a summary on the convergence rates given in statistical learning and in deterministic ill-posed inverse problems for spectral regularization methods under the general source condition (2.19) and according to different assumptions on the operator A [4, 8, 6, 12, 18]. Furthermore, in Table 2.2 we make explicit the polynomial convergence rates obtained by considering the Hölder-type source condition. Whereas convergence rates for ill-posed inverse problems are independent of the assumption on the operator A, these hypotheses are crucial to improve the convergence rates in the case of statistical learning. First, assuming a polynomial decay of the eigenvalues of the operator $B = A^*A$ with exponent $b > 1$, convergence rates improve and they become faster and faster as b increases (this case is treated in [8] for Tikhonov regularization under the Hölder-type source condition, generalized for spectral regularization in [6] and under general source conditions in [32]). When b goes to 1, the rate corresponds to the one obtained without assuming any further condition on the eigenvalue decay (this latter case is studied in [33] for Tikhonov regularization under the Hölder-type source condition and in [4] for spectral regularization under general source conditions). Second, assuming that the set of nonzero eigenvalues is finite, i.e., B has finite rank, the fastest possible convergence rate is reached. Such a rate is the limit rate achieved when b goes to infinity under the polynomial eigenvalue decay hypothesis. Indeed, if we consider the upper rates in the latter case as b goes to infinity, the eigenvalues are $\mu_1 \leq d$ and $\mu_j = 0$ for $j \geq 2$, i.e., the rank is finite. The finite rank case is treated in [8] for Tikhonov regularization under the Hölder-type source condition, and it can be generalized for spectral regularization under general source conditions using results in [6, 32] and considering the effective dimension finite.

Table 2.1

Existing upper rates under the general source condition (see (2.19)) in statistical learning and ill-posed inverse problems with increasing n and decreasing δ, respectively.

Assumption on A	$\mathbb{E}_{\rho^{\otimes n}}(\Vert\hat{f}^{\lambda}_{n,\mathrm{learn}} - f^\dagger\Vert_{\mathcal{H}_1})$	$\Vert f^{\lambda}_{\delta} - f^\dagger\Vert_{\mathcal{H}_1}$
–	$\Phi(\Psi_1^{-1}(n^{-\frac{1}{2}}))$ with $\Psi_1(t) := t\Phi(t)$	
Eigenvalue decay (2.20)	$\Phi(\Psi_b^{-1}(n^{-\frac{1}{2}}))$ with $\Psi_b(t) := t^{\frac{1}{2}+\frac{1}{2b}}\Phi(t)$	$\Phi(\Psi^{-1}(\delta))$
Finite rank	$\Phi(\Psi^{-1}(n^{-\frac{1}{2}}))$ with $\Psi(t) := \sqrt{t}\Phi(t)$	

Table 2.2

Existing upper rates under the Hölder-type source condition (see (2.19) with $\Phi(t) = t^r$, $0 < r \leq q$, denoting q the qualification of the spectral regularization method) in statistical learning and ill-posed inverse problems with increasing n and decreasing δ, respectively.

Assumption on A	$\mathbb{E}_{\rho^{\otimes n}}(\Vert\hat{f}^{\lambda}_{n,\mathrm{learn}} - f^\dagger\Vert_{\mathcal{H}_1})$	$\Vert f^{\lambda}_{\delta} - f^\dagger\Vert_{\mathcal{H}_1}$
–	$\left(\frac{1}{n}\right)^{\frac{r}{2r+2}}$	
Eigenvalue decay (2.20)	$\left(\frac{1}{n}\right)^{\frac{r}{2r+1+\frac{1}{b}}}$	$\delta^{\frac{2r}{2r+1}}$
Finite rank	$\left(\frac{1}{n}\right)^{\frac{r}{2r+1}}$	

3. Connection between convergence rates.

The main difference between the study of the convergence rates in statistical learning and ill-posed inverse problems with deterministic noise lies in the independent variable on which the error depends. Whereas for learning problems the independent variable is the number of examples n, for inverse problems it is the noise level δ of infinite dimensional noisy data. The relation between the rates provided in these two settings under the same source condition is not straightforward. It is evident that there is no direct transformation between n

and δ. To establish such a relation we need to introduce the following estimator:

$$\hat{f}_n^\lambda := s_\lambda(A^*A)A_n^*\mathbf{y}. \tag{3.1}$$

We refer to estimator (3.1) as the *hybrid* estimator as it is halfway between the spectral regularization for ill-posed problems and for statistical learning. Indeed, it is the composition of two terms: a regularization function s_λ depending on the exact operator A^*A, and the empirical backprojection of the data depending on the sampling operator A_n. This estimator has been introduced in the literature in [24]. We are interested in this estimator as it has the following two properties:

(i) The error given by this estimator is always larger than the error given by the standard spectral regularized solution, provided that a suitable relation between n and δ holds true.

(ii) It has the same upper rates of the spectral regularized estimator $\hat{f}^\lambda_{n,\text{learn}}$.

The first property allows us to convert upper convergence rates depending on n to upper convergence rates depending on δ and vice versa to convert lower convergence rates depending on δ to lower convergence rates depending on n (subsection 3.1). The second property assures that this estimator is the same as the spectral estimator in statistical learning in terms of upper rates (subsection 3.2).

3.1. A link between the number of examples n and the noise level δ. The first property of the hybrid estimator $\hat{f}_n^\lambda$ is summarized in the following.

PROPOSITION 3.1. *Consider the spectral regularization f_δ^λ defined in* (2.13) *and the hybrid estimator $\hat{f}_n^\lambda$ defined in* (3.1). *Let $\mathcal{H}_1$ be embedded in the space of square integrable functions. Let us consider n samples identically and independently drawn according to a distribution ρ as in section* 2.3. *Let*

$$\varepsilon(\lambda) := \|f^\lambda - f^\dagger\|_{\mathcal{H}_1}\|s_\lambda(A^*A)A^*\|_{HS}^{-1}, \tag{3.2}$$

where $f^\lambda := R(A, y, \lambda) = s_\lambda(A^*A)A^*y$ and $\|\cdot\|_{HS}$ denotes the Hilbert–Schmidt norm. For each $n > 0$, there exists a function

$$\Delta(n,\lambda) = \frac{1}{\sqrt{\frac{\sigma^2}{n} + \varepsilon(\lambda)^2} + \varepsilon(\lambda)} \frac{\sigma^2}{n} \ , \tag{3.3}$$

such that for each $0 < \delta \le \Delta(n,\lambda)$ and infinite dimensional noisy data y^δ such that $\|y^\delta - y\|_{\mathcal{H}_2} \le \delta$, the following inequality holds:

$$\|f_\delta^\lambda - f^\dagger\|^2_{\mathcal{H}_1} \le \mathbb{E}_{\rho^{\otimes n}}\left(\|\hat{f}_n^\lambda - f^\dagger\|^2_{\mathcal{H}_1}\right) \ . \tag{3.4}$$

Conversely, for each $\delta > 0$ there exists a function

$$N(\delta,\lambda) = \frac{\sigma^2}{\delta^2 + 2\delta\varepsilon(\lambda)}. \tag{3.5}$$

such that for each $n \in \mathbb{N}$ such that $0 < n \le N(\delta,\lambda)$ inequality (3.4) *applies.*

Thanks to the result in Proposition 3.1 we can relate a given upper convergence rate computed with respect to n (for the hybrid estimator $\hat{f}_n^\lambda$) to the one computed with respect to δ (for the spectral regularized solution f_δ^λ). It is worth noticing that results in Theorem 3.3 and Theorem 3.4 are independent of the choice of the source condition. From now on, in order to express asymptotic behaviors we make use of the Landau symbols O, Ω, and Θ, defined as follows.

DEFINITION 3.2. *Let $g, h : (0, \infty) \to \mathbb{R}$ be real positive valued functions.*

1. *$g(\delta) \in O(h(\delta))$ as $\delta \to 0$ if there exist $M > 0$ and $\delta_0 > 0$ such that $g(\delta) \leq Mh(\delta)$ for each $0 < \delta < \delta_0$.*
2. *$g(\delta) \in \Omega(h(\delta))$ as $\delta \to 0$ if there exist $M > 0$ and $\delta_0 > 0$ such that $g(\delta) \geq Mh(\delta)$ for each $0 < \delta < \delta_0$.*
3. *$g(\delta) \in \Theta(h(\delta))$ as $\delta \to 0$ if $g(\delta) \in O(h(\delta))$ and $g(\delta) \in \Omega(h(\delta))$, as $\delta \to 0$.*

We define $g(n^{-1}) \in O(h(n^{-1}))$, $g(n^{-1}) \in \Omega(h(n^{-1}))$, and $g(n^{-1}) \in \Theta(h(n^{-1}))$ as $n \to \infty$, as in Definition 3.2 taking $\delta = n^{-1}$.

THEOREM 3.3. *Let α and τ be two continuous increasing functions $\alpha, \tau : (0, \infty) \to (0, \infty)$ such that $\alpha(z), \tau(z) \to 0$ as $z \to 0$. Let the upper rate of the hybrid estimator $\hat{f}_n^\lambda$ be described by*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}^2) \in O\left(\alpha\left(n^{-1}\right)\right) \quad \text{as } n \to \infty \tag{3.6}$$

for a given $\lambda = \lambda_n \in \Theta\left(\tau\left(n^{-1}\right)\right)$ as $n \to \infty$. Then the upper rate of the error of the estimator f_δ^λ defined in (2.13) with respect to the noise level $\delta \to 0$ is given by

$$\|f_\delta^\lambda - f^\dagger\|_{\mathcal{H}_1}^2 \in O\left(\alpha(k(\delta))\right), \tag{3.7}$$

where y^δ is such that $\|y^\delta - y\|_{\mathcal{H}_2} \leq \delta$; $\lambda = \lambda_\delta$ has the rate

$$\lambda_\delta \in \begin{cases} \Theta\left(\tau(\delta^2)\right) & \text{if} \quad \epsilon(\lambda_n) \in O(n^{-\frac{1}{2}}), \\ \Theta\left(\tau(c^{-1}(\delta))\right) & \text{if} \quad \epsilon(\lambda_n) \in \Omega(n^{-\frac{1}{2}}), \end{cases} \tag{3.8}$$

where c^{-1} denotes the inverse of the function $c : z \in (0, \infty) \mapsto c(z) \in (0, \infty)$, which is a continuous and strictly increasing function with respect to the variable z, and c is such that $c(n^{-1}) \in \Theta((n\epsilon(\lambda_n))^{-1})$ as $n \to \infty$; and

$$k(\delta) = \begin{cases} \delta^2 & \text{if} \quad \epsilon(\lambda_\delta) \in O(\delta), \\ (d(\delta))^{-1} & \text{if} \quad \epsilon(\lambda_\delta) \in \Omega(\delta), \end{cases} \tag{3.9}$$

where $d : z \in (0, \infty) \mapsto d(z) \in (0, \infty)$ is a continuous strictly decreasing function with respect to z and d is such that $d(\delta) \in \Theta((\delta\epsilon(\lambda_\delta))^{-1})$ as $\delta \to 0$.

Now we give the converse result on lower rates.

THEOREM 3.4. *Let α and φ be two continuous increasing functions $\alpha, \varphi : (0, \infty) \to (0, \infty)$ such that $\alpha(z), \varphi(z) \to 0$ as $z \to 0$. Let f_δ^λ be defined as in (2.13). Let the lower rate of f_δ^λ be described by*

$$\|f_\delta^\lambda - f^\dagger\|_{\mathcal{H}_1}^2 \in \Omega(\alpha(\delta)) \text{ as } \delta \to 0, \tag{3.10}$$

where y^δ is such that $\|y^\delta - y\|_{\mathcal{H}_2} \leq \delta$ and $\lambda = \lambda_\delta \in \Theta(\varphi(\delta))$, as $\delta \to 0$. Then the lower rate of the hybrid estimator $\hat{f}_n^\lambda$ with respect to the number of samples $n \to \infty$ is given by

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}^2) \in \Omega\left(\alpha(k(n^{-1}))\right), \tag{3.11}$$

where $\lambda = \lambda_n$ has the rate

$$\lambda_n \in \begin{cases} \Theta\left(\varphi(n^{-\frac{1}{2}})\right) & \text{if} \quad \epsilon(\lambda_\delta) \in O(\delta), \\ \Theta\left(\varphi(d^{-1}(n^{-1}))\right) & \text{if} \quad \epsilon(\lambda_\delta) \in \Omega(\delta), \end{cases} \tag{3.12}$$

where d^{-1} denotes the inverse of the function $d : z \in (0,\infty) \mapsto d(z) \in (0,\infty)$, which is a continuous strictly decreasing function with respect to z, and d is such that $d(\delta) \in \Theta((\delta\epsilon(\lambda_\delta))^{-1})$ as $\delta \to 0$, and

$$k(n^{-1}) = \begin{cases} n^{-\frac{1}{2}} & \text{if} \quad \epsilon(\lambda_n) \in O(n^{-\frac{1}{2}}), \\ c(n^{-1}) & \text{if} \quad \epsilon(\lambda_n) \in \Omega(n^{-\frac{1}{2}}), \end{cases} \tag{3.13}$$

where $c : z \in (0,\infty) \mapsto c(z) \in (0,\infty)$ is a continuous strictly increasing function with respect to z, and c is such that $c(n^{-1}) \in \Theta((n\epsilon(\lambda_n))^{-1})$ as $n \to \infty$.

Proofs of Proposition 3.1, Corollary 3.5, and Corollary 3.6 are given in section 5. In the case of polynomial convergence rates, results in Theorem 3.3 and Theorem 3.4 are remarkably simplified and we have the following.

Corollary 3.5. *Let the upper rate of the hybrid estimator $\hat{f}_n^\lambda$ be equal to $n^{-\alpha}$ for some $\alpha > 0$, i.e.,*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|^2_{\mathcal{H}_1}) \in O\left(\left(\frac{1}{n}\right)^\alpha\right) \quad \text{as } n \to \infty \tag{3.14}$$

for a given $\lambda = \lambda_n \in \Theta(n^{-p})$, as $n \to \infty$ with $p > 0$ and $\varepsilon(\lambda) \in \Theta(\lambda^\gamma)$, and as $\lambda \to 0$ with $\gamma > 0$. Then the upper rate of the error of the estimator f_δ^λ defined in (2.13) *with respect to the noise level $\delta \to 0$ is given by*

$$\|f_\delta^\lambda - f^\dagger\|^2_{\mathcal{H}_1} \in O\left(\delta^{\min\left(2\alpha, \frac{\alpha}{1-p\gamma}\right)}\right), \tag{3.15}$$

where y^δ is such that $\|y^\delta - y\|_{\mathcal{H}_2} \le \delta$, and $\lambda = \lambda_\delta$ has the following rate:

$$\lambda_\delta \in \Theta\left(\delta^{\min\left(2p, \frac{p}{1-p\gamma}\right)}\right) \quad \text{as } \delta \to 0. \tag{3.16}$$

Now we give the converse result on lower rates.

Corollary 3.6. *Let f_δ^λ be defined in* (2.13)*. Let the lower rate of f_δ^λ be equal to δ^α for some $\alpha > 0$, i.e.,*

$$\|f_\delta^\lambda - f^\dagger\|^2_{\mathcal{H}_1} \in \Omega(\delta^\alpha) \quad \text{as } \delta \to 0, \tag{3.17}$$

where y^δ is such that $\|y^\delta - y\|_{\mathcal{H}_2} \le \delta$, $\lambda = \lambda_\delta \in \Theta(\delta^{p^*})$ as $\delta \to 0$ with $p^* > 0$, and $\varepsilon(\lambda) \in \Theta(\lambda^\gamma)$ as $\lambda \to 0$ with $\gamma > 0$. Then the lower rate of the hybrid estimator $\hat{f}_n^\lambda$ with respect to the number of samples $n \to \infty$ is given by

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|^2_{\mathcal{H}_1}) \in \Omega\left(n^{-\max\left(\frac{\alpha}{2}, \frac{\alpha}{1+p^*\gamma}\right)}\right), \tag{3.18}$$

where $\lambda = \lambda_n$ has the following rate:

$$\lambda_n \in \Theta\left(n^{-\max\left(\frac{p^*}{2}, \frac{p^*}{1+p^*\gamma}\right)}\right) \quad \text{as } n \to \infty. \tag{3.19}$$

It is worth noting that not all such rate exponents α are possible and occur: in many cases existence requires specific additional conditions.

3.2. Upper rates of the hybrid estimator. We now present the result on the upper rates of the hybrid estimator $\hat{f}_n^\lambda$ under the general source condition (2.19) and according to different assumptions on the operator A. Let the model be described by (2.15) and (2.16). If we do not make any assumption on the singular values of A, we have the following result.

LEMMA 3.7. *Under the general source condition* (2.19) *we have*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}) \in O\left(\Phi\left(\Psi_1^{-1}\left(n^{-\frac{1}{2}}\right)\right)\right) \quad \text{as } n \to \infty, \tag{3.20}$$

with $\lambda \in \Theta(\Psi_1^{-1}(n^{-\frac{1}{2}}))$ *as* $n \to \infty$ *and* $\Psi_1(t) = t\Phi(t)$. *If we consider the Hölder-type source condition (i.e.,* (2.19) *with* $\Phi(t) = t^r$, *with* $r > 0$*), the result reduces to*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|^2_{\mathcal{H}_1}) \in O\left(\left(\frac{1}{n}\right)^{\frac{2r}{2r+2}}\right) \quad \text{as } n \to \infty, \tag{3.21}$$

with $\lambda \in \Theta((\frac{1}{n})^{\frac{1}{2r+2}})$ *as* $n \to \infty$.

Under the hypothesis of the polynomial eigenvalue decay of A^*A we have the following result.

LEMMA 3.8. *Under the general source condition* (2.19) *and assumption* (2.20) *we have*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}) \in O\left(\Phi\left(\Psi_b^{-1}\left(n^{-\frac{1}{2}}\right)\right)\right) \quad \text{as } n \to \infty, \tag{3.22}$$

with $\lambda \in \Theta(\Psi_b^{-1}(n^{-\frac{1}{2}}))$ *as* $n \to \infty$ *and* $\Psi_b(t) = t^{\frac{1}{2}+\frac{1}{2b}}\Phi(t)$. *If we consider the Hölder-type source condition (i.e.,* (2.19) *with* $\Phi(t) = t^r$, *with* $r > 0$*), the result reduces to*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|^2_{\mathcal{H}_1}) \in O\left(\left(\frac{1}{n}\right)^{\frac{2r}{2r+1+\frac{1}{b}}}\right) \quad \text{as } n \to \infty, \tag{3.23}$$

with $\lambda \in \Theta((\frac{1}{n})^{\frac{1}{2r+1+\frac{1}{b}}})$ *as* $n \to \infty$.

Finally, in the case the operator A^*A has finite rank we have the following.

LEMMA 3.9. *Under the general source condition* (2.19) *and under the hypothesis that the operator* A^*A *has finite rank* ($\mu_j = 0$ *for all* $j > F$, *where* F *is called effective dimension) we have*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}) \in O\left(\Phi\left(\Psi^{-1}\left(n^{-\frac{1}{2}}\right)\right)\right) \quad \text{as } n \to \infty, \tag{3.24}$$

with $\lambda \in \Theta(\Psi^{-1}(n^{-\frac{1}{2}}))$ *as* $n \to \infty$ *and* $\Psi(t) = \sqrt{t}\Phi(t)$. *If we consider the Hölder-type source condition (i.e.,* (2.19) *with* $\Phi(t) = t^r$, *with* $r > 0$*), the result reduces to*

$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|^2_{\mathcal{H}_1}) \in O\left(\left(\frac{1}{n}\right)^{\frac{2r}{2r+1}}\right) \quad \text{as } n \to \infty, \tag{3.25}$$

with $\lambda \in \Theta((\frac{1}{n})^{\frac{1}{2r+1}})$ *as* $n \to \infty$.

Proofs of Lemmas 3.7, 3.8, and 3.9 are given in section 5. We remark that the upper rates given in (3.22), (3.20), and (3.24) under general source conditions (and upper rates given in (3.23), (3.21), and (3.25) under the Hölder-type source condition) are the same ones of the classical spectral estimator $\hat{f}^{\lambda}_{n,\mathrm{learn}}$ defined in (2.17) (see Table 2.1 under general source conditions and Table 2.2 under the Hölder-type source condition).

3.3. Conversion of convergence rates. In the previous section we have shown that the estimator $\hat{f}^{\lambda}_{n}$ defined in (3.1) has the same upper rates of the standard statistical learning estimator $\hat{f}^{\lambda}_{n,\mathrm{learn}}$ defined in (2.17) for the same choice of the sequence λ_n. This allows us to use Corollary 3.5 to transform the upper rates depending on n in Table 2.2 to upper rates for the classical spectral regularization depending on δ. Let $f^{\lambda}_{\delta} := s_{\lambda}(A^*A)A^*y^{\delta}$ and γ be defined as in Corollary 3.5.

3.3.1. Upper rates. We now focus on the hybrid estimator and on its upper rates in the three cases considered in the last subsection. We have the following results.

COROLLARY 3.10. *Consider the hybrid estimator $\hat{f}^{\lambda_n}_{n}$ and its upper rates in* (3.20), (3.22)*, and* (3.24) *(Lemmas* 3.7*,* 3.8*, and* 3.9*). Let c and d be functions defined as in Theorem* 3.3*. Let $\tilde{\Psi}$ be defined according to the following cases:*

1. $\tilde{\Psi}(t) := \Psi_1(t) = t\Phi(t)$ *under assumption* (2.19)*;*
2. $\tilde{\Psi}(t) := \Psi_b(t) = t^{\frac{1}{2}+\frac{1}{2b}}\Phi(t)$ *under assumptions* (2.19) *and* (2.20)*;*
3. $\tilde{\Psi}(t) := \Psi(t) = \sqrt{t}\Phi(t)$ *under assumption* (2.19) *and assuming that the rank A^*A is finite.*

Let $\lambda_n \in \Theta(\tilde{\Psi}^{-1}(n^{-\frac{1}{2}}))$ as $n \to \infty$. Then, for the spectral regularization f^{λ}_{δ} we have the following cases:

- *If $\epsilon(\lambda_n) \in O(n^{-\frac{1}{2}})$ as $n \to \infty$, then*

$$\|f^{\lambda}_{\delta} - f^{\dagger}\|_{\mathcal{H}_1} \in O\left(\Phi(\tilde{\Psi}^{-1}(\delta))\right), \ \textit{with}\ \lambda \in \Theta\left(\tilde{\Psi}^{-1}(\delta)\right) \ \textit{as} \ \ \delta \to 0. \tag{3.26}$$

- *If $\epsilon(\lambda_n) \in \Omega(n^{-\frac{1}{2}})$ as $n \to \infty$, then*

$$\|f^{\lambda}_{\delta} - f^{\dagger}\|_{\mathcal{H}_1} \in O\Big(\Phi(\tilde{\Psi}^{-1}((d(\delta^{-1}))^{-\frac{1}{2}}))\Big), \ \textit{with}\ \lambda \in \Theta\Big(\tilde{\Psi}^{-1}((c^{-1}(\delta))^{\frac{1}{2}})\Big) \ \textit{as} \ \ \delta \to 0. \tag{3.27}$$

Now we apply these results to the case of the Hölder-type source condition. From the upper rates of the estimator $\hat{f}^{\lambda}_{n}$ given in (3.21), (3.23), and (3.21) (see Lemmas 3.7, 3.8, and 3.9) and Corollary 3.5 we obtain the following.

COROLLARY 3.11. *Let ϵ be defined in* (3.2)*, and let $\epsilon(\lambda) \in \Theta(\lambda^{\gamma})$, $\gamma \geq 0$, as in Corollary* 3.5*.*

1. *Under Hölder-type source conditions*

$$\|f^{\lambda}_{\delta} - f^{\dagger}\|_{\mathcal{H}_1} \in \begin{cases} O\left(\delta^{\frac{2r}{2r+2}}\right), & \gamma \geq r+1 \quad \textit{and} \quad \lambda \in \Theta\left(\delta^{\frac{2}{2r+2}}\right) \\ O\left(\delta^{\frac{r}{2r+2-\gamma}}\right), & \gamma < r+1 \quad \textit{and} \quad \lambda \in \Theta\left(\delta^{\frac{1}{2r+2-\gamma}}\right) \end{cases} \quad \textit{as}\ \delta \to 0. \tag{3.28}$$

2. *Under Hölder-type source conditions and assumption* (2.20)

(3.29)
$$\|f_\delta^\lambda - f^\dagger\|_{\mathcal{H}_1} \in \begin{cases} O\left(\delta^{\frac{2r}{2r+1+\frac{1}{b}}}\right), & \gamma \geq r+\frac{1}{2}+\frac{1}{2b} \quad \textit{and} \quad \lambda \in \Theta\left(\delta^{\frac{2}{2r+1+\frac{1}{b}}}\right) \\ O\left(\delta^{\frac{r}{2r+1+\frac{1}{b}-\gamma}}\right), & \gamma < r+\frac{1}{2}+\frac{1}{2b} \quad \textit{and} \quad \lambda \in \Theta\left(\delta^{\frac{1}{2r+1+\frac{1}{b}-\gamma}}\right) \end{cases} \quad \textit{as } \delta \to 0.$$

3. *Under Hölder-type source conditions and assuming the rank of A^*A is finite*

(3.30)
$$\|f_\delta^\lambda - f^\dagger\|_{\mathcal{H}_1} \in \begin{cases} O\left(\delta^{\frac{2r}{2r+1}}\right), & \gamma \geq r+\frac{1}{2} \quad \textit{and} \quad \lambda \in \Theta\left(\delta^{\frac{2}{2r+1}}\right) \\ O\left(\delta^{\frac{r}{2r+1-\gamma}}\right), & \gamma < r+\frac{1}{2} \quad \textit{and} \quad \lambda \in \Theta\left(\delta^{\frac{1}{2r+1-\gamma}}\right) \end{cases} \quad \textit{as } \delta \to 0.$$

In all three cases, if γ is sufficiently large, the upper rates are independent of γ. Otherwise, they are bounded from below by the γ-independent upper rates. We remark that in the first two cases (3.28) and (3.29), the upper rates are always slower than the classical optimal one, i.e., $O(\delta^{\frac{2r}{2r+1}})$ [12]. This rate can be achieved only in the case that A^*A has finite rank (3.30). The same remarks apply when considering the general source condition (see Corollary 3.10): the rate $\Phi(\Psi^{-1}(\delta))$ in the inverse problems setting [18] can be achieved only in the case that A^*A has finite rank. Here we point out that the conversion procedure involves only the tight inequality in Lemma 5.1 based on the Jensen inequality and therefore, it yields the best possible upper rates under the statistical noise assumption [5]. The reason the converted rates are not as good as the classical ones (see Table 2.1) is due to the difference between the assumptions on which rates rely: the statistical noise assumption turns out to be weaker than the deterministic one, as the converted upper rates are slower than $\Phi(\Psi^{-1}(\delta))$. Further comments on this point are provided at the end of section 3.

3.3.2. Lower rates. Now we exploit Theorem 3.4 to convert the lower rate of the spectral regularized solution f_δ^λ in a lower rate depending on n for the estimator $\hat{f}_n^\lambda$. As shown in Table 2.2, the lower rate of f_δ^λ depends only on the source condition (2.19) and it is independent of the eigenvalue decay of A^*A. Therefore, under assumption (2.19) we have the following.

COROLLARY 3.12. *Consider the spectral regularized solution $f_\delta^{\lambda_\delta}$ and its lower rate $\Omega(\Phi(\Psi^{-1}(\delta)))$, with $\Psi(t) = \sqrt{t}\Phi(t)$ and $\lambda_\delta \in \Theta(\Psi^{-1}(\delta))$ as $\delta \to 0$. Let c and d be functions defined as in Theorem 3.4. Then, for the hybrid estimator $\hat{f}_n^\lambda$ we have the following:*

- *If $\epsilon(\lambda_\delta) \in O(\delta)$ as $\delta \to 0$, then*
(3.31)
$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}) \in \Omega\left(\Phi(\Psi^{-1}(n^{-\frac{1}{2}}))\right), \ \textit{with } \lambda \in \Theta\left(\Psi^{-1}(n^{-\frac{1}{2}})\right) \ \textit{as } n \to \infty.$$

- *If $\epsilon(\lambda_\delta) \in \Omega(\delta)$ as $\delta \to 0$, then*
(3.32)
$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}) \in \Omega\left(\Phi(\Psi^{-1}(c(n^{-1})))\right), \ \textit{with } \lambda \in \Theta\left(\Psi^{-1}(d^{-1}(n^{-1}))\right) \ \textit{as } n \to \infty.$$

Now we apply the conversion procedure to the case of Hölder-type source conditions.

COROLLARY 3.13. *Under the Hölder-type source condition the lower rate of the spectral regularized solution* f_δ^λ *is given by* $\Omega(\delta^{\frac{2r}{2r+1}})$. *Then, thanks to Corollary* 3.6 *for the hybrid estimator* $\hat{f}_n^\lambda$, *we have*

(3.33)
$$\mathbb{E}_{\rho^{\otimes n}}(\|\hat{f}_n^\lambda - f^\dagger\|_{\mathcal{H}_1}) \in \begin{cases} \Omega\left(\left(\frac{1}{n}\right)^{\frac{r}{2r+1}}\right), & \gamma \geq r+\frac{1}{2} \text{ and } \lambda \in \Theta\left(\left(\frac{1}{n}\right)^{\frac{1}{2r+1}}\right) \\ \Omega\left(\left(\frac{1}{n}\right)^{\frac{r}{r+\gamma+\frac{1}{2}}}\right), & \gamma < r+\frac{1}{2} \text{ and } \lambda \in \Theta\left(\left(\frac{1}{n}\right)^{\frac{1}{r+\gamma+\frac{1}{2}}}\right) \end{cases} \quad \text{as } n \to \infty.$$

We remark that rates obtained in Corollaries 3.12 and 3.13 are lower bounds of the classical lower rates of the spectral regularized estimator $\hat{f}_{n,\mathrm{learn}}^\lambda$. It is worth noting that when $\gamma \geq r + \frac{1}{2}$ by converting the lower rate $\delta^{\frac{2r}{2r+1}}$ obtained under the deterministic noise assumption, we get $(\frac{1}{n})^{\frac{r}{2r+1}}$ which is equal to the best possible lower rate obtained under the statistical noise and the polynomial eigenvalue decay assumptions, i.e., it is equal to $\inf_{b>0}\left(\frac{1}{n}\right)^{\frac{r}{2r+1+\frac{1}{b}}}$.

Results in Corollaries 3.10–3.13 involve the behavior of $\epsilon(\lambda)$ which is the rate between the noise-free term of the regularization error $\|f^\lambda - f^\dagger\|_{\mathcal{H}_1}$ and the Hilbert–Schmidt norm of the spectral operator $\|s_\lambda(A^*A)A^*\|_{HS}$ (see (3.2)). Upper bounds of the noise-free term which goes to zero with decreasing λ are called profile functions [18]. The behavior of profile functions depends on the smoothness of the solution, and under general source conditions we have

$$\|f^\lambda - f^\dagger\| \leq RC_s\Phi(\lambda), \tag{3.34}$$

where C_s is a constant defined in (5.45). For more details see section 5 of the present work and [18]. From (3.34) and the definition of $\epsilon(\lambda)$ we have

$$\epsilon(\lambda) = \frac{\|f^\lambda - f^\dagger\|}{\|s_\lambda(A^*A)A^*\|_{HS}} \leq \frac{RC_s\Phi(\lambda)}{\|s_\lambda(A^*A)A^*\|}. \tag{3.35}$$

The term $\|s_\lambda(A^*A)A^*\|$ depends on the choice of the regularization function s_λ, and for the three classical examples of the spectral regularization (Tikhonov regularization, spectral cut-off, and Landweber regularization) it can be estimated as follows.

- Tikhonov regularization: in this case $s_\lambda(t) = (t+\lambda)^{-1}$. We have

$$\|s_\lambda(A^*A)A^*\| = \sup_{t \in \tau(A^*A)} \frac{\sqrt{t}}{\lambda + t} = \frac{1}{2\sqrt{\lambda}}. \tag{3.36}$$

  Therefore, under general source conditions $\epsilon(\lambda) \leq \frac{RC_s\Phi(\lambda)\sqrt{\lambda}}{2}$.
- Truncated singular value decomposition: in this case

$$s_\lambda(t) = \begin{cases} \frac{1}{t} & \text{if } t \geq \lambda, \\ 0 & \text{if } t < \lambda. \end{cases} \tag{3.37}$$

  We have

$$\|s_\lambda(A^*A)A^*\| = \sup_{t \geq \lambda} \frac{1}{\sqrt{t}} = \frac{1}{\sqrt{\lambda}}. \tag{3.38}$$

  Therefore, under general source conditions $\epsilon(\lambda) \leq RC_s\Phi(\lambda)\sqrt{\lambda}$.

- Landweber iteration: in this case the iteration k plays the role of the regularization parameter ($k = \lfloor \frac{1}{\lambda} \rfloor \in \mathbb{Z}_+$), and therefore, $s_{\frac{1}{k}}(t) = \sum_{j=0}^{k-1}(1-\varsigma t)^j$, with $\varsigma \in (0, \|A^*A\|^{-1})$. We have

(3.39)
$$\|s_{\frac{1}{k}}(A^*A)A^*\| = \sup_{t\in\sigma(A^*A)} \sum_{j=0}^{k-1}(1-\varsigma t)^j\sqrt{t}$$
$$= \sup_{t'\in(0,1]} \frac{1}{\sqrt{\varsigma}} \frac{1-(1-t')^k}{\sqrt{t'}} \geq \frac{\sqrt{k}}{\sqrt{\varsigma}}(1-(1-k^{-1})^k) \geq \frac{\sqrt{k}}{2\sqrt{\varsigma}},$$

having substituted $t' = \varsigma t$, taken $t' = 1/k$ to obtain the lower bound, and used the reverse Bernoulli inequality as follows: $(1-\frac{h}{k})^k \leq 1-(1-\frac{1}{e})h \leq 1-\frac{1}{2}h$ for all $h \in [0,1]$. Therefore, under general source conditions $\epsilon(\lambda) \leq 2RC_s\sqrt{\varsigma}\Phi(\lambda)\sqrt{\lambda}$.

Under the general source condition for the three above-mentioned cases, we have $\epsilon(\lambda) \in O(\sqrt{\lambda}\Phi(\lambda))$ as $\lambda \to 0$. Moreover, in the case of the Hölder-type source condition, we have $\epsilon(\lambda) \in O(\lambda^{r+\frac{1}{2}})$ as $\lambda \to 0$; i.e., the parameter γ in Corollaries 3.11 and 3.13 satisfies $\gamma \geq r + \frac{1}{2}$. We show in Table 3.1 a summary of the conversion of upper rates and lower rates under the Hölder-type source condition in order to show explicit computations.

TABLE 3.1
Conversion of convergence rates in the case $\gamma = r + \frac{1}{2}$.

	Upper rates (from n to δ)			
Assumption	Hybrid	λ_n	λ_δ	Spectral reg.
–	$\left(\frac{1}{n}\right)^{\frac{r}{2r+2}}$	$\left(\frac{1}{n}\right)^{\frac{1}{2r+2}}$	$\delta^{\frac{1}{r+\frac{3}{2}}}$	$\delta^{\frac{r}{r+\frac{3}{2}}}$
Eigenvalue decay (2.20)	$\left(\frac{1}{n}\right)^{\frac{r}{2r+1+\frac{1}{b}}}$	$\left(\frac{1}{n}\right)^{\frac{1}{2r+1+\frac{1}{b}}}$	$\delta^{\frac{1}{r+\frac{1}{2}+\frac{1}{b}}}$	$\delta^{\frac{r}{r+\frac{1}{2}+\frac{1}{b}}}$
Finite rank	$\left(\frac{1}{n}\right)^{\frac{r}{2r+1}}$	$\left(\frac{1}{n}\right)^{\frac{1}{2r+1}}$	$\delta^{\frac{2}{2r+1}}$	$\delta^{\frac{2r}{2r+1}}$
	Lower rates (from δ to n)			
Assumption	Spectral reg.	λ_δ	λ_n	Hybrid
–	$\delta^{\frac{2r}{2r+1}}$	$\delta^{\frac{2}{2r+1}}$	$\left(\frac{1}{n}\right)^{\frac{1}{2r+1}}$	$\left(\frac{1}{n}\right)^{\frac{r}{2r+1}}$

We remark that, in the finite rank hypothesis, the rates of the hybrid estimator and the spectral regularization match each other and, in this case, the number of samples n turns out to be inversely proportional to the noise level δ^2. However, this is not true if the rank of A^*A is not finite: in such a case, convergence rates given in statistical learning are weaker than the ones given for ill-posed deterministic inverse problems. Indeed, the conversion of statistical learning rates yields slower rates than the classical $\delta^{\frac{2r}{2r+1}}$ resulting from the inverse problems theory [12]. The fact that learning rates are generally slower should not be surprising: as the noise level δ goes to zero, the assumption $\|y - y^\delta\|_{\mathcal{H}_2} \leq \delta$ implies that there exists a subsequence of noisy data that converges to the exact data y on the set $\mathcal{X}$ almost everywhere; by contrast, taking the set of samples $\{(X_i, Y_i)\}_{i=1}^n$ as n goes to infinity is an assumption on the set $\{X_i\}_{i=1}^n \subset \mathcal{X}$, which is at most countable.

4. Conclusion. In this paper we provided a comparison between the convergence rates of spectral regularized methods in ill-posed inverse problems, where they are studied with respect to the noise level $\delta \to 0$, and the ones in statistical learning, where they are studied with respect to the number of samples $n \to \infty$. The comparison is based on a conversion procedure (Theorems 3.3 and 3.4) independent of source conditions, which can be applied under any smoothness conditions (not only general source conditions), provided that convergence rates for the hybrid estimator and $\epsilon(\lambda)$ (defined in (3.1) and (3.2), respectively) have been computed. In this paper, we performed the comparison by assuming in both contexts general source conditions, and an additional hypothesis on the eigenvalue decay of the operator A^*A in the statistical learning context. We quantified the difference between optimal rates resulting from the two frameworks by proving a result which allows us to convert upper rates depending on n to upper rates depending on δ and lower rates depending on δ to lower rates depending on n. The comparison led to the conclusion that the optimal rates of the spectral regularized learning estimator are slower than those of the spectral regularization in inverse problems: indeed, the hybrid estimator rates are the same as those of the spectral regularized learning, while they are slower than the classical ones when converted in the inverse problems framework. Moreover, the improvement of the rates of the spectral regularized learning estimator due to the addition of the hypothesis on the eigenvalue decay is not sufficient to reach the rates obtained in the inverse problems framework. Nonetheless, we showed that the finite rank hypothesis can make the spectral regularized learning estimator as fast as the spectral methods in inverse problems, which is the case for the Tikhonov regularization, spectral cut-off, and Landweber regularization.

5. Proofs. We prove Proposition 3.1; the main theorems of the paper, Theorems 3.3 and 3.4; Corollaries 3.5 and 3.6; and the upper convergence rates given in Lemmas 3.7, 3.8, and 3.9.

5.1. Proofs of Proposition 3.1; Theorems 3.3 and 3.4; and Corollaries 3.5 and 3.6. The first property of the hybrid estimator depends on the fact that it can be seen as an empirical version of the standard spectral regularization. To see this, we now introduce a linear regularization operator family L^λ as follows. We consider a positive and finite measure μ. We suppose that $\mathcal{H}_1$ is the Hilbert space of square integrable functions on a set $\mathcal{T}$ with respect to the measure μ, $L^2(\mathcal{T},\mu)$. We recall that $\mathcal{H}_2 = L^2(\mathcal{X},\nu)$. Let the linear regularization operator family $L^\lambda : \mathcal{H}_2 \to \mathcal{H}_1$, with $\lambda > 0$, be of the form

$$L^\lambda y = \int_{\mathcal{X}} \ell_x^\lambda \; y(x) \; d\nu(x), \tag{5.1}$$

where $\ell_x^\lambda \in \mathcal{H}_1$, $\ell_x^\lambda(t) := \ell^\lambda(x,t)$, and $\ell^\lambda(\cdot,t) \in \mathcal{H}_2$ for each $x \in \mathcal{X}$ and for each $t \in \mathcal{T}$. Thanks to this assumption the integral in (5.1) is finite. Moreover, we assume $\sup_{t\in\mathcal{T}} \|\ell^\lambda(\cdot,t)\|_{\mathcal{H}_2} < +\infty$ in such a way that L^λ is uniformly bounded and then is bounded in supremum norm for each $y \in \mathcal{H}_2$ $L^\lambda y$ which assures that $L^\lambda y \in \mathcal{H}_1$. We denote with F^λ the regularized solution given by the linear regularization operator L^λ applied to the noise-free data y, i.e.,

$$F^\lambda = L^\lambda y, \tag{5.2}$$

and with F_δ^λ the regularized solution given by the noisy data y^δ, i.e.,

$$F_\delta^\lambda = L^\lambda y^\delta, \tag{5.3}$$

when $\|y - y^\delta\|_{\mathcal{H}_2} \leq \delta$. We introduce the following empirical estimator:

$$\hat{F}_n^\lambda = L_{\mathbf{x}}^\lambda \mathbf{y} = \frac{1}{n}\sum_{i=1}^n \ell_{X_i}^\lambda \; Y_i, \tag{5.4}$$

where $\mathbf{x} = (X_1, \ldots, X_n)$ and $\mathbf{y} = (Y_1, \ldots, Y_n)$ denote the samples for which we consider the model assumptions in (2.15) and (2.16). The spectral regularization can be seen as a special case of linear regularization L^λ (5.3) by setting

$$\ell_x^\lambda = s_\lambda(A^*A)\phi_x, \tag{5.5}$$

with $x \in \mathcal{X}$. With this choice hypotheses on ℓ_x^λ are satisfied since $\sup_{t\in\mathcal{T}} \|\phi(\cdot, t)\|_{\mathcal{H}_2} < +\infty$, and we have

$$f_\delta^\lambda = F_\delta^\lambda \quad \text{and} \quad \hat{f}_n^\lambda = \hat{F}_n^\lambda. \tag{5.6}$$

We start by proving an inequality which will be used in the proof of Proposition 3.1. In what follows, to make the writing easier, we avoid the subscript of the norms and we denote with $\mathbb{E}$ the mean computed with respect to the measure $\rho^{\otimes n}$.

Lemma 5.1. *Let $\hat{F}_n^\lambda$ be defined in* (5.4)*. Under assumptions* (2.15) *and* (2.16) *we have*

$$\mathbb{E}(\|\hat{F}_n^\lambda - f^\dagger\|^2) \geq \frac{\sigma^2}{n}\|L^\lambda\|_{HS}^2 + \|F^\lambda - f^\dagger\|^2, \tag{5.7}$$

where $\|\cdot\|_{HS}$ denotes the Hilbert–Schmidt norm.

Proof. Denote with ϑ_n the difference between the estimate $\hat{F}_n^\lambda$ obtained with n samples and the sought solution $f^\dagger$. For any $t \in \mathcal{T}$ we have

$$\begin{aligned}\vartheta_n^2(t) &= \left(\frac{1}{n}\sum_{i=1}^n \ell_{X_i}^\lambda(t)Y_i - f^\dagger(t)\right)^2\\ &= \frac{1}{n^2}\sum_{i,j=1}^n \ell_{X_i}^\lambda(t)Y_i\ell_{X_j}^\lambda(t)Y_j - \frac{2}{n}f^\dagger(t)\sum_{i=1}^n \ell_{X_i}(t)Y_i + (f^\dagger(t))^2.\end{aligned} \tag{5.8}$$

By integrating over $\mathcal{Y}^n$, we get

$$\begin{aligned}\int_{\mathcal{Y}^n} \vartheta_n^2(t)\; d\rho(\cdot|\cdot)^{\otimes n} &= \frac{1}{n^2}\sum_{i=1}^n(\ell_{X_i}^\lambda(t))^2\sigma^2 + \frac{1}{n^2}\sum_{i,j=1}^n \ell_{X_i}^\lambda(t)\ell_{X_j}^\lambda(t)\; y(X_i)y(X_j)\\ &\quad -\frac{2}{n}\; f^\dagger(t)\sum_{i=1}^n \ell_{X_i}^\lambda(t)y(X_i) + (f^\dagger(t))^2,\end{aligned} \tag{5.9}$$

where $d\rho(\cdot|\cdot)^{\otimes n} = d\rho(Y_1|X_1)\cdots d\rho(Y_n|X_n)$ and by using that $\rho(\cdot|\cdot)$ is a probability measure on $\mathcal{Y}$. Then, by integrating over $\mathcal{X}^n$ we obtain

$$\begin{aligned}
\int_{\mathcal{X}^n}\int_{\mathcal{Y}^n} \vartheta_n^2(t)\ d\rho(\cdot|\cdot)^{\otimes n} d\nu^{\otimes n} &= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\int_{\mathcal{X}} (\ell^\lambda_{X_i}(t))^2 d\nu(X_i) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n^2-n}\left(\int_{\mathcal{X}} \ell^\lambda_{X_i}(t) y(X_i) d\nu(X_i)\right)^2 \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\int_{\mathcal{X}} \left(\ell^\lambda_{X_i}(t) y(X_i)\right)^2 d\nu(X_i) \\
&\quad - \frac{2}{n} f^\dagger(t)\sum_{i=1}^{n}\int_{\mathcal{X}} \ell^\lambda_{X_i}(t) y(X_i) d\nu(X_i) + (f^\dagger(t))^2 \\
&\geq \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\int_{\mathcal{X}} (\ell^\lambda_{X_i}(t))^2 d\nu(X_i) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n^2}\left(\int_{\mathcal{X}} \ell^\lambda_{X_i}(t) y(X_i) d\nu(X_i)\right)^2 \\
&\quad - \frac{2}{n} f^\dagger(t)\sum_{i=1}^{n}\int_{\mathcal{X}} \ell^\lambda_{X_i}(t) y(X_i) d\nu(X_i) + (f^\dagger(t))^2 \\
&= \frac{\sigma^2}{n}\int_{\mathcal{X}} (\ell^\lambda_{X}(t))^2 d\nu(X) \\
&\quad + \left(F^\lambda(t)\right)^2 - 2f^\dagger(t)F^\lambda(t) + (f^\dagger(t))^2, \qquad (5.10)
\end{aligned}$$

where we used that ν is a probability measure on $\mathcal{X}$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left(\|\hat{F}^\lambda_n - f^\dagger\|^2\right) &\geq \int_{\mathcal{T}} \frac{\sigma^2}{n}\int_{\mathcal{X}} (\ell^\lambda_X(t))^2\ d\nu(X) + \left(F^\lambda(t) - f^\dagger(t)\right)^2 d\mu(t) \\
&= \frac{\sigma^2}{n}\|L^\lambda\|^2_{HS} + \|F^\lambda - f^\dagger\|^2, \qquad (5.11)
\end{aligned}$$

as required. ∎

In the following we prove Proposition 3.1.

Proof. We start from the result of Lemma 5.1. Easy manipulation of formula (5.7) leads to

$$\sqrt{\mathbb{E}\left(\|\hat{F}^\lambda_n - f^\dagger\|^2\right)} \geq \Delta(n,\lambda)\|L^\lambda\|_{HS} + \|F^\lambda - f^\dagger\|, \qquad (5.12)$$

where $\Delta(n,\lambda)$ is defined in (3.3). For each $\delta > 0$, let y^δ be such that $\|y^\delta - y\| \leq \delta$; then a simple calculation gives

$$\|F^\lambda_\delta - f^\dagger\| \leq \delta\|L^\lambda\| + \|F^\lambda - f^\dagger\|. \qquad (5.13)$$

Further, for each $\delta \leq \Delta(n,\lambda)$ we have

$$\sqrt{\mathbb{E}\left(\|\hat{F}^\lambda_n - f^\dagger\|^2\right)} \geq \delta\|L^\lambda\| + \|F^\lambda - f^\dagger\| \qquad (5.14)$$

as $\|\cdot\|_{HS} \geq \|\cdot\|$. From (5.13) and (5.14) we obtain for all $\delta \leq \Delta(n,\lambda)$

$$\|F_\delta^\lambda - f^\dagger\|^2 \leq \mathbb{E}\left(\|\hat{F}_n^\lambda - f^\dagger\|^2\right) \tag{5.15}$$

for each y^δ such that $\|y^\delta - y\| \leq \delta$.

Conversely, let $\delta > 0$. For each $n \leq N(\delta,\lambda)$, with $N(\lambda,\delta)$ defined by (3.5), we have

$$\delta \leq \Delta(n,\lambda) \tag{5.16}$$

and so the thesis is proved. ∎

Functions $\Delta(n,\lambda)$ and $N(\delta,\lambda)$ express the dependency between the noisy level δ and the number of samples n. To make explicit this dependency we need to specify the rate of convergence of $\lambda \to 0$ considered as a function of both δ and n. For the sake of convenience, we introduce the following

NOTATION 5.2. *For any given λ_n we define*

$$\tilde{\delta}(n) := \Delta(n,\lambda_n), \tag{5.17}$$

where $\Delta(n,\lambda_n)$ is defined in (3.3). *Conversely, for any given λ_δ we define*

$$\tilde{n}(\delta) := \lfloor N(\delta,\lambda_\delta)\rfloor, \tag{5.18}$$

where the symbol $\lfloor\cdot\rfloor$ denotes the integer part and $N(\delta,\lambda_\delta)$ is defined in (3.5).

LEMMA 5.3. *Let τ and φ be two continuous strictly increasing functions $\tau,\varphi : (0,\infty) \to (0,\infty)$ such that $\tau(z) \to 0$ and $\varphi(z) \to 0$ as $z \to 0$.*

Let $\lambda_n \in \Theta(\tau(n^{-1}))$ as $n \to \infty$; then,

$$\tilde{\delta}(n) \in \begin{cases} \Theta(n^{-\frac{1}{2}}) & \text{if } \epsilon(\lambda_n) \in O(n^{-\frac{1}{2}}),\\ \Theta(c(n^{-1})) & \text{if } \epsilon(\lambda_n) \in \Omega(n^{-\frac{1}{2}}),\end{cases} \tag{5.19}$$

where $c : z \in (0,\infty) \mapsto c(z) \in (0,\infty)$ is a continuous strictly increasing function with respect to the variable z and c is such that $c(n^{-1}) \in \Theta(\frac{1}{n\epsilon(\lambda_n)})$ as $n \to \infty$.

Let $\lambda_\delta \in \Theta(\varphi(\delta))$ as $\delta \to 0$; then

$$\tilde{n}(\delta) \in \begin{cases} \Theta(\delta^{-2}) & \text{if } \epsilon(\lambda_\delta) \in O(\delta),\\ \Theta(d(\delta)) & \text{if } \epsilon(\lambda_\delta) \in \Omega(\delta),\end{cases} \tag{5.20}$$

where $d : z \in (0,\infty) \mapsto d(z) \in (0,\infty)$ is a continuous strictly decreasing function with respect to the variable z and d is such that $d(\delta) \in \Theta(\frac{1}{\delta\epsilon(\lambda_\delta)})$ as $\delta \to 0$.

Proof. We start by proving (5.20). If $\epsilon(\lambda_n) \in O(n^{-\frac{1}{2}})$, then there exist $M_1 > 0$ and $n_0 > 0$ such that $\epsilon(\lambda_n) \leq M_1 n^{-\frac{1}{2}}$ for each $n > n_0$. Then, for each $n > n_0$,

$$D(n,\lambda_n) := \sqrt{\frac{\sigma^2}{n} + \epsilon(\lambda_n)^2} + \epsilon(\lambda_n) \leq \frac{\left(\sqrt{\sigma^2 + M_1^2} + M_1\right)}{\sqrt{n}}, \tag{5.21}$$

i.e., $D(n,\lambda_n) \in O(n^{-\frac{1}{2}})$. Furthermore, by noticing that

$$D(n,\lambda_n) \geq \sqrt{\frac{\sigma^2}{n}}, \tag{5.22}$$

we have $D(n,\lambda_n) \in \Theta(n^{-\frac{1}{2}})$. Therefore, for the definition of $\tilde{\delta}(n)$ in (5.17), we obtain

$$\tilde{\delta}(n) = \frac{\sigma^2}{nD(n,\lambda_n)} \in \Theta(n^{-\frac{1}{2}}). \tag{5.23}$$

If $\epsilon(\lambda_n) \in \Omega(n^{-\frac{1}{2}})$, then there exist $M_2 > 0$ and $n_0 > 0$ such that $\epsilon(\lambda_n) \geq M_2 n^{-\frac{1}{2}}$ for each $n > n_0$. Then, for each $n > n_0$ we have $n\epsilon(\lambda_n)^2 \geq M_2^2$; therefore

$$D(n,\lambda_n) = \left(\sqrt{\frac{\sigma^2}{n\epsilon(\lambda_n)^2} + 1} + 1\right)\epsilon(\lambda_n) \leq \left(\sqrt{\frac{\sigma^2}{M_2^2} + 1} + 1\right)\epsilon(\lambda_n), \tag{5.24}$$

i.e., $D(n,\lambda_n) \in O(\epsilon(\lambda_n))$. Furthermore, by noticing that

$$D(n,\lambda_n) \geq \epsilon(\lambda_n), \tag{5.25}$$

we have $D(n,\lambda_n) \in \Theta(\epsilon(\lambda_n))$. Therefore, we obtain

$$\tilde{\delta}(n) = \frac{\sigma^2}{nD(n,\lambda_n)} \in \Theta\left(\frac{1}{n\epsilon(\lambda_n)}\right). \tag{5.26}$$

In particular, taking $c : z \in (0,\infty) \to c(z) \in (0,\infty)$ a continuous strictly increasing function with respect to z such that $c(n^{-1}) \in \Theta(\frac{1}{n\epsilon(\lambda_n)})$, we can write that $\tilde{\delta}(n) \in \Theta(c(n^{-1}))$.

Now we prove (5.20). If $\epsilon(\lambda_\delta) \in O(\delta)$, then there exist $P_1 > 0$ and $\delta_0 > 0$ such that $\epsilon(\lambda_\delta) \leq P_1\delta$ for each $0 < \delta < \delta_0$. Therefore, for each $0 < \delta < \delta_0$,

$$\delta^2 + 2\delta\epsilon(\lambda_\delta) \leq (1 + 2P_1)\delta^2, \tag{5.27}$$

i.e., $\delta^2 + 2\delta\epsilon(\lambda_\delta) \in O(\delta^2)$. Furthermore, since

$$\delta^2 + 2\delta\epsilon(\lambda_\delta) \geq \delta^2, \tag{5.28}$$

then $\delta^2 + 2\delta\epsilon(\lambda_\delta) \in \Theta(\delta^2)$. Therefore, $N(\delta,\lambda_\delta) \in \Theta(\delta^{-2})$. We conclude that $\tilde{n}(\delta) \in \Theta(\delta^{-2})$, since $N(\delta,\lambda_\delta) - 1 \leq \tilde{n}(\delta) \leq N(\delta,\lambda_\delta)$ and $N(\delta,\lambda_\delta) \in \Theta(\delta^{-2})$, as $\delta \to 0$.

If $\epsilon(\lambda_\delta) \in \Omega(\delta)$, then there exist $P_2 > 0$ and $\delta_0 > 0$ such that $\epsilon(\lambda_\delta) \geq P_2\delta$ for each $0 < \delta < \delta_0$. Therefore, for each $0 < \delta < \delta_0$, we have $\frac{\delta}{\epsilon(\lambda_\delta)} \leq \frac{1}{P_2}$, and

$$\delta^2 + 2\delta\epsilon(\lambda_\delta) = \left(\frac{\delta}{\epsilon(\lambda_\delta)} + 2\right)\delta\epsilon(\lambda_\delta) \leq \left(\frac{1}{P_2} + 2\right)\delta\epsilon(\lambda_\delta), \tag{5.29}$$

i.e., $\delta^2 + 2\delta\epsilon(\lambda_\delta) \in O(\delta\epsilon(\lambda_\delta))$. Furthermore, since

$$\delta^2 + 2\delta\epsilon(\lambda_\delta) \geq \delta\epsilon(\lambda_\delta), \tag{5.30}$$

then $\delta^2 + 2\delta\epsilon(\lambda_\delta) \in \Theta(\delta\epsilon(\lambda_\delta))$. Therefore, $\tilde{n}(\delta) \in \Theta(\frac{1}{\delta\epsilon(\lambda_\delta)})$, since $N(\delta,\lambda_\delta) - 1 \leq \tilde{n}(\delta) \leq N(\delta,\lambda_\delta)$, $N(\delta,\lambda_\delta) \in \Theta(\frac{1}{\delta\epsilon(\lambda_\delta)})$, and $\delta\epsilon(\lambda_\delta) \to 0$ as $\delta \to 0$. In particular, taking $d : z \in (0,\infty) \to d(z) \in (0,\infty)$ a continuous strictly decreasing function with respect to z such that $d(\delta) \in \Theta(\frac{1}{n\epsilon(\lambda_n)})$, we can write that $\tilde{n}(\delta) \in \Theta(d(\delta))$. ∎

In the case that the convergence rates have a polynomial form, results in Lemma 5.3 simplify as follows.

LEMMA 5.4. *Let $\varepsilon(\lambda) \in \Theta(\lambda^\gamma)$ as $\lambda \to 0$ with $\gamma \geq 0$. If $\lambda_n \in \Theta(n^{-p})$ as $n \to \infty$, with $p > 0$, then*

$$\tilde{\delta}(n) \in \Theta\left(n^{-\max(\frac{1}{2},1-p\gamma)}\right) \quad \text{as } n \to \infty\,. \tag{5.31}$$

If $\lambda_\delta \in \Theta(\delta^{p^*})$ as $\delta \to 0$ with $p^* > 0$, then

$$\tilde{n}(\delta) \in \Theta\left(\delta^{-\min(2,p^*\gamma+1)}\right) \quad \text{as } \delta \to 0\,. \tag{5.32}$$

Proof. The proof follows an argument similar to the proof of Lemma 5.3. In particular, (5.31) follows from the definition of $\tilde{\delta}$ in (5.17) and from hypotheses $\lambda_n \in \Theta(n^{-p})$ as $n \to \infty$ and $\varepsilon(\lambda) \in \Theta(\lambda^\gamma)$ as $\lambda \to 0$. In the same way (5.32) follows from the definition of $\tilde{n}$ in (5.18) and from hypotheses $\lambda_\delta \in \Theta(\delta^{p^*})$ as $\delta \to 0$ and $\varepsilon(\lambda) \in \Theta(\lambda^\gamma)$ as $\lambda \to 0$. ∎

LEMMA 5.5. *Given λ_n there exists a unique λ_δ such that*

$$\tilde{\delta} \circ \tilde{n} = id_{\Im(\tilde{\delta})}, \tag{5.33}$$

where $id_{\Im(\tilde{\delta})}$ indicates the identity on the set $\Im(\tilde{\delta}) = \{\delta > 0 \mid \frac{\sigma^2}{\delta^2+2\delta\varepsilon(\lambda_\delta)} \in \mathbb{N}\}$ and

$$\Lambda^n = \Lambda^\delta \circ \tilde{\delta}, \tag{5.34}$$

where $\Lambda^n : \mathbb{N} \to \mathbb{R}$ and $\Lambda^\delta : \mathbb{R} \to \mathbb{R}$ are such that $\lambda_n = \Lambda^n(n)$ and $\lambda_\delta = \Lambda^\delta(\delta)$. Furthermore,

$$\tilde{n} \circ \tilde{\delta} = id_{\mathbb{N}}. \tag{5.35}$$

Proof. The existence and uniqueness of a sequence λ_δ, such that (5.33) and (5.34) are verified, follow by defining $\lambda_\delta := \Lambda^n(\tilde{n}(\delta))$. With straightforward calculus it can be verified that (5.34) implies (5.35). ∎

Similarly, we give the converse result.

LEMMA 5.6. *Given λ_δ, there exists a unique λ_n such that*

$$\tilde{n} \circ \tilde{\delta} = id_{\mathbb{N}}$$

and

$$\Lambda^\delta = \Lambda^n \circ \tilde{n} \tag{5.36}$$

where we have used the same notation as in Lemma 5.5. Furthermore,

$$\tilde{\delta} \circ \tilde{n} = id_{\Im(\tilde{\delta})}. \tag{5.37}$$

The proof is analogous to that of Lemma 5.5 by defining $\lambda_n = \Lambda^\delta(\tilde{\delta}(n))$.

Now we prove the main theorems. In the following we prove the result in Theorem 3.3.

Proof. Given $\lambda_n = \Lambda^n(n)$, we define $\lambda_\delta = \Lambda^\delta(\delta)$ according to Lemma 5.5, so that (5.33) and (5.34) hold. The rate of λ_δ given in (3.16) can be found by using the hypothesis $\lambda_n \in \Theta(\tau(n^{-1}))$ and Lemma 5.3. Now we prove (3.15). Thanks to

Proposition 3.1 and Lemma 5.3, for each $\lambda > 0$ and $\delta > 0$ there exists $\tilde{n}(\delta)$ such that for all $n \leq \tilde{n}(\delta)$

$$\|F_\delta^\lambda - f^\dagger\|^2 \leq \mathbb{E}(\|\hat{F}_n^\lambda - f^\dagger\|^2). \tag{5.38}$$

Let $n = \tilde{n}(\delta)$, then

$$\|F_\delta^\lambda - f^\dagger\|^2 \leq \mathbb{E}(\|\hat{F}_{\tilde{n}(\delta)}^\lambda - f^\dagger\|^2). \tag{5.39}$$

Let $\lambda = \lambda_\delta$. Then there exist $n_0 \in \mathbb{N}$ and $M > 0$ such that

$$\|F_\delta^{\lambda_\delta} - f^\dagger\|^2 \leq \mathbb{E}(\|\hat{F}_{\tilde{n}(\delta)}^{\Lambda^\delta(\delta)} - f^\dagger\|^2) = \mathbb{E}(\|\hat{F}_{\tilde{n}(\delta)}^{\Lambda^n(\tilde{n}(\delta))} - f^\dagger\|^2) \leq M\alpha\left((\tilde{n}(\delta))^{-1}\right), \tag{5.40}$$

for all $\tilde{n}(\delta) > n_0$. From (3.8) and (5.40), and by using Lemma 5.3 we obtain that

- if $\epsilon(\lambda_n) \in O(n^{-\frac{1}{2}})$, then $\lambda_\delta \in \Theta(\tau(\delta^2))$, and therefore from Proposition 3.1 we have $\tilde{n}(\delta) \in \Theta(\delta^{-2})$ and from (5.40) we obtain $\|F_\delta^\lambda - f^\dagger\|^2 \in O(\alpha(\delta^2))$;
- if $\epsilon(\lambda_n) \in \Omega(n^{-\frac{1}{2}})$, then $\lambda_\delta \in \Theta(\tau(c^{-1}(\delta)))$, and therefore from Proposition 3.1 we have $\tilde{n}(\delta) \in \Theta(d(\delta))$ and from (5.40) we obtain $\|F_\delta^\lambda - f^\dagger\|^2 \in O(\alpha(\frac{1}{d(\delta)}))$.

This completes the proof. ∎

We give the proof of the result in Theorem 3.4.

Proof. The proof exploits an argument similar to the one used for Theorem 3.3. Given $\lambda_\delta = \Lambda^\delta(\delta)$, by defining $\lambda_n = \Lambda^n(n)$ according to Lemma 5.6, it can be proved that the rate of λ_n is given by (3.12). To prove (3.11) one has to reverse the role of n and δ in the proof of Theorem 3.3 and use Proposition 3.1 and hypothesis (3.10). In this way one obtains that for each $n \in \mathbb{N}$, there exist $\delta_0 > 0$ and $M' > 0$ such that

$$\mathbb{E}(\|\hat{F}_n^{\lambda_n} - f^\dagger\|^2) \geq \|F_{\tilde{\delta}(n)}^{\Lambda^\delta(\tilde{\delta}(n))} - f^\dagger\|^2 \geq M'\alpha(\tilde{\delta}(n)) \tag{5.41}$$

for all $\tilde{\delta}(n) < \delta_0$. The thesis follows from (3.12) and (5.41) and Lemma 5.3. ∎

In the following we prove the result in Corollary 3.5.

Proof. Given $\lambda_n = \Lambda^n(n)$, we define $\lambda_\delta = \Lambda^\delta(\delta)$ according to Lemma 5.5, so that (5.33) and (5.34) hold. The rate of λ_δ given in (3.16) can be found by using the hypothesis $\lambda_n = \Theta(n^{-p})$ and Lemma 5.4. Now we prove (3.15). Thanks to Proposition 3.1, Lemma 5.4, and repeating the same argument as in the proof of Theorem 3.3 we have the following.

Let $\lambda = \lambda_\delta$. Then there exist $n_0 \in \mathbb{N}$ and $M > 0$ such that

$$\|F_\delta^{\lambda_\delta} - f^\dagger\|^2 \leq \mathbb{E}(\|\hat{F}_{\tilde{n}(\delta)}^{\Lambda^\delta(\delta)} - f^\dagger\|^2) \leq M\left(\frac{1}{\tilde{n}(\delta)}\right)^\alpha, \tag{5.42}$$

for all $\tilde{n}(\delta) > n_0$. From (3.16) and (5.42), and by using Lemma 5.4 we obtain that

- if $p\gamma \geq \frac{1}{2}$, then $\lambda_\delta = \Theta(\delta^{2p})$, and therefore from Proposition 3.1 we have $\tilde{n}(\delta) \in \Theta(\delta^2)$ and from (5.42) we obtain $\|F_\delta^\lambda - f^\dagger\|^2 \in O(\delta^{2\alpha})$;
- if $p\gamma < \frac{1}{2}$, then $\lambda_\delta \in \Theta(\delta^{\frac{p}{1-p\gamma}})$, and therefore from Proposition 3.1 we have $\tilde{n}(\delta) \in \Theta(\delta^{\frac{1}{1-p\gamma}})$ and from (5.42) we obtain $\|F_\delta^\lambda - f^\dagger\|^2 \in O(\delta^{\frac{\alpha}{1-p\gamma}})$.

This completes the proof. ∎

We give the proof of the result in Corollary 3.6.

Proof. The proof exploits an argument similar to the one used for Corollary 3.5. Given $\lambda_\delta = \Lambda^\delta(\delta)$, by defining $\lambda_n = \Lambda^n(n)$ according to Lemma 5.6, it can be proved that the rate of λ_n is given by (3.19). To prove (3.18) one has to reverse the role of n and δ in the proof of Corollary 3.5 and use Proposition 3.1 and hypothesis (3.17). In this way one obtains that for each $n \in \mathbb{N}$, there exist $\delta_0 > 0$ and $M' > 0$ such that

$$\mathbb{E}(\|\hat{F}_n^{\lambda_n} - f^\dagger\|^2) \geq \|F_{\tilde{\delta}(n)}^{\Lambda^\delta(\tilde{\delta}(n))} - f^\dagger\|^2 \geq M'(\tilde{\delta}(n))^\alpha \tag{5.43}$$

for all $\tilde{\delta}(n) < \delta_0$. The thesis follows from (3.19) and (5.43) and Lemma 5.4. ◻

5.2. Proofs of upper rates of the hybrid estimator. We remark that $\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2)$ satisfies the bias-variance decomposition as follows:

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) = B(\hat{f}_\lambda)^2 + \mathbb{E}(\|\hat{f}_\lambda - \mathbb{E}(\hat{f}_\lambda)\|^2), \tag{5.44}$$

where $B(\hat{f}_\lambda) := \|\mathbb{E}(\hat{f}_n^\lambda) - f^\dagger\|$ is the bias term and $\mathbb{E}(\hat{f}_n^\lambda) = f^\lambda$. We remark that the property of regularization function s_λ in (2.11) represents the qualification of a method in terms of the Hölder-type source conditions [5]. Under general source conditions the following property is considered.

Definition 5.7. *There exists a constant C_s such that*

$$\sup_{t\in\sigma(A^*A)} |\Phi(t)(1 - ts_\lambda(t))| \leq C_s\Phi(\lambda). \tag{5.45}$$

We remark that the property (5.45) is satisfied when $C_s = C_q$, the constant in the property (2.11), provided that the function $t \mapsto \frac{t^q}{\Phi(t)}$ is nondecreasing. For details see [5] and references therein. Therefore, under the source condition (2.19) the bias term can be bounded by

$$B(\hat{f}_\lambda) \leq C_s R\Phi(\lambda). \tag{5.46}$$

We remark that, under the Hölder-type source condition,

$$B(\hat{f}_\lambda) \leq C_r R\lambda^r, \tag{5.47}$$

with $r \leq q$. The estimation of the variance term needs more manipulations. In detail, to bound the variance term we use the argument given in [5] where a more general mixed type noise model is considered and the stochastic part of the noise is modeled as an Hilbert-space process. We consider the model in (2.15) and (2.16). Therefore, by defining $\bar{\epsilon} = Y - Af^\dagger(X)$, we have that $\mathbb{E}(\bar{\epsilon}|X = x) = 0$ and $Var(\bar{\epsilon}|X = x) = \sigma^2$ for ν-almost $x \in \mathcal{X}$. We introduce the Hilbert-space noise process $\tilde{\epsilon}$, defined by

$$\langle \tilde{\epsilon}, \psi \rangle := \frac{1}{n}\sum_{i=1}^n Y_i\psi(X_i) - \langle Af^\dagger, \psi\rangle_{\mathcal{H}_2} \tag{5.48}$$

for $\psi \in \mathcal{H}_2$. Then, by remarking that

$$\langle A^*\tilde{\epsilon}, \varsigma\rangle_{\mathcal{H}_1} = \langle \tilde{\epsilon}, A\varsigma\rangle_{\mathcal{H}_2} = \frac{1}{n}\sum_{i=1}^n Y_i\langle \varsigma, \phi_{X_i}\rangle_{\mathcal{H}_1} - \langle A^*Af^\dagger, \varsigma\rangle_{\mathcal{H}_1} \tag{5.49}$$

$$= \Big\langle \frac{1}{n}\sum_{i=1}^n Y_i\phi_{X_i}, \varsigma\Big\rangle_{\mathcal{H}_1} - \langle A^*Af^\dagger, \varsigma\rangle_{\mathcal{H}_1} = \langle A_n^*\mathbf{y} - A^*Af^\dagger, \varsigma\rangle \tag{5.50}$$

for $\varsigma \in \mathcal{H}_1$, using (2.2), (2.5), and (2.6). Therefore, $A^*\tilde{\epsilon} = A_n^*\mathbf{y} - A^*Af^\dagger$. From [5] we have that the noise $\tilde{\tilde{\epsilon}} := \frac{\tilde{\epsilon}}{\tilde{\sigma}}$, where $\tilde{\sigma} := \frac{\sqrt{C}}{\sqrt{n}}$ and $C := \sigma + \|Af^\dagger\|^2_{L^\infty(\mathcal{X},\nu)} + \|Af^\dagger\|^2_{L^2(\mathcal{X},\nu)}$, is such that

$$\mathbb{E}(\|\hat{f}_n^\lambda - \mathbb{E}(\hat{f}_n^\lambda)\|^2) = \mathbb{E}(\|s_\lambda(A^*A)A^*\tilde{\sigma}\tilde{\tilde{\epsilon}}\|^2) \tag{5.51}$$

(cf. Theorem 3 and Proposition 8 in [5]). We remark that $C < \infty$, since assumption (2.1) holds. In the following we provide proofs of results in Lemmas 3.7–3.9. The proofs mainly consist of bounding the term in (5.51) in different ways according to the hypothesis on the eigenvalue decay. We start to prove Lemma 3.7.

Proof. From (5.51) we have

$$\begin{aligned}\mathbb{E}(\|\hat{f}_n^\lambda - \mathbb{E}(\hat{f}_n^\lambda)\|^2) &\le \frac{C}{n}\sum_{j:\mu_j\in\sigma(A^*A)} s_\lambda^2(\mu_j)\mu_j \\ &\le \frac{C}{n}\left(\sup_{\mu_j\in\sigma(A^*A)} s_\lambda^2(\mu_j)\right)\sum_{j:\mu_j\in\sigma(A^*A)}\mu_j \le \frac{C}{n}\frac{E^2}{\lambda^2}C',\end{aligned} \tag{5.52}$$

where we have used the property (2.10) of the regularization function s_λ and the fact that the operator A^*A is of trace class where C' represents a constant which bounds the trace norm of A^*A. Therefore, under assumption (2.19) we obtain

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) \le C_s^2(\Phi(\lambda))^2R^2 + \frac{CE^2C'}{n\lambda^2}. \tag{5.53}$$

By balancing terms in the right-hand side (r.h.s.) of (5.53) we have the thesis. Trivially, we notice that under the Hölder-type source condition the bound in (5.53) boils down to

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) \le C_r^2\lambda^{2r}R^2 + \frac{CE^2C'}{n\lambda^2}. \tag{5.54}$$ ◻

Now we prove the result in Lemma 3.8.

Proof. We consider the decay of the eigenvalues in (2.20). Then we have that

$$\#\{j : \mu_j \ge \lambda\} \in \Theta(S(\lambda)),$$

where $S(\lambda) = \lambda^{-\frac{1}{b}}$. Moreover, $S(\lambda)$ satisfies the following conditions (cf. equation (3.8) in [5]):

- $S(\lambda) \in C^2((0,\bar{\lambda}])$, for some $\bar{\lambda} > 0$.
- $S'(\lambda) = -\frac{1}{b}\lambda^{-\frac{1}{b}-1} < 0$.
- $\lambda S'(\lambda) = \frac{1}{b}\lambda^{-\frac{1}{b}}$ is integrable in $(0,\bar{\lambda}]$, since $b > 1$.
- $\lim_{\lambda\to 0}\lambda S(\lambda) = \lim_{\lambda\to 0}\lambda^{1-\frac{1}{b}} = 0$, since $b > 1$.
- There exists a constant $\gamma_S \in (0,2)$ such that $\frac{S''(\lambda)}{-S'(\lambda)} \le \frac{\gamma_S}{\lambda}$, since $\frac{S''(\lambda)}{-S'(\lambda)} = \frac{\frac{1}{b}+1}{\lambda}$, and therefore the property is satisfied taking $\gamma_S = \frac{1}{b} + 1 \in (1,2)$, since $b > 1$.

From Theorem 3 in [5] we have that S is such that

$$\mathbb{E}(\|\hat{f}_n^\lambda - \mathbb{E}(\hat{f}_n^\lambda)\|^2) = \mathbb{E}(\|s_\lambda(A^*A)A^*\tilde{\sigma}\tilde{\tilde{\epsilon}}\|^2) \in O\left(\frac{C}{n}L\frac{1}{\lambda^2}\int_0^\lambda S(\beta)d\beta\right) \in O\left(\frac{C}{n}L\frac{1}{\lambda^{1+\frac{1}{b}}}\right) \tag{5.55}$$

as $\lambda \to 0$, where L is a constant which depends on D and E (see properties (2.9) and (2.10)) and constants in the assumption (2.20). Therefore, under assumption (2.19) and from (5.55) we obtain that there exist $M > 0$ and $n_0 > 0$ such that for each $n > n_0$

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) \leq C_s^2(\Phi(\lambda))^2 R^2 + M\frac{C}{n}L\frac{1}{\lambda^{1+\frac{1}{b}}}. \tag{5.56}$$

By balancing terms in the r.h.s. of (5.56) we have the thesis. Trivially, if we consider the Hölder-type source condition, the bound in (5.57) is

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) \leq C_r^2\lambda^{2r} R^2 + M\frac{C}{n}L\frac{1}{\lambda^{1+\frac{1}{b}}} \tag{5.57}$$

for each $n > n_0$. $\square$

Finally, we give the proof of Lemma 3.9.

Proof. The variance term can be bounded as follows:

$$\mathbb{E}(\|\hat{f}_n^\lambda - \mathbb{E}(\hat{f}_n^\lambda)\|^2) \leq \frac{C}{n}\sum_{j:\mu_j\in\sigma(A^*A)} s_\lambda^2(\mu_j)\mu_j \tag{5.58}$$

$$\leq \frac{C}{n}\left(\sup_{\mu_j\in\sigma(A^*A)} s_\lambda(\mu_j)\right)\sum_{j:\mu_j\in\sigma(A^*A)} s_\lambda(\mu_j)\mu_j \tag{5.59}$$

$$\leq \frac{C}{n}F\frac{E}{\lambda}\left(\sup_{\mu_j\in\sigma(A^*A)} s_\lambda(\mu_j)\mu_j\right) \leq \frac{C}{n}\frac{E}{\lambda}FD \tag{5.60}$$

for properties (2.9) and (2.10) of the regularization function. Therefore,

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) \leq C_s^2(\Phi(\lambda))^2 R^2 + \frac{CEFD}{n\lambda}. \tag{5.61}$$

By balancing terms in the r.h.s. of (5.61) we obtain the thesis. Under the Hölder-type source condition the bound in (5.62) becomes the following:

$$\mathbb{E}(\|\hat{f}_n^\lambda - f^\dagger\|^2) \leq C_r^2\lambda^{2r} R^2 + \frac{CEFD}{n\lambda}. \tag{5.62}$$

$\square$

Acknowledgment. The authors are grateful to the two anonymous referees for their careful reading and valuable comments which have considerably improved the exposition of this paper.

REFERENCES

[1] J. Adler and O. Öktem, *Solving ill-posed inverse problems using iterative deep neural networks*, Inverse Problems, 33 (2017), 124007.

[2] V. Albani, P. Elbau, M. V. de Hoop, and O. Scherzer, *Optimal convergence rates results for linear inverse problems in Hilbert spaces*, Numer. Funct. Anal. Optim., 37 (2016), pp. 521–540.

[3] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[4] F. Bauer, S. Pereverzev, and L. Rosasco, *On regularization algorithms in learning theory*, J. Complexity, 23 (2007), pp. 52–72.

[5] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart, *Convergence rates of general regularization methods for statistical inverse problems and applications*, SIAM J. Numer. Anal., 45 (2007), pp. 2610–2636, https://doi.org/10.1137/060651884.

[6] G. Blanchard and N. Mücke, *Optimal rates for regularization of statistical inverse learning problems*, Found. Comput. Math., 18 (2018), pp. 971–1013.

[7] A. Böttcher, B. Hofmann, U. Tautenhahn, and M. Yamamoto, *Convergence rates for Tikhonov regularization from different kinds of smoothness conditions*, Appl. Anal., 85 (2006), pp. 555–578.

[8] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math., 7 (2007), pp. 331–368.

[9] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[10] E. De Vito, L. Rosasco, and A. Caponnetto, *Discretization error analysis for Tikhonov regularization*, Anal. Appl. (Singap.), 4 (2006), pp. 81–99.

[11] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone, *Learning from examples as an inverse problem*, J. Mach. Learn. Res., 6 (2005), pp. 883–904.

[12] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Springer, Dordrecht, Boston, 1996.

[13] J. Flemming, *Generalized Tikhonov Regularization and Modern Convergence Rate Theory in Banach Spaces*, Ber. Math., Shaker, 2012.

[14] F. Girosi, M. Jones, and T. Poggio, *Regularization theory and neural networks architectures*, Neural Comput., 7 (1995), pp. 219–269.

[15] S. Guastavino and F. Benvenuto, *A consistent and numerically efficient variable selection method for sparse Poisson regression with applications to learning and signal recovery*, Stat. Comput., 29 (2019), pp. 501–516.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Ser. Statist., Springer-Verlag, New York, 2001.

[17] B. Hofmann, S. Kindermann, and P. Mathé, *Penalty-based smoothness conditions in convex variational regularization*, J. Inverse Ill-Posed Probl., 27 (2019), pp. 283–300.

[18] B. Hofmann and P. Mathé, *Analysis of profile functions for general linear regularization methods*, SIAM J. Numer. Anal., 45 (2007), pp. 1122–1141, https://doi.org/10.1137/060654530.

[19] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer-Verlag, New York, 2005.

[20] R. A. Kress, *Linear Integral Equations*, Springer, New York, 2014.

[21] J. Lin and V. Cevher, *Optimal Rates for Spectral-regularized Algorithms with Least-Squares Regression over Hilbert Spaces*, preprint, https://arxiv.org/abs/1801.06720v1, 2018.

[22] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri, *Spectral algorithms for supervised learning*, Neural Comput., 20 (2008), pp. 1873–1897.

[23] S. Lu and S. V. Pereverzev, *Regularization Theory for Ill-posed Problems: Selected Topics*, Inverse Ill-posed Probl. Ser. 58, De Gruyter, 2013.

[24] B. A. Mair and F. H. Ruymgaart, *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math., 56 (1996), pp. 1424–1444, https://doi.org/10.1137/S0036139994264476.

[25] P. Mathé and B. Hofmann, *How general are general source conditions?*, Inverse Problems, 24 (2008), 015009.

[26] P. Mathé and S. Pereverzev, *Discretization strategy for linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 1263–1277.

[27] P. Mathé and S. Pereverzev, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.

[28] P. Mathé and S. Pereverzev, *Regularization of some linear ill-posed problems with discretized random noisy data*, Math. Comp., 75 (2006), pp. 1913–1929.

[29] M. T. McCann, K. H. Jin, and M. Unser, *Convolutional neural networks for inverse problems in imaging: A review*, IEEE Signal Process. Mag., 34 (2017), pp. 85–95.

[30] S. Mendelson and J. Neeman, *Regularization in kernel learning*, Ann. Statist., 38 (2010), pp. 526–565.

[31] V. I. Norkin and M. A. Keyzer, *On convergence of kernel learning estimators*, SIAM J. Optim., 20 (2009), pp. 1205–1223, https://doi.org/10.1137/070696817.

[32] A. Rastogi and S. Sampath, *Optimal rates for the regularized learning algorithms under general source condition*, Front. Appl. Math. Stat., 3 (2017), 3.

[33] S. Smale and D.-X. Zhou, *Learning theory estimates via integral operators and their approximations*, Constr. Approx., 26 (2007), pp. 153–172.

[34] A. J. Smola, B. Schölkopf, and K.-R. Müller, *The connection between regularization operators and support vector kernels*, Neural Netw., 11 (1998), pp. 637–649.
[35] B. Sprung and T. Hohage, *Higher order convergence rates for Bregman iterated variational regularization of inverse problems*, Numer. Math., 141 (2019), pp. 215–252.
[36] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Science & Business Media, 2008.
[37] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2013.
[38] C. R. Vogel, *Computational Methods for Inverse Problems*, Frontiers Appl. Math. 23, SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898717570.
[39] C. Wang and D.-X. Zhou, *Optimal learning rates for least squares regularized regression with unbounded sampling*, J. Complexity, 27 (2011), pp. 55–67.
[40] F. Werner and B. Hofmann, *Convergence analysis of (statistical) inverse problems under conditional stability estimates*, Inverse Problems, 36 (2019), 015004.
[41] Y. Yao, L. Rosasco, and A. Caponnetto, *On early stopping in gradient descent learning*, Constr. Approx., 26 (2007), pp. 289–315.